University of Alberta

Development and Applications of Stable Isotope Labelling Liquid Chromatography Mass Spectrometry for Quantitative Proteomics

by

Andy Lo

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry

Spring 2011
Edmonton, Alberta

Thesis Examination Committee

Dr. Liang Li, Professor, Department of Chemistry

Dr. John S. Klassen, Professor, Department of Chemistry

Dr. Frederick G. West, Professor, Department of Chemistry

Dr. Joel H. Weiner, Professor, Department of Biochemistry

Dr. Hélène Perreault, Professor, Department of Chemistry, University of Manitoba

for my parents

Abstract

This thesis describes the characterization, automation, and applications of a stable isotope labelling method for quantitative proteomics by mass spectrometry. Known as 2MEGA, the method uses guanidinylation to convert peptide lysine residues to homoarginine followed by reductive methylation of free peptide N-termini with isotopically encoded formaldehyde. 2MEGA was shown to be applicable to the identification of membrane proteins by increasing the percentage of lysine-containing peptides identified and by the observation of diagnostic a_1-ions for >95% of glycine N-terminal peptides and >99% of non-glycine N-terminated peptides. Subsequent work demonstrated that 2MEGA was readily automatable with a commercially available liquid handler. With minor reagent substitutions, the 2MEGA labelling method was used to simultaneously process twelve samples. Over 98% labelling efficiency was observed, with the most common side reaction products being N-terminal guanidinylation (~2%) for glycine and alanine N-terminal peptides. Various front-end protein sample preparation methods were found to be compatible with the procedure.

Reciprocal labelling was used to evaluate the internal consistency from quantitative peptide sample comparisons by switching the original isotopic labelling assignment and analyzing the resultant sample mixture. With approximately 60% overlap in peptide identifications, reversal of the isotopic labels was not found significantly affect the observed quantification ratios, as evidenced by the internal consistency in the peptide quantification ratios (an average of 1.29-fold relative difference for the entire *E. coli* dataset). For over 90% of peptides, the relative error was less than 50%. After

discarding 1% of the quantified peptides, approximately 1% of protein matches were lost, but with significant gains in the internal consistency of quantification values. The large-scale quantitative analysis suggested that data processing can greatly influence the overall consistency observed in proteomics experiments. Reciprocal labelling was also applied to the analysis of a human carcinoma cell line deficient in Bax, a key protein in stimuli-induced apoptosis, in which 200 proteins with a significant change abundance difference were identified from Bax-expressing and Bax-deficient samples.

Overall, the thesis highlights the potential of the 2MEGA labelling method, its applicability to high throughput MS-based proteomics applications, and suggests an increased role for quantitative mass spectrometry as a routine bioanalytical methodology.

Acknowledgements

I would like to express my gratitude to Dr. Liang Li for his guidance during my Ph. D. studies. In addition to providing a research assistantship that allowed me to focus on laboratory work, his enthusiasm and spirited approach to research will leave a lasting impression for many years to come.

I wish to recognize the members of the thesis examination committee, Dr. J. S. Klassen, Dr. H. Perreault, Dr. J. H. Weiner, and Dr. F. G. West, in addition to Dr. R. L. McCreery for their careful reading of the thesis and thoughtful commentary.

All of the work in this thesis was carried out with the support of many other individuals, including Dr. R. Lai, Dr. M. Pasdar, Dr. J. H. Weiner, and members of their respective research groups. I thank them for their contributions and assistance, particularly for their expertise in biological matters.

I consider myself fortunate to have been a student in the Department of Chemistry, with its excellent support staff. The members of the Mass Spectrometry Facility, Biological Services, Shipping/Receiving Department, Machine Shop, and Electronics Shop have created an environment highly conducive to research by taking care of an array of difficult issues with the utmost professionalism.

Over the course of my studies, I have had the opportunity to interact with many members of Dr. Li's research group, both past and present. While too numerous to enumerate, I would like to acknowledge them for the many intellectually stimulating and challenging discussions we have had.

On a personal note, the enduring support and good humour of my parents and brothers (J, S, and D) remain a constant in my life and have made my time in Edmonton bearable. I am grateful for their honest advice and kind words. To my chagrin, I missed their company more than I ever could have expected.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2D	two dimensional
2MEGA	dimethylation after guanidinylation
ACN	acetonitrile
AALS	anionic acid labile surfactant
AQUA	absolute quantification
CHCA	α –cyano-4-hydroxycinnamic acid
CID	collision induced dissociation
CV	coefficient of variation
DHB	2, 5-dihydroxybenzoic acid
DIABLO	direct inhibitor of apoptosis binding protein with low pI
DR4/DR5	death receptor 4/death receptor 5
DISC	death-inducing signaling complex
ECL	enhanced chemiluminescence
ESI	electrospray ionization
ETD	electron transfer dissociation
FADD	Fas-associated death domain
GADPH	glyceraldehyde-3-phosphate dehydrogenase
GRAVY	grand average of hydropathy
HCCA	α-cyanohydroxycinnamic acid
HMGB1	high mobility group protein B1
HSP	heat shock protein
ICAT	isotope coded affinity tag
ICR	ion cyclotron resonance
IEF	isoelectric focussing

LC	liquid chromatography
LRPPRC	leucine-rich PPR motif-containing protein
m/z	mass to charge ratio
MALDI	matrix assisted laser desorption ionization
MIF	migration inhibitory factor
MMR	mismatch repair
MPT	mitochondrial permeability transition pore
MRM	multiple reaction monitoring
MS	mass spectrometry/mass spectrometric
MS/MS	tandem mass spectrometry
PrP^{C}	cellular prion protein
PrP^{res}	protease resistant PrP
PrP^{Sc}	prion protein from scrapie
QQQ	triple quadrupole mass spectrometer
QTOF	quadrupole time-of-flight
RP	reverse phase
RPLC	reverse phase liquid chromatography
S/N	signal-to-noise
SCX	strong cation exchange
SDS	sodium dodecyl sulfate
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SILAC	stable isotope labelling by amino acids in cell culture
SISCAPA	stable isotope standards and capture by anti-peptide antibodies
SPE	solid phase extraction
SRM	selected reaction monitoring

TCEP	tris(carboxyethyl) phosphine
TCP-1	T-complex protein-1
TFA	trifluoroacetic acid
TNF	tumor necrosis factor
TRAIL	tumor necrosis factor-related apoptosis inducing ligand
VDAC1	voltage-dependent anion channel 1
VDAC2	voltage-dependent anion channel 2

Chapter 1: Introduction

Proteomics involves the simultaneous analysis of hundreds to thousands of proteins from a defined system at a given time, with the ultimate goal being the complete description of all protein components. Studies focussed on peptide and protein identification were burgeoning in the early 2000s; a direct consequence was the development and refinement of robust mass spectrometry (MS) based identification technologies.[1, 2] As these strategies have matured, recent research efforts have sought to expand the information from proteomics experiments beyond protein identification. One logical extension in the ultimate goal of proteome characterization is proteome quantification. Shifting patterns in protein expression are an expected outcome during the cellular lifecycle in response to various endogenous and exogenous stimuli. Although the primary focus of this thesis is on method development for quantification of proteomes, it is particularly instructive to describe considerations in standard protein identification experiments to understand the function of each process and how it can influence quantification experiments.

Multiple dimensions of separation are generally required for the complete identification of all proteins within a sample when considering the dynamic range of proteins within a sample, protein sequence variations (splicing variants), and post translational modifications.[3, 4] The inherent dynamic range and the various physicochemical properties of proteins necessitate the use of chromatographic methods to separate them, which aids in detection of low abundance proteins. Even under optimal conditions, the dynamic range of current mass spectrometers is on the order of 10^3 - 10^4,[5] while the dynamic range of proteins for complex matrices, such as plasma, is believed to be on the order of at least 10^7.[6] This analytical challenge is reflected in the undersampling problem: protein and protein digests are highly complex mixtures and re-analysis of the same sample leads to identification of previously unidentified proteins. Multiple rounds of preparative separation and analytical chromatography are employed to remove interfering species and to simplify peptide mixtures by fractionation. Development of increasingly sophisticated

chromatographic and bioanalytical strategies to enrich specific proteins and peptides has led to continuing advances in protein identification. Elaborate front-end sample preparation has a profound impact on quantification accuracy, as each experimental procedure introduces some variability.[7, 8] No one quantification method is able to address protein samples from all study systems and a "toolbox" of techniques is required to adequately address the diversity of proteome samples. Furthermore, the inherent measurement variability of mass spectrometry, both in sample introduction (ionization) and measurement, presents its own challenges.

1.1 Protein Purification

Most samples for proteomics experiments are derived from cultured cellular samples. In this case, protein extraction often involves rupturing the cellular membrane via physical action (e.g., sonication or pressure), the use of extraction buffers containing detergents that weaken cellular structure, or a combination of both. Proteins from other sources often require more elaborate sample preparation procedures to remove cellular components and extracellular material that may interfere with MS analysis. Removal of lipids and other small organic molecules can be effectively achieved with solvent[9] or trichloroacetic acid protein precipitation[10]. For studies specific to particular organelles[11] (e.g., nuclei or mitochondria), selective extraction and centrifugation methods can be used to specifically enrich organelles of interest before downstream processing.

Once reasonably pure protein samples are obtained, there are a variety of methods used for the analytical separation of proteins. The most widely used protein level separation method, sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), is noted for its high resolution and fair loading capacity (~50 µg/lane). While digestion and extraction of peptides from the gel matrix can be tedious, the process can be partially automated using robotic instrumentation. Although laborious, a recent report demonstrated that more proteins were identified from honeybee head samples separated by SDS-PAGE than by isoelectric focussing (IEF) or protein reverse phase liquid chromatography (protein RPLC), likely due to the fair peptide recovery

(~90%) and high resolving power of SDS-PAGE.[12] Solution based separation methods, such as IEF and protein RPLC, offer the advantage that proteins are recovered in solution and can be subjected to additional sample work-up procedures at both protein and peptide level. Certain polishing methods using antibodies have been used for specific enrichment of certain classes of proteins. Antibodies have been used to immunoprecipitate phosphoproteins[13] or ubiquitinylated[14] proteins to specifically enrich for these post-translational modifications and to separate them from relatively abundant unmodified proteins. Similarly, lectins and periodate oxidation-hydrazide resins have been used to select glycoproteins from samples[15]. More recently, protein interactions have been studied using tandem affinity purification tags. Tandem affinity purification uses a combination of two affinity purification tags (e.g., 6x His and biotin/biotin ligation sequence) on an expressed bait protein to co-purify interacting proteins under native conditions.[16-18]

1.2 Protein Digestion and Peptide Level Separation

Once purified proteins of interest are obtained, samples are typically digested into smaller peptides using enzymatic or chemical methods. The most common protein digestion method utilizes trypsin, an enzyme which cleaves C-terminally to lysine and arginine in peptides, unless the next residue is proline. While other protein structural features and residues may influence cleavage efficiency, this informal rule generally holds true.[19] Other enzymes that are commonly used include chymotrypsin (cleavage C-terminally to F, W, Y, unless the next residue is P), Lys-C (C-terminal to K), Lys-N (N-terminal to K), and Glu-C/V8 (cleavage C-terminally to E; sometimes D). All of these enzymes are characterized by relatively high specificity to generate peptides reproducibly from a given sample. Consistent enzyme performance is crucial to prevent production of non-specific cleavage products which would otherwise reduce the effective concentration of anticipated digestion products.

Chemical methods are less frequently employed for proteomics experiments and are typically used to address specific challenges. Although infrequently used nowadays due to reagent toxicity, cyanogen bromide selectively cleaves after

methionine residues, unless the next residue is serine. Cyanogen bromide can be effective for dealing with proteins that have lysine/arginine free regions, such as the transmembrane domains of membrane proteins.[20] Similarly, 2-nitro-5-thiocyanobenzoic acid has been used for selective cleavage at cysteine.[21] Microwave assisted acid hydrolysis of proteins using trifluoroacetic acid has been suggested for sequencing hydrophobic proteins insoluble in most aqueous digestion buffers[22], while hydrolysis with acetic acid generates peptides with cleavages on either side of aspartic acid residues.[23]

As with proteins, peptides can be fractionated using similar separation methods, such as isoelectric focussing[24] and capillary electrophoresis. Due to the compatibility of most reverse phase solvents and additives to electrospray ionization, reverse phase separation is the most common and typically last dimension of separation. Strong cation exchange at low pH is commonly used for peptide fractionation of tryptic peptides, since the unblocked N-terminus amine or side chain of lysine/arginine ensures retention of most peptides.[25, 26] Modified peptides that are blocked N-terminally (e.g., acetylation) and phosphorylated peptides will have decreased retention, since these modifications neutralize positive groups.[26] Strong cation exchange has been used to selectively enrich for these types of peptides.[27] Specific enrichment techniques for post-translational modifications include immobilized metal affinity chromatography or metal oxide for enrichment of phosphopeptides and antibody enrichment of phosphotyrosine peptides.[28] For MALDI-based proteomics studies, beads coated with anti-peptide antibodies have been directly spotted onto MALDI targets for direct analysis without LC purification.[29, 30] Boronic acid functionalized MALDI targets have also been used for on-target glycopeptide enrichment[31] and polymerized titanium dioxide coated wafers have been used for phosphopeptide enrichment.[32]

1.3 MS Analysis

1.3.1 Ionization Methods

Before peptides are analyzed by mass spectrometry, they need to be introduced into the mass spectrometer after becoming ionized. While the majority of proteomics studies use electrospray ionization (ESI), matrix assisted laser desorption ionization (MALDI) can provide complementary information, since the mechanisms of ionization are fundamentally different.

1.3.1.1 Electrospray Ionization (ESI)

Electrospray ionization begins when a solution is sprayed through a narrow bore steel or conducting capillary under the influence of an applied voltage,[33] typically between 2 to 3 kV. For larger bore tips, the solution is typically forced through the narrow tip at the end of the capillary under the influence of pressure, but for narrow bore tips, capillary action can slowly draw the analyte solution from the tip. Under the influence of the high electric field at the capillary tip, a Taylor cone is formed at a threshold voltage and the solution is sprayed in a fine mist of charged droplets. The solvent of the charged droplets evaporates, occasionally aided by the influence of a heated source region and/or coaxial gas. As the solvent evaporates, the increasing charge-to-volume density on the droplet leads to the production of ionized gas phase analytes. There are two main theoretical models rationalizing the formation of gas phase analyte ions: the charged residue model[34] and the ion evaporation model.[35]

The charged residue model predicts the formation of smaller fission droplets from the main droplet, in which the fission products contain a high proportion of the excess charge in comparison to the volume of the smaller fission product. If there is an analyte molecule within this smaller droplet, continuing desolvation eventually leads to charge transfer to the analyte and formation of the gas phase ion. Alternatively, the ion evaporation model suggests that gas phase ions are formed directly from droplets, without the production of the smaller fission products. As solvent evaporates, the excessive surface charges begin to repel the charged analytes on the droplet surface. Once the charge density is high enough to overcome the Rayleigh limit, the repulsive force of the excessive surface charges leads to production of a gas phase analyte ion. Depending the analyte, it has been suggested that either model may be better at rationalizing the formation of gas phase ions. For large proteins, the charged residue

model is considered to be a better approximation for ion formation, whereas for smaller analytes, the ion evaporation model is believed to more accurately model ion formation. A key consideration, implied by both models, is that electrospray ionization is a competitive process and the analyte composition will directly affect the observed response through factors such as gas phase basicity and surface activity.

1.3.1.2 Matrix Assisted Laser Desorption Ionization (MALDI)

Matrix assisted laser desorption ionization involves formation of gas phase ions from a solid analyte/matrix mixture that has been on a spotted onto a sample target.[36, 37] Since reverse phase LC separation is essentially required for analysis of proteome digests, the LC eluate containing peptides is deposited directly onto a solid target. Depending on the instrumental configuration, a matrix may be added with the LC eluent[38], added after the sample has dried[39], or the LC eluent may be spotted onto pre-dried matrix spots[40]. While the structure of MALDI matrices vary, they are generally characterized as small, organic molecules with labile protons (i.e., carboxylic acid moieties) that have reasonable absorption coefficients for UV light and can incorporate analyte molecules upon crystallization. The two most common MALDI matrices are α-cyanohydroxycinnamic acid (CHCA/HCCA) and 2, 5-dihydroxybenzoic acid (DHB). CHCA is a pale yellow solid that forms a uniformly flat sample spot, whereas DHB is white solid that forms small needles on the target. Depending on the analytes, one matrix may give better signal response or cleaner spectra than the other. Since the ionization mechanism in MALDI is not perfectly understood, this is generally determined empirically. Samples are acidified before matrix addition on the plate in order to add ionization. Additives may also be used to promote ionization of certain compounds or reduce formation of matrix clusters and salt adducts during ionization. Phosphoric acid is a common additive for the analysis of phosphopeptides when using DHB and ammonium citrate is a common additive during peptide and oligonucleotide analysis to prevent formation of matrix clusters and sodium adducts. UV light from a laser is irradiated onto the sample spots to promote ionization. Both CHCA and DHB have reasonable absorbance coefficients near 337 nm, which is the primary emission wavelength of the nitrogen laser commonly used in most MALDI interfaces.

1.4 MS Instrumentation

There are a variety of mass analyzers available that can be used in different configurations for the analysis of peptides and proteins. The main instruments used for the described work were both quadrupole time-of-flight instruments (QTOFs) and will be discussed in further detail. Other types of instruments that are used for modern proteomics research such as one dimensional ion traps, two dimensional ion traps (Orbitrap)[5], and ion cyclotron resonance (ICR) mass spectrometers[41] have been the subjects of some reviews. Each type of mass spectrometer has its own performance characteristics in terms of mass accuracy, mass resolution, and spectral acquisition rate. While the discussion will be limited to a discussion about collisionally induced dissociation (CID), newer dissociation techniques are gaining prominence in proteomics, such as electron transfer dissociation (ETD), and have recently been reviewed.[42]

1.4.1. Quadrupole Mass Analyzer

Quadrupoles are four cylindrical metal rods that are arranged in a square or near-square configuration. The diametrically opposed rods are paired and the same voltage and a radiofrequency are applied for each pair. Depending on the specific voltage and frequency applied, ions of a particular *m/z* ratio will be able to transit down the entire length of the rods; other ions have unstable trajectories and are lost. When the mass filtering of the quadrupole is not required, such as during MS analysis, the voltages are turned off, allowing the quadrupole to act as a broad bandpass filter that transmits ions over a wide *m/z* range. By setting one pair of rods as a low mass filter (rejecting ions below a specific *m/z* threshold) and the second pair of rods as a high mass filter (rejecting ions above a specific *m/z* threshold), ions within a mass range of ±0.7 Da can be selected for MS/MS.

1.4.2. Time-of-Flight Mass Spectrometer

Time-of-flight mass analyzers are field free regions; no electric or magnetic fields are applied across the length of the flight tube. The ions are pulsed in by use of an extraction voltage that pushes the ion packet into the flight tube with minimal

distance dispersion in the direction of the flight path. The ions travel down the length of the tube and are typically reflected back by a reflectron. The reflectron is used to minimize the kinetic energy dispersion originating from distance dispersion when the ions are pulsed into the time-of-flight tube. The reflectron partially compensates for this energy dispersion by using successive sets of plates within which an electric field gradient is created. As ions with different kinetic energy enter the field, higher energy ions penetrate more deeply into the reflectron, increasing their flight path length and observed flight time. The net effect is improved mass resolution (m/Δm = ~10,000 to 20,000) with minimal losses in sensitivity. The reflectron is typically tilted at an angle to insure that the ions do not transit back exactly towards the extraction plate/pusher. The angled ions are then measured by the detector, which is usually placed just beside the extraction plate/pusher at a slight angle. A multichannel plate is used to amplify the signal from impinging ions and an analog to digital converter or a time to digital converter is used to determine the number of ions and total flight time. Based on the flight time, the *m/z* ratio of the ions can be calculated:

$$\frac{m}{z} = 2\ e\ V\ \left(\frac{t}{L}\right)^2$$

where:

m is the mass of the ion

z is the charge state of the ion

e is the elementary charge

V is the acceleration potential

t is the flight time

L is the length of the flight tube

1.4.3. Hybrid Quadrupole – ToF Mass Spectrometer

A generalized schematic of a hybrid quadrupole time-of-flight mass spectrometer is shown in Figure 1.1. Analytes are separated by RPLC and are ionized with an electrospray source. In the case of MALDI-based instruments, the target plate is loaded into the vacuum region and a laser pulse is brought in either by a fibre optic

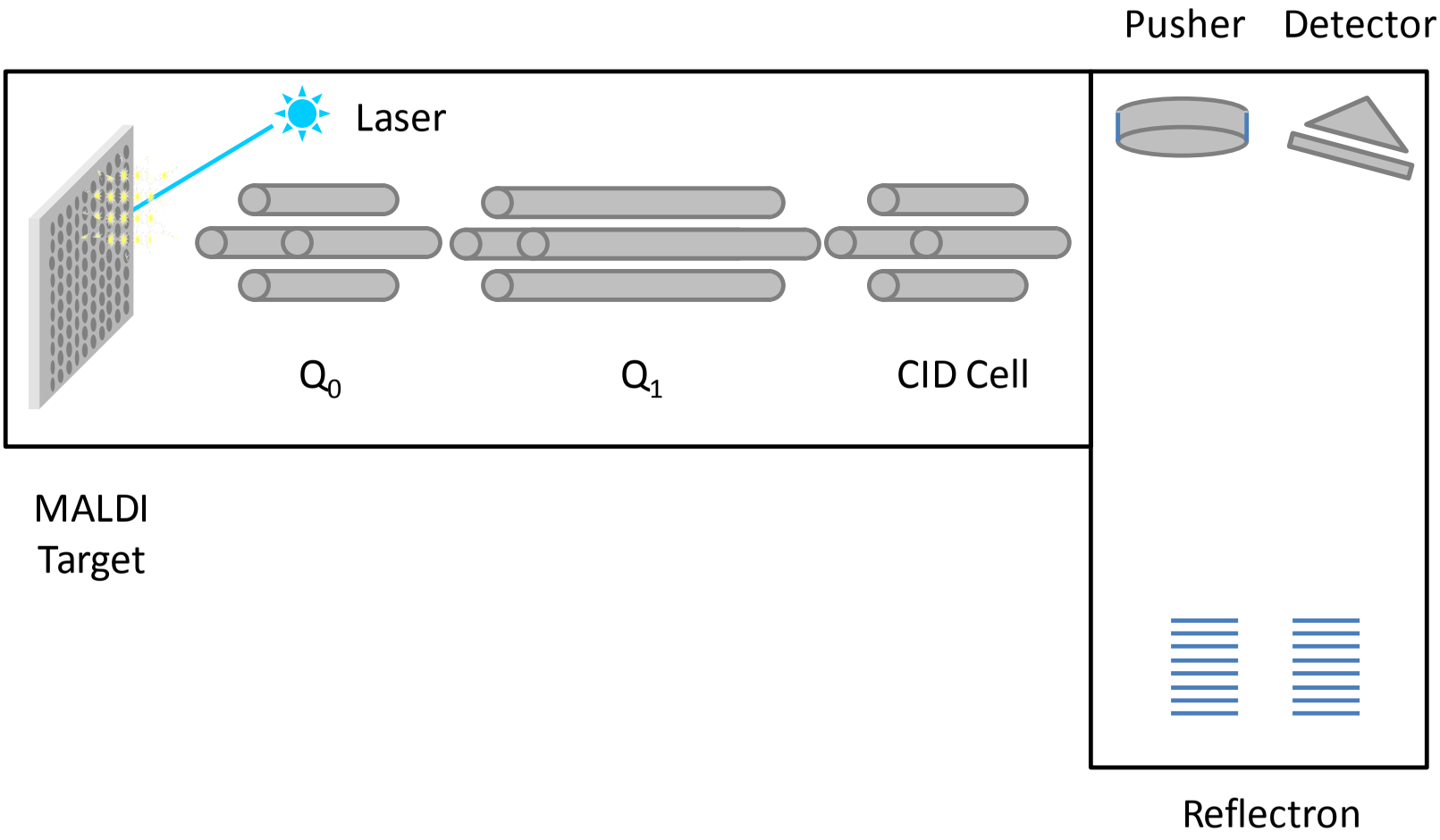


Figure 1.1 A simplified schematic of a quadrupole – time of flight hybrid mass spectrometer with a MALDI source

cable or mirrors to produce gas phase ions. Guiding optics (typically a short RF-only focussing quadrupole, Q_0) generate a narrow beam of ions for mass filtering by the analyzing quadrupole, Q_1. After passing the collision cell, the ions (regardless if they are precursor or product ions) are focussed into a narrow ion beam by collisional cooling or additional ion optics. A pusher is then pulsed to introduce ions into the orthogonally situated time-of-flight tube. An orthogonal configuration is used to minimize the distance dispersion of the ions, which improves mass resolution. The ions then travel towards the reflectron and are then reflected back where they are measured by the detector, as previously described. Here, the total flight time is defined as the transit time between the pusher and the detector.

1.5 MS/MS

A typical duty cycle during mass spectrometric analysis of peptides is initiated with the acquisition of an MS spectrum. The quadrupole is set as a broad bandpass filter, allowing ions over the entire *m/z* range (*m/z* 300 to 2000 for ESI-based instruments; *m/z* 700 to 4000 for MALDI-based instruments) to be measured by the time-of-flight mass spectrometer. Peaks are then quickly processed and the most intense peaks are selected for MS/MS fragmentation. To collect an MS/MS spectrum, the quadrupole is set to allow only ions with a particular *m/z* ratio through. Precursor ions are fragmented in the collision cell and the product ions are separated by the time-of-flight mass spectrometer and measured at the detector. The timescale for each MS and MS/MS spectrum is typically on the order of one second for the instrumentation used in this study.

1.5.1 Collision Induced Dissociation (CID) and Product Ion Series Nomenclature

MS/MS spectral acquisition begins when ions selected by the quadrupole enter the collision cell to be fragmented. The collision cell is a hexapole or octupole within a set of acceleration plates. A slight voltage is applied across the plates that accelerate the ions through the cell. Collisions with a neutral, inert bath gas, typically nitrogen or argon, increases the internal energy within the ions, leading to intramolecular reactions that result in bond fragmentation. A hexapole or octupole is used to focus

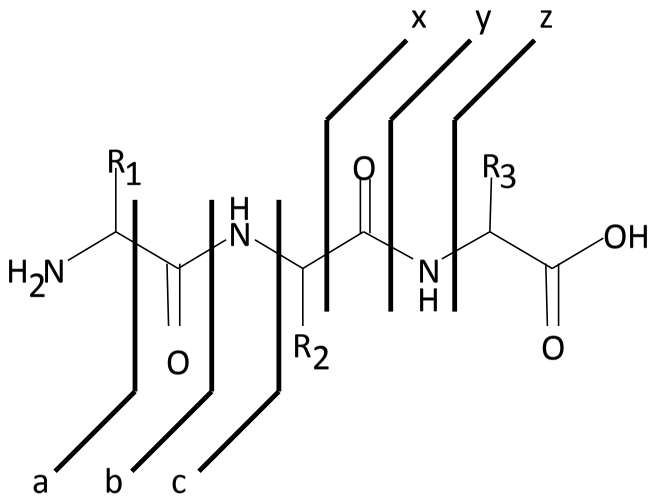


Figure 1.2 Product ion nomenclature

ions within the center of the cell to prevent losses from scattering. Although there are many bonds within a peptide, the most common fragmentations occur along the amide backbone of the peptide. Using the general nomenclature, the most commonly observed ions are b- and y- ions resulting from cleavage of the C-N bond in the amide backbone as shown in Figure 1.2. Other ion fragment series are less common and may be enhanced in higher energy fragmentation methods. Neutral losses of water (-18 Da) can be observed for serine, threonine, aspartic acid, and glutamic acid containing fragment ions and ammonia (-17 Da) can be observed for lysine, arginine, asparagine, and glutamine containing fragments, but these are generally weaker than the parent fragment.

1.5.2 Database Searching

Database searching uses both the precursor *m/z* and the product ion *m/z* ratios in the MS/MS spectrum. From a genomic or sequence database, masses of theoretical peptide sequences are calculated based on the digestion method used. The *m/z* ratio of the precursor and its charge state are used to calculate the mass of the peptide and only mass matched theoretical peptides, within a stated tolerance, are interrogated further. Each of these potential matches is then evaluated to determine how well the theoretical sequence accounts for features in the experimental MS/MS spectrum. Various scoring algorithms are used in database search engines (e.g., X!Tandem,[43] MASCOT,[44] and SEQUEST[45]) and each have their own features. In general, higher confidence matches will have strong and medium intensity peaks assigned to appropriate fragment ions (typically b- and y- ions) and fewer assignments to neutral loss peaks (i.e., loss of water or ammonia) in the absence of the parent fragment. Missing theoretical peaks are generally not heavily penalized in scoring, since favourable fragmentations may dominate a spectrum. Once a score is determined for each potential match, the highest scoring match is considered against the other candidates to determine whether the match is statistically significant or potentially due to random matching. As the size of the database increases, the number of candidate peptides also increases, leading to stricter matching criteria. Not all spectra will meet the threshold and those which cannot are considered unassigned. Alternative matching strategies using sequence tags (short “sequenced” stretches of amino acids

from an MS/MS spectrum) are comparatively rare, but such algorithms, such as Paragon[46], have been used for database searching. These algorithms perform an initial *de novo* sequencing of the MS/MS spectrum and the highest confidence *de novo* portions, typically 3-4 amino acid residues long, are used as the basis for the database search. When the same string of amino acids is found in the database, the algorithm attempts to align the *de novo* sequence string with the existing sequence, considering expected peptide products as well as non-specific cleavages, single amino acid substitutions, and post-translational modifications to rationalize the precursor *m/z* ratio and MS/MS fragmentation pattern observed. Regardless of the search strategy method employed, identified peptides are combined together to generate a final protein list. While generating the parsimonious list of proteins for a given set of peptides remains the *de facto* standard, newer clustering algorithms have sought to use weighting functions to generate protein lists that best rationalize the majority of the data. Once peptides and proteins have been identified, quantification results can be extracted and processed from the raw data.

1.6 Quantification Strategies

As a direct result of variations in their primary structure, each protein will produce peptides of variable length, amino acid composition, and amino acid sequence. One practical consequence is that the differing physicochemical properties of a protein's digestion products will lead to differential ionization efficiencies by electrospray ionization or matrix-assisted laser desorption ionization. Absolute peak intensities, as measured by mass spectrometry, are not necessarily indicative of peptide concentrations and are further affected by other simultaneously ionized components such as peptides, LC eluent buffer components, and other chemical interferences. In order to address this issue, a variety of strategies for the identification and quantification of peptides from complex digests have been developed.

1.6.1 Label Free Methods

1.6.1.1 Spectral Counting

While signal response may not be directly related to peptide concentration, there is nevertheless an increase in observed response as peptide concentration increases.[47] Unlike many other approaches that will be discussed *vide infra* that report relative concentrations of components, spectral counting strategies use a variety of weighting equations to approximate protein concentration.[48, 49] Spectral counting exploits the inherent MS and MS/MS duty cycle and uses the frequency of MS/MS sequencing events as a surrogate for protein concentration. A typical duty cycle used for peptide identification is initiated with a MS or survey scan, which identifies precursor ions of suitable intensity for subsequent MS/MS sequencing. Since several suitable peaks are typically observed in a single MS spectrum, peak selection algorithms will generally select peaks based on intensity to maximize signal response during MS/MS analysis. This is particularly important in LC-MS experiments of complex protein digests, where complete chromatographic resolution of sample peptides is difficult to achieve.

After a set of peaks are sequentially analyzed by MS/MS, the mass spectrometer begins the duty cycle again and acquires another MS spectrum to select more peaks for further analysis. In spectral counting strategies, the number of MS/MS sequencing events for a given protein's peptides is considered indicative of overall concentration. As a protein's concentration increases, more of its constituent peptides are selected for MS/MS analysis or redundantly analyzed. A variety of equations have been used to correlate protein concentration to the number of MS/MS sequencing events. While initial models normalized the number of sequenced residues in a protein to its concentration, recent improvements have shown an exponential dependency on protein concentration to the number of sequencing events.[48]

Recent reports have also demonstrated that the dynamic exclusion of peptides has only a minimal effect on protein abundance estimation if the sample is sufficiently complex.[50] In order to prevent redundant analysis of peaks with high intensity, most peak selection algorithms use dynamic exclusion, which prevents reanalysis of a given *m/z* ratio for a particular amount of time once it has been selected for MS/MS analysis. Since spectral counting methods do not utilize information beyond that required for

protein identification (successfully assigned MS/MS spectra), the quantitative data generated can be readily used as a first estimate for protein quantification. Despite the universal applicability of spectral counting strategies, there are inherent drawbacks. With their unique primary structures, proteins with several short or long peptides may be under-represented due to infrequent selection of these peaks for MS/MS. Short peptides (< 6 residues) may not reach the minimum *m/z* cutoff (typically *m/z* 300-350 in ESI-based instruments) used during the MS survey scan, whereas long peptides (>30 residues) suffer from relatively low MS response and limited fragmentation in MS/MS, which prevents successful MS/MS identification. Compositional biases within a protein can also produce peptides that are easily ionized, which can lead to artificially high abundance estimations. Decreased sample complexity in simple mixtures will also increase the average number of sequencing events for each protein in the sample, without reflecting a change in absolute protein concentrations. Spectral counting strategies are particularly sensitive to the proteins analyzed and are generally considered semi-quantitative methods for protein estimation.[51]

1.6.1.2 Ion Current Intensity

Another method for label free analysis utilizes the ion current intensity of a peptide, typically presented as an extracted ion chromatogram when used for the comparison of samples.[52] While the experiment uses the same data obtained during standard LC-MS proteomics experiments, a significant amount of data processing is required for quantification. Samples to be compared are prepared under identical conditions and run under the same LC-MS conditions. One of the samples is set as the reference and its MS/MS peak lists are processed and searched against the appropriate database to identify peptides. The elution profile of each identified peptide is reconstructed by using the MS survey data around the retention time from when the peptide's MS/MS spectrum was taken. Using the precursor *m/z* of identified peptides and their chromatographic retention times, identical peaks/features within each of the other samples are identified. Provided that the feature is considered sufficiently similar, as defined by tolerances for *m/z* ratio mass accuracy and potential retention time shift, the matching peak feature in the comparison sample is reconstructed in a

similar fashion. The reconstructed ion intensities are then integrated over the elution profiles for each peptide and the relative response is reported.

The direct comparison of peptide signals across multiple runs is deceptively simple: by comparing the extracted ion chromatograms of each component, it should be facile to gauge the relative response and relative intensity of the same component in different samples. However, the number of features from a single LC-MS run is typically on the order of several thousand and represents a significant challenge. Accurate feature alignment and retention time correction are active areas of research with continual refinements to existing algorithms.[53] Given the thousands of peptide components and multiple charge states accessible for each component, there can be many similar peaks within a reasonable *m/z* ratio error and time range that could be the feature of interest. One advantage of ion current methods is that peptide identifications from either a sample or reference run can be used to identify features for *m/z* and retention time matching, which can increase overall quantified proteome coverage.

One drawback of ion current based methods is a significant dependency on consistent chromatography and mass spectrometer behaviour. Variations in LC or MS instrument performance will adversely affect the separation efficiency and peptide response observed.[54, 55] Temperature fluctuations have been previously cited as a major concern influencing both LC separation and MS acquisition and needs to be controlled. Retention time drift and mass accuracy shifts can be significant on the scale of regular laboratory temperature fluctuations. One corrective approach uses spiked peptides standards eluting during analysis of samples in order to aid in chromatographic retention time alignment.[56]

Performance characteristics of ion current intensity methods are currently difficult to establish. Systematic approaches to determine the accuracy of ion current based methods, against both absolute values and other quantification methods, have produced a range of results. As the undersampling problem suggests, there is some inherent variability in peptide response due to small changes in peptide elution during

LC separation. With 70% similarity in the peptide identifications between replicate analyses of the identical sample, it can be seen that the LC separation and MS analysis of samples is not an entirely deterministic process. Even when considering peak features (non-identified peptides), only 53% of observed peaks were consistently observed across five replicate analysis of the same sample.[57] Given the sensitivity of electrospray ionization to components in the eluent, changes in ionization efficiency due to local variations will be inappropriately reflected as changes in abundance. Lastly, ion current or intensity based approaches are unsuitable for quantitative analysis of peptides by MALDI MS. Variations in signal response resulting from inhomogeneous matrix spotting are typically significant and cannot be corrected in a straightforward way. Depending on the matrix used, "hot spots" can give spikes in analyte response that are otherwise not observed over other areas of the same MALDI sample spot.

1.6.2. Label Based

Label based methods are characterized by the incorporation of stable isotopes to produce isotopomers that have nearly identical chemical behaviour, but can be readily distinguished by the mass spectrometry. By simultaneously analyzing both a sample and reference, variations from post-labelling sample preparation, chromatographic separation, and MS behaviour are largely corrected. Depending on when isotopes are introduced, differentially labelled proteins and peptides are subjected to the same sample workup procedure, which further minimizes sample processing variations. However, since introduction of isotopes is itself an experimental procedure, errors in isotope incorporation can lead to deviations not reflective of actual quantitative differences. Another disadvantage to label based methods is the cost of isotopic reagent. Depending on the isotopes used and the complexity of the precursors, the cost for a single experiment can range from <$1 per sample to over $30,000. Here, the two general classes of label-based approaches for quantitative discovery proteomics are outlined.

1.6.2.1 Metabolic Methods

By growing cells, plants, and animals on isotopically enriched sources, the resultant protein samples are coded based on their growth conditions. With the variety of isotopically labelled small molecules and inorganic salts available, a litany of study systems can be metabolically labelled for quantitative proteomics experiments. Metabolic methods offer a variety of advantages and present their own challenges. Unlike all other quantification methods, metabolically labelled samples can be combined and subjected to protein level separation methods, such as SDS-PAGE or protein RPLC, and affinity chromatography methods, such as phosphoprotein enrichment[58] or immunoprecipitation[7]. This is particularly important given the desire to study low-stoichiometry post-translational modifications of proteins. Early sample mixing prevents samples from being subjected to separate downstream processing steps, ultimately reducing experimental variation. One major limitation of metabolic methods is that samples need to be cultured on a specific medium. Samples of clinical significance, such as human blood and plasma, or arising from other naturally occurring sources cannot be metabolically labelled. Subtle effects on protein expression profile due to growth on isotopically enriched media also need to evaluated carefully. Lastly, metabolic methods tend to be relatively costly as lengthy growth periods are required to insure thorough incorporation. While simple systems such as *E. coli* or yeast can be grown on media with inorganic or simple precursors, heterotrophs, such as fish or mice, require a food source that has already been grown on enriched media for the duration of their lifetime, which can be an expensive proposition. Despite these shortcomings, metabolic labelling strategies are widely employed due to their many advantages. Metabolic approaches can be broadly divided into two groups, depending on type of incorporation.

1.6.2.1.1 Complete Label Incorporation

Species ranging from bacteria to rats[59] have been successfully ^{15}N labelled, by either directly using isotope sources such as inorganic salts (e.g., $K^{15}NO_3$) to spirulina (cyanobacteria) grown on ^{15}N salts. After a sufficiently long growth period, replacement of the endogenous ^{14}N with ^{15}N at levels close to the enrichment level of the feedstock can be obtained. For unicellular systems, this typically occurs after five to ten doublings; for mammals such as mice, two generations fed ^{15}N enriched diets can

approach >95% incorporation. ^{15}N is particularly important in plant proteomics as a variety of inorganic salts are comparatively easy to introduce. Complete label incorporation methods have been used in "label chase" or "pulse chase" time course studies where a system is grown on media that contains exclusively either a light or heavy isotope. As the system is subjected to a particular stimulus, the medium is simultaneously switched to the alternate isotope. System response to the stimulus is measured by the incorporation rate of the added isotope over time. One salient example is this strategy is the measurement of protein turnover rate.[60]

While ^{15}N imparts a kinetic isotope effect during reverse phase separation of peptides and ^{13}C does not, the difference in cost between the isotopes is a practical consideration. Correction for incomplete labelling of samples is another experimental consideration. Assuming an isotopic enrichment level of 99%, a peptide with ten amino acid residues is expected to contain at least ten nitrogens from N-terminus and each of the backbone amide groups. While the major product will be an entirely ^{15}N coded peptide, the ~1% chance of ^{14}N incorporation at each nitrogen atom, will lead to a -1 Da peak at approximately 10% height of the fully-labelled peak. Proper "deisotoping" of peaks due to this ^{14}N peak contribution is required before analysis of MS and MS/MS spectra to ensure proper peak selection.

1.6.2.1.2 Stable Isotope Labelling by Amino Acids in Cell Culture (SILAC)

Stable isotope labelling by amino acids in cell culture (SILAC) was first reported in 2003. While the initial report describing SILAC used d_{10}-leucine[61], the kinetic isotope effect of deuterium incorporation during reverse phase chromatography can be severe when deuteriums are incorporated onto hydrophobic groups, such as the isobutyl side chain of leucine. Modern versions typically use a combination of uniformly ^{13}C and/or ^{15}N enriched lysine or arginine in the growth medium of bacteria, yeast, and human cells.[62] Lysine and arginine are used because trypsin selectively cleaves C-terminally to lysine and arginine during protein digestion. The incorporation of a single isotopic label at the C-terminus of peptides after complete digestion allows for facile data analysis and removes ambiguity for quantification ratio reporting once the peptide is successfully sequenced. Yeast is a common study system for SILAC experiments, since

strains that are lysine and/or arginine auxotrophic can be used to force uptake of isotopically labelled lysine or arginine.[63] It has been noted that interconversion of isotopically labelled arginine to proline does occur *in vivo* and a correction factor for proline containing peptides is often applied based on calculated percent conversion.[63, 64] Expansion of the SILAC methodology to higher organisms was achieved through appropriate diet supplementation for insects[58] and mice[65]. While reports are relatively recent, this method can provide organ, tissue, and other biofluid samples, which are typically inaccessible by cell culture based methods. The samples can be processed in conjunction with animal disease model systems to follow protein changes.

1.6.2.2 Chemical and Enzymatic Labelling Methods

The primary advantage of chemical labelling approaches is their general applicability to all types of samples, regardless of their origin. Intact proteins or protein digests are labelled with isotopically coded variants of the same reagent, ensuring equal reactivity and near identical response to downstream sample workup procedures. Chemical labelling approaches can generally be divided into two classes: non-isobaric tags and isobaric tags. Non-isobaric mass tags, also known as precursor-based methods, extract quantification information from the MS scan of the duty cycle. Peptides are reacted with non-isobaric, isotopic variants of the same reagent, which allows them to be distinguished in the survey scan based on the number of labels present and mass shift that each label imparts. Isobaric tags produce peptides of the same apparent mass in MS; quantification results are not determined until peptides are fragmented in MS/MS. Reagents are structurally similar, but differing patterns of isotope incorporation yield fragment ions of different isotopic composition and observed *m/z* ratio. Both methods offer their own advantages, features, and drawbacks and can be tailored for the study requirements.

1.6.2.2.1 Non-Isobaric Tags

Most chemical reagents for quantitative proteomics target the reactive functionalities of proteins:[66] the primary amine of the N-terminus, the side chain amine of lysine, and the thiol group of cysteine. The nucleophilic character of these groups is well-defined and side products can be minimized through careful stepwise reactions.

Other protocols convert the acidic groups of aspartic and glutamic acid to amides using a carbodiimide and $^{13}C_6$ aniline[67] or *p*-sulfanilic acid[68]. One of the most sophisticated tags is the isotope coded affinity tag (iCAT) reagent.[69, 70] The cleavable iCAT tag has four portions to its structure: a reactive group, an isotope coding group, an acid-cleavable linker and a purification tag. Disulfide bonds in the protein are reduced using tris(carboxyethyl) phosphine (TCEP) before the reagent is added. Free thiol groups react with the iodoalkyl group to form a thioether bond. Proteins are then digested and peptides with the labelling reagent are then selectively purified through the biotin purification tag using streptavidin beads. Since cysteine is a relatively rare amino acid, the purification step assists in enriching peptides that will provide quantitative information. The purification tag is then released after workup in acid leaving the cysteine residue derivatized with only the isotope coding group. Specific post-translational modifications have also been studied quantitatively by using the inherent reactivity of modifications as handles for labelling. One protocol describes the β-elimination of a phosphate group from a phosphoserine to produce dehydroalanine, which is then alkylated with an isotopically coded thiol in a Michael addition reaction.[71]

Enzymatic methods can also been used to incorporate stable isotopes. The most widely used method is tryptic digestion in ^{18}O water.[72, 73] Samples are separately digested with trypsin and dried down before reconstitution in either $H_2{}^{16}O$ or $H_2{}^{18}O$ and fresh trypsin is added to the sample. Although proteins are already digested, the hydrolysis mechanism of trypsin exchanges both of the ^{16}O atoms at the C-terminus of the peptide with ^{18}O. The samples are then mixed and subjected to downstream processing. The main issues with enzymatic labelling are potential back exchange if active trypsin remains once the samples are mixed[74] and the slow conversion the ^{18}O atoms under mildly acidic aqueous solutions.[75] Glycosylation has also been studied using the enzymatic action of PNGase F, an exoglycosidase that hydrolyzes N-linked glycans on asparagines.[76, 77] Asparagine is converted to an aspartic acid during the hydrolysis, which can be used to introduce an ^{18}O atom to track the site and frequency of glycosylation.

1.6.2.2.2 Isobaric Tags

The adoption of tandem mass tags has been widespread, as evidenced by an increasing number of publications since the initial report describing tandem mass tags was published in 2003.[78] While a few research based reagents have been developed, most reports use one of two commercially available products: iTRAQ and TMT. The generalized structure of the tandem mass tag has three features: the reactive group, the mass balance group, and the reporter group. The reactive group is a succinimidyl ester that reacts primarily with amines, although side products with the tyrosine are also observed. As with precursor based methods, amines are an attractive labelling group due to the ubiquity of free peptide N-termini and lysine side chain amines. Reporter groups typically contain functionalities (e.g., tertiary amines) that have higher gas phase basicity than other peptide functional groups (i.e., primary amines or amide carbonyls) to promote protonation. The reporter groups are generally located beside a bond that is labile during CID fragmentation. The ease of fragmentation and protonation aids in the formation of the reporter ions used for quantification. The balance and reporter groups have isotopes (^{13}C, ^{15}N, and ^{18}O) incorporated at various atoms in order to maintain the nominal isobaric mass of the modifying group. However, the pattern of isotopic incorporation in the reporter group will produce a fragment with a different *m/z* value. Isobaric tag designs capable of simultaneous comparison of six to eight samples are available.

An advantage of isobaric tags is that non-reporter peptide fragments are additive and contribute to overall signal intensity for peptide sequencing. Reduced instrument analysis times are also realized, since up to eight samples can be measured and quantified simultaneously. Since most precursor-based methods work on a duplex or triplex system, analyzing eight samples concomitantly can drastically improve throughput. A distinctive feature of isobaric tags is that reporter ions are measured in the MS/MS spectrum. Due to the filtering of precursor ions for MS/MS analysis, the spectral density of MS/MS spectra is far less than for MS spectra. As a result, fragments are generally observed with higher signal to noise ratios, due to the reduction of noise in the spectra. However, isobaric tags do have some noted shortcomings. Since sufficient fragmentation of the tags is required for quantification, the labile group bond may be comparatively weak to the peptide backbone bonds, leading to formation of

only reporter ions, hindering identification. Conversely, the labile bond in the tag may be comparatively strong to one or two peptide backbone cleavages. If insufficient numbers of reporter fragments are produced, particularly when relative abundance ratios are significantly different, no quantification ratio can be reported.[79] Precursor based methods are not sensitive to the particular fragmentation characteristics of a peptide and the quantification ratio can always be reported from the precursor intensities. More recently, isobaric tags have been found to "overcharge" unmodified and phosphorylated peptides by producing an increased percentage of +3 and +4 charge peptides, which have comparatively poor fragmentation and identification efficiency to +2 peptides.[80] Reagent cost can also be prohibitive for the high throughput analysis, since the tags need to be synthesized with costly isotopes in a very specific pattern.

1.7 Quantification of Known Peptides

The methods already discussed can be used for discovery-based proteomics experiments, in which hundreds to thousands of proteins can be identified and quantified in a single study. However, when the proteins and peptides of interest are known, targeted MS-based approaches can be used to sensitively and accurately measures changes in abundance.

1.7.1. Multiple Reaction Monitoring

Multiple reaction monitoring (MRM) can be used to selectively measure ion intensities from known peptides with improved quantification accuracy and precision. Triple quadrupole (QQQ) mass spectrometers are typically used for MRM experiments. Within the retention time range of a peptide, the first quadrupole is set to allow the *m/z* ratio of the precursor ions through and the second quadrupole is set to monitor a fragment ion without scanning the entire mass range. This significantly increases signal response, since only the particular transitions of interest are measured. Transitions are selected for their uniqueness to avoid interferences from isobaric peaks. MRM timetables are used to schedule specific transitions during the separation of peptide mixtures in order to maximize the number of peptides that can be measured in a single

chromatographic run. There is at least a 10-fold enhancement in sensitivity when using MRM for known peptides and the increased sampling along the elution profile of a peptide improves quantification accuracy and the dynamic range. While only useable for previously identified peptides, the ability to sensitively monitor hundreds of transitions (> 500) in a single MRM experiment makes it a valuable tool in well-defined study systems.

1.7.1.1 Synthesized Standards

Using solid phase peptide synthesis methods, isotopically labelled amino acid residues can be introduced into synthetic peptide standards and spiked into samples for concentration monitoring. This method is known as AQUA (absolute quantification) and was described in 2003.[81] With MRM monitoring of the endogenous and synthetic peptide, absolute quantification can be determined if the amount of synthetic peptide added is known. Peptide standards are chosen with a few design guidelines. Residues that are easily modified or altered under regular laboratory conditions are generally avoided; for example, oxidation of methionine/tryptophan and deamidation of asparagines/glutamine are not infrequently observed and are deprecated when designing peptide standards.

A related approach known as stable isotope standards and capture by anti-peptide antibodies (SISCAPA) uses antibodies to specifically enrich target peptides from a complex protein digest.[82] Protein samples are enzymatically digested and synthetic peptide standards are spiked in. Antibodies are used to enrich endogenous and standard peptides. After releasing the peptides from the antibodies, they are analyzed in a standard MRM experiment. The use of antibody enrichment, in conjunction with standard peptides and multiple reaction monitoring, greatly increases the dynamic range available and has been suggested for the analysis of low-abundance proteins without the need for additional fractionation. Reports have suggested up to an average of 120-fold increase in sensitivity over a dynamic range of 3 orders[82] and the development of anti-peptide antibody panels is currently being pursued for diagnostic applications.

The QconCAT method for absolute quantification was described in 2006.[83, 84] For proteins of interest, two or three peptides are chosen for monitoring the entire protein. An expression vector with all of the selected peptide sequences concatenated is transfected into a cell and grown with isotopically labelled media to produce a chimeric protein that, when digested, will produce the isotopically peptides for quantification. The chimeric protein can be purified, quantified, and spiked into protein samples subjected to downstream sample preparation steps, such as digestion.

1.8 Thesis Overview

This thesis seeks to further characterize and develop the 2MEGA labelling method for applications in quantitative proteomics.[85] The 2MEGA protocol uses guanidinylation to block the side chain amine of lysine before differential reductive methylation of free peptide N-termini with either light ($^{12}CH_2O$) or heavy ($^{13}CD_2O$) formaldehyde. The differentially labeled peptide N-termini ($(CH_3)_2$ vs. $(^{13}CD_2H)_2$) yields a 6.032 Da mass difference that can be clearly distinguished during MS analysis and that minimizes contributions from peaks in the isotopic distribution.

In Chapter 2, the 2MEGA labelling method was characterized by examining differences in the identified peptides from 2MEGA labelled and unlabelled samples from the *E. coli* membrane fraction. Since modification of peptides can alter ESI response and fragmentation patterns, the chapter examines potential differences between the two datasets. It was found that an increased percentage of lysine-containing peptides were identified in the 2MEGA labelled dataset, suggesting a reduction in the bias toward arginine-containing peptides. Furthermore, it was found through manual inspection of MS/MS spectra that the 2MEGA labelled samples have a strong a_1 ion (>98% of spectra), which is largely absent in the MS/MS spectra of the unlabelled samples. Glycine N-terminated peptides had a slightly lower frequency of a_1 ions. In the spectra of lysine or arginine N-terminated peptides, a_1 or a_1 related ions were often observed. The consistent appearance of the a_1 ion is suggested as a useful quality criterion for assignment of the N-terminal residue.

Chapter 3 is primarily concerned with issues arising from quantitative proteomics experiments of complex proteomes by LC-MALDI. Here, reciprocal labelling and targeted peak selection were used to increase the confidence in the quantitative results and to reduce the overall analysis time. One of the major challenges in large scale proteomics experiments is difficulty validating all of the results obtained with an orthogonal method. By using the LC-MALDI platform, single peaks and peak pairs from peptides indicating differential abundance were identified for MS/MS from the mostly unchanged peptides to preferentially identify differentially expressed proteins. This is particularly important for MALDI-based applications as the spectral acquisition rate of MALDI instruments is generally slower than for ESI instruments. Since a low error rate in the quantification accuracy across a large dataset can generate many proteins with supposedly differential expression, reciprocal labeling was used as an additional criterion for peak selection. Only peptide peaks and pairs suggesting differential abundance in both the forward and reciprocal labelled samples were selected for MS/MS analysis. While the approach may be considered overly conservative, this was done to minimize the impact of inherent variation in the quantification method. By comparing the proteomes of Bax-expressing and Bax-deficient human cells, 200 proteins with significantly changed abundance were identified. Many proteins identified in this list were found to be consistent with previous literature reports.

Chapter 4 attempts to address a fundamental issue in quantitative proteomics. Many protocols for differential isotopic labelling are complex and require a significant amount of analyst time. Furthermore, reagent cost is another concern, since the cost for incorporation of isotopes into specialized derivatization reagents can be significant. As a demonstration of the potential of the 2MEGA labelling method, the chemistry was modified to be compatible for automated labeling using a commercial liquid handler. Peptide samples from a variety of protein preparation buffers were simulated to evaluate the robustness of the chemistry. Modifications to the protocol included standardizing the guanidinylation conditions and changing the reducing agent from the toxic sodium cyanoborohydride to a safer alternative (2-picoline borane). Considering the amount of the formaldehyde required for ten samples (80 µL of 4% solution) and the current cost of the isotopically labeled formaldehyde ($^{13}CD_2O$, $500 for 1 mL of

20% solution), the current cost is ~$1/sample. With the potential to label several samples simultaneously, the application of 2MEGA labelling for high-throughput applications appears feasible.

In Chapter 5, the application of reciprocal labelling for the comparison of *E. coli* grown under aerobic and anaerobic conditions using 2D-LC MS was considered. By analyzing two different control mixtures (1 to 1 mixtures of the light and heavy labelled versions of the same peptide digest), it was found that the internal consistency in the quantification values from the common peptides between replicate analyses were within 50% for 95% of the peptides. Based on the distribution of the protein ratios after data analysis, it was found that over 95% of proteins lie within the range of 1.50 and 0.67 and proteins outside of this range can be considered significantly different than 1. The comparison experiments between aerobically and anaerobically grown *E. coli* using both forward and reciprocal labelling found that the percentage of overlapping peptide identifications between the replicates of the forward labelling were only slightly higher (~5%) than against the reciprocal datasets. The quantitative results from the forward and reciprocal datasets demonstrated that while most of the protein level data is qualitatively consistent, as evidenced by the near ideal behavior in the log-log plot of the protein ratios, a small percentage of proteins do not follow the expected trend. Reciprocal labelling was successfully used to identify peptides with inconsistent values by calculating the relative difference between the reported quantification results. This was used as a data quality metric to eliminate a small percentage of peptides. Over 280 proteins were found to be consistently differentially abundant in both the forward and reciprocal datasets and were used for bioinformatics analysis.

Chapter 6 briefly examines the potential of protein level separation prior to labelling at peptide level for quantitative proteomics. The study system was the plasma of elk infected with chronic wasting disease; the goal was to identify putative biomarkers of infection. Protein level separation by an affinity method or SDS-PAGE was necessitated by the high abundance proteins found in plasma. A blended reference was used in order to create an allowed range of quantification values that

could be used to detect outliers. A combination of database searching and partial *de novo* sequencing was successfully applied to identify proteins, but *de novo* sequencing followed by sequence homology analysis was found to be less successful. Here, the challenging combination of a lack of sequence database and the significant dynamic range in plasma protein concentrations gave less than promising results for biomarker detection. Chapter 7 briefly reviews the work and suggests future avenues of research, given the key findings in the thesis.

1.9 Conclusion

Mass spectrometry stands at the forefront of protein identification technologies, given its sensitivity and speed. The development of quantification techniques has added a dimension to the biological challenges that MS-based methods can successfully address. Research focused on increasing peptide identification efficiency, especially related to spectral acquisition rates, instrument sensitivity, and peptide chromatography will likely to lead to proportional improvements in the applicability of quantification-based MS-methods. When considering the number and dynamic range of protein components in a given biological system, the scope of analytical challenge cannot be easily addressed through a singular focus on any one portion of the research pipeline alone.

1.10 Literature Cited

(1) Wolters, D. A.; Washburn, M. P.; Yates, J. R. *Anal. Chem.* **2001**, *73*, 5683-5690.

(2) Washburn, M. P.; Wolters, D.; Yates, J. R. *Nat. Biotechnol.* **2001**, *19*, 242-247.

(3) de Godoy, L. M. F.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Froehlich, F.; Walther, T. C.; Mann, M. *Nature* **2008**, *455*, 1251-1254.

(4) Wang, H.; Chang-Wong, T.; Tang, H.-Y.; Speicher, D. W. *J. Proteome Res.* **2010**, *9*, 1032-1040.

(5) Gorshkov, M. V.; Good, D. M.; Lyutvinskiy, Y.; Yang, H.; Zubarev, R. A. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1846-1851.

(6) Anderson, N. L.; Anderson, N. G. *Mol. Cell. Proteomics* **2002**, *1*, 845-867.

(7) Zhang, G.; Fenyo, D.; Neubert, T. A. *J. Proteome Res.* **2009**, *8*, 1285-1292.

(8) Ji, C.; Zhang, N.; Damaraju, S.; Damaraju, V. L.; Carpenter, P.; Cass, C. E.; Li, L. *Anal. Chim. Acta* **2007**, *585*, 219-226.

(9) Simpson, D. M.; Beynon, R. J. *J. Proteome Res.* **2010**, *9*, 444-450.

(10) Wright, A. P. H.; Bruns, M.; Hartley, B. S. *Yeast* **1989**, *5*, 51-53.

(11) Wang, J.; Gutierrez, P.; Edwards, N.; Fenselau, C. *J. Proteome Res.* **2007**, *6*, 4601-4607.

(12) Fang, Y.; Robinson, D. P.; Foster, L. J. *J. Proteome Res.* **2010**, *9*, 1902-1912.

(13) Schumacher, J. A.; Crockett, D. K.; Elenitoba-Johnson, K. S. J.; Lim, M. S. *J. Mol. Diagn.* **2007**, *9*, 169-177.

(14) Vasilescu, J.; Smith Jeffrey, C.; Ethier, M.; Figeys, D. *J. Proteome Res.* **2005**, *4*, 2192-2200.

(15) McDonald, C. A.; Yang, J. Y.; Marathe, V.; Yen, T.-Y.; Macher, B. A. *Mol. Cell. Proteomics* **2009**, *8*, 287-301.

(16) Graumann, J.; Dunipace, L. A.; Seol, J. H.; McDonald, W. H.; Yates, J. R.; Wold, B. J.; Deshaies, R. J. *Mol. Cell. Proteomics* **2004**, *3*, 226-237.

(17) Wang, P.; Wu, F.; Ma, Y.; Li, L.; Lai, R.; Young, L. C. *J. Biol. Chem.* **2010**, *285*, 95-103.

(18) Wu, F.; Wang, P.; Young, L. C.; Lai, R.; Li, L. *Am. J. Pathol.* **2009**, *174*, 361-370.

(19) Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. *J. Proteome Res.* **2008**, *7*, 300-305.

(20) Blackler, A. R.; Speers, A. E.; Ladinsky, M. S.; Wu, C. C. *J. Proteome Res.* **2008**, *7*, 3028-3034.

(21) Iwasaki, M.; Masuda, T.; Tomita, M.; Ishihama, Y. *J. Proteome Res.* **2009**, *8*, 3169-3175.

(22) Zhong, H.; Marcus, S. L.; Li, L. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 471-481.

(23) Swatkoski, S.; Gutierrez, P.; Wynne, C.; Petrov, A.; Dinman, J. D.; Edwards, N.; Fenselau, C. *J. Proteome Res.* **2008**, *7*, 579-586.

(24) Slebos, R. J. C.; Brock, J. W. C.; 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. *J. Proteome Res.* **2008**, *7*, 5286-5294.

(25) Alpert, A. J.; Andrews, P. C. *J. Chromatogr., B* **1988**, *443*, 85-96.

(26) Villen, J.; Gygi Steven, P. *Nat. Protoc.* **2008**, *3*, 1630-1638.

(27) Helbig, A. O.; Gauci, S.; Raijmakers, R.; van Breukelen, B.; Slijper, M.; Mohammed, S.; Heck, A. J. R. *Mol. Cell. Proteomics* **2010**, *9*, 928-939.

(28) Boersema, P. J.; Foong, L. Y.; Ding, V. M. Y.; Lemeer, S.; van Breukelen, B.; Philp, R.; Boekhorst, J.; Snel, B.; den Hertog, J.; Choo, A. B. H.; Hecka, A. J. R. *Mol. Cell. Proteomics* **2010**, *9*, 84-99.

(29) Jiang, J.; Parker, C. E.; Hoadley, K. A.; Perou, C. M.; Boysen, G.; Borchers, C. H. *Proteomics: Clin. Appl.* **2007**, *1*, 1651-1659.

(30) Reid, J. D.; Holmes, D. T.; Mason, D. R.; Shah, B.; Borchers, C. H. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1680-1686.

(31) Tang, J.; Liu, Y.; Qi, D.; Yao, G.; Deng, C.; Zhang, X. *Proteomics* **2009**, *9*, 5046-5055.

(32) Wang, W.-H.; Palumbo, A. M.; Tan, Y.-J.; Reid, G. E.; Tepe, J. J.; Bruening, M. L. *J. Proteome Res.* **2010**, *9*, 3005-3015.

(33) Yamashita, M.; Fenn, J. B. *J. Phys. Chem.* **1984**, *88*, 4451-4459.

(34) Dole, M.; Mack, L. L.; Hines, R. L.; Mobley, R. C.; Ferguson, L. D.; Alice, M. B. *J. Chem. Phys.* **1968**, *49*, 2240-2249.

(35) Iribarne, J. V.; Thomson, B. A. *J. Chem. Phys.* **1976**, *64*, 2287-2294.

(36) Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yohida, T. *Rapid Commun. Mass Spectrom.* **1988**, *2*, 151-153.

(37) Karas, M.; Hillenkamp, F. *Anal. Chem.* **1988**, *60*, 2299-2301.

(38) Tegeler, T. J.; Mechref, Y.; Boraas, K.; Reilly, J. P.; Novotny, M. V. *Anal. Chem.* **2004**, *76*, 6698-6706.

(39) Young, J. B.; Li, L. *Anal. Chem.* **2007**, *79*, 5927-5934.

(40) Mirgorodskaya, E.; Braeuer, C.; Fucini, P.; Lehrach, H.; Gobom, J. *Proteomics* **2005**, *5*, 399-408.

(41) Bogdanov, B.; Smith, R. D. *Mass Spectrom. Rev.* **2005**, *24*, 168-200.

(42) Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E. P.; Shabanowitz, J.; Hunt, D. F. *Biochim. Biophys. Acta, Proteins Proteomics* **2006**, *1764*, 1811-1822.

(43) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466-1467.

(44) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.

(45) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976-989.

(46) Shilov, I. V.; Seymour, S. L.; Patel, A. A.; Loboda, A.; Tang, W. H.; Keating, S. P.; Hunter, C. L.; Nuwaysir, L. M.; Schaeffer, D. A. *Mol. Cell. Proteomics* **2007**, *6*, 1638-1655.

(47) Liu, H.; Sadygov, R. G.; Yates, J. R. *Anal. Chem.* **2004**, *76*, 4193-4201.

(48) Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M. *Mol. Cell. Proteomics* **2005**, *4*, 1265-1272.

(49) Lu, P.; Vogel, C.; Wang, R.; Yao, X.; Marcotte, E. M. *Nat. Biotechnol.* **2007**, *25*, 117-124.

(50) Zhang, Y.; Wen, Z.; Washburn Michael, P.; Florens, L. *Anal. Chem.* **2009**, *81*, 6317-6326.

(51) Xia, Q.; Hendrickson, E. L.; Wang, T.; Lamont, R. J.; Leight, J. A.; Hackett, M. *Proteomics* **2007**, *7*, 2904-2919.

(52) Ryu, S.; Gallis, B.; Goo, Y. A.; Shaffer, S. A.; Radulovic, D.; Goodlett, D. R. *Cancer Inf.* **2008**, *4*, 243-255.

(53) Savitski, M. M.; Fischer, F.; Mathieson, T.; Sweetman, G.; Lang, M.; Bantscheff, M. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1668-1679.

(54) Paulovich, A. G.; Billheimer, D.; Ham, A.-J. L.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C. *Mol. Cell. Proteomics* **2010**, *9*, 242-254.

(55) Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A.-J. L.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C. *J. Proteome Res.* **2010**, *9*, 761-776.

(56) Christin, C.; Hoefsloot, H. C. J.; Smilde, A. K.; Suits, F.; Bischoff, R.; Horvatovich, P. L. *J. Proteome Res.* **2010**, *9*, 1483-1495.

(57) Lengqvist, J.; Andrade, J.; Yang, Y.; Alvelius, G.; Lewensohn, R.; Lehtioe, J. *J. Chromatogr., B* **2009**, *877*, 1306-1316.

(58) Hilger, M.; Bonaldi, T.; Gnad, F.; Mann, M. *Mol. Cell. Proteomics* **2009**, *8*, 1908-1920.

(59) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Matthews, D. E.; Yates, J. R. *Anal. Chem.* **2004**, *76*, 4951-4959.

(60) Bunner, A. E.; Williamson, J. R. *Methods* **2009**, *49*, 136-141.

(61) Jiang, H.; English Ann, M. *J. Proteome Res.* **2002**, *1*, 345-350.

(62) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen Dan, B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell. Proteomics* **2002**, *1*, 376-386.

(63) Bicho, C. C.; de Lima Alves, F.; Chen, Z. A.; Rappsilber, J.; Sawin, K. E. *Mol. Cell. Proteomics* **2010**, *9*, 1567-1577.

(64) Bendall, S. C.; Hughes, C.; Stewart, M. H.; Doble, B.; Bhatia, M.; Lajoie, G. A. *Mol. Cell. Proteomics* **2008**, *7*, 1587-1597.

(65) Kruger, M.; Moser, M.; Ussar, S.; Thievessen, I.; Luber Christian, A.; Forner, F.; Schmidt, S.; Zanivan, S.; Fassler, R.; Mann, M. *Cell* **2008**, *134*, 353-364.

(66) Julka, S.; Regnier, F. *J. Proteome Res.* **2004**, *3*, 350-363.

(67) Panchaud, A.; Hansson, J.; Affolter, M.; Bel Rhlid, R.; Piu, S.; Moreillon, P.; Kussmann, M. *Mol. Cell. Proteomics* **2008**, *7*, 800-812.

(68) Panchaud, A.; Guillaume, E.; Affolter, M.; Robert, F.; Moreillon, P.; Kussmann, M. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 1585-1594.

(69) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994-999.

(70) Li, J.; Steen, H.; Gygi, S. P. *Mol. Cell. Proteomics* **2003**, *2*, 1198-1204.

(71) Qian, W.-J.; Goshe, M. B.; Camp, D. G., II; Yu, L.-R.; Tang, K.; Smith, R. D. *Anal. Chem.* **2003**, *75*, 5441-5450.

(72) Heller, M.; Mattou, H.; Menzel, C.; Yao, X. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 704-718.

(73) Fenselau, C.; Yao, X. *J. Proteome Res.* **2009**, *8*, 2140-2143.

(74) Petritis, B. O.; Qian, W.-J.; Camp, D. G., II; Smith, R. D. *J. Proteome Res.* **2009**, *8*, 2157-2163.

(75) Niles, R.; Witkowska, H. E.; Allen, S.; Hall, S. C.; Fisher, S. J.; Hardt, M. *Anal. Chem.* **2009**, *81*, 2804-2809.

(76) Liu, Z.; Cao, J.; He, Y.; Qiao, L.; Xu, C.; Lu, H.; Yang, P. *J. Proteome Res.* **2010**, *9*, 227-236.

(77) Shakey, Q.; Bates, B.; Wu, J. *Anal. Chem.* **2010**, *82*, 7722-7728.

(78) Thompson, A.; Schaefer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. *Anal. Chem.* **2003**, *75*, 1895-1904.

(79) Kuzyk, M. A.; Ohlund, L. B.; Elliott, M. H.; Smith, D.; Qian, H.; Delaney, A.; Hunter, C. L.; Borchers, C. H. *Proteomics* **2009**, *9*, 3328-3340.

(80) Thingholm, T. E.; Palmisano, G.; Kjeldsen, F.; Larsen, M. R. *J. Proteome Res.* **2010**, *9*, 4045-4052.

(81) Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 6940-6945.

(82) Anderson, N. L.; Anderson, N. G.; Haines, L. R.; Hardie, D. B.; Olafson, R. W.; Pearson, T. W. *J. Proteome Res.* **2004**, *3*, 235-244.

(83) Pratt, J. M.; Simpson, D. M.; Doherty, M. K.; Rivers, J.; Gaskell, S. J.; Beynon, R. J. *Nat. Protoc.* **2006**, *1*, 1029-1043.

(84) Rivers, J.; Simpson, D. M.; Robertson, D. H. L.; Gaskell, S. J.; Beynon, R. J. *Mol. Cell. Proteomics* **2007**, *6*, 1416-1427.

(85) Ji, C.; Guo, N.; Li, L. *J. Proteome Res.* **2005**, *4*, 2099-2108.

Chapter 2 - Effect of 2MEGA Labelling on Membrane Proteome Analysis Using LC-ESI-QTOF MS*

2.1 Introduction

Mass spectrometry (MS)-based peptide sequencing is one of the most reliable techniques for the identification of proteins and their post-translational modifications. Recent advancements in instrumentation,[1-4] database searching engines,[5-7] and high performance separation techniques[8-11] have led to an emergence of high-throughput approaches designed to identify thousands of peptides from a variety of biologically complex protein mixtures. For example, the "shotgun proteomics" method,[12] which involves the direct analysis of complex peptide mixtures derived from proteolytic digestion of heterogeneous mixtures of proteins, has been widely used for rapid generation of a global profile of the protein complement within the mixture. In this method, complex protein mixtures are typically digested by the enzyme trypsin to produce extremely complex peptide mixtures. The peptide mixtures are then subjected to extensive separations, such as strong cation exchange (SCX) chromatography, coupled online or offline with reversed-phase (RP) capillary liquid chromatography (LC). Peptides eluting from the RP LC column are commonly analyzed by electrospray ionization (ESI) MS and selected ions are subjected to fragmentation by collision-induced dissociation (CID) to produce MS/MS spectra. Peptide identifications are usually made by comparing the experimental MS/MS spectra with predicted MS/MS spectra generated from a set of possible proteins in a database using an automated database searching algorithm, such as SEQUEST[6] or MASCOT.[7] Finally, a long list of

---

*A version of this chapter was published as: Ji, C.; Lo, A.; Marcus, S.; Li, L. *Journal of Proteome Research* **2006**, *5*, 2567-2576. Dr. C. Ji and Dr. L. Li were responsible for experimental design and manuscript preparation. Dr. S. Marcus cultured the cells and prepared the membrane fraction. Mr. A. Lo performed the desalting using solid-phase extraction, strong cation exchange separation, and assisted with the data analysis and manuscript editing. Dr. C. Ji performed the LC-MS analysis and the majority of the data analysis.

peptides and proteins is reported in terms of a probability score, as is the case for the MASCOT search engine and a recently modified version of SEQUEST.[13]

One of the greatest challenges associated with large-scale proteome analysis using MS/MS and automated database searching is to determine the reliability of these protein hits, i.e., to reduce the number of proteins that are false positives without compromising the number of correct identifications. To reduce false positive identifications, several "rules of thumb" can be used to carry out critical manual evaluation of large-scale proteomic experimental results.[14] However, the extensive knowledge of peptide fragmentation required and time-consuming manual evaluation involved have resulted in the majority of large-scale proteome analysis results published or reported to be without critical manual interpretation. Therefore, any information that can be used to evaluate the reliability of protein identification via an automated method is highly desirable. Recent reports have begun to raise key issues relating to the confidence of these identifications.[15-18] Additional supporting information from a peptide's LC elution time or isoelectric point has been developed to reduce further the false positive rates for peptide identifications.[19, 20]

Chemical modifications of the N-termini and C-termini of peptides have been developed to simplify and direct the fragmentation of peptides to facilitate the interpretation of the obtained MS/MS spectra.[21-28] For example, divinyl sulfone has been used as a post-digestion modifier to enhance the intensity of the signal of the a_1 ion produced in MS/MS and post-source decay.[29] This enhanced signal can be used to fingerprint the N-terminal amino acid of a peptide. This information, which is normally not present in low-energy CID spectra, is advantageous for de novo sequencing and could also be used as a filter to reduce false positive identifications. However, divinyl sulfone may label the N-terminus or certain amino acids (lysine, histidine, cysteine) and produce isomeric products; all these factors complicate CID spectra and hinder the interpretation of peptide sequences. Several studies have reported that dimethyl labelling of amino groups on the N-termini and side chains of lysine residues in peptide sequences also leads to the enhancement of a_1 ions in the corresponding CID spectra.[28, 30-32] The enhanced a_1 signals in the CID spectra have been

demonstrated to provide higher confidence in the identification of proteins performed by either de novo sequencing or database-assisted searching by providing a universal a_1 tag for mapping the N-terminal amino acid through a precursor ion scan with a small set of data.[28] However, the similar masses of a_1 ions derived from N-terminal lysine (157.1705 Da) and arginine (157.1453 Da) residues are indistinguishable when using low-resolution instruments. In addition, the possible multiple labelling of a peptide containing 1−3 lysine residues, which is normally observed in a tryptic digest from a complex protein mixture, complicates data analysis. More recently, Reilly and co-workers reported a novel derivatization strategy that utilizes both guanidinylation and amidination to assist peptide sequencing.[27] A unique characterization of this derivatization is that abundant y_{n-1} and b_1 ions are typically observed in MS/MS spectra.[27] This feature can also be used as a constraint to reduce false positive identifications. However, it was also reported that the N-terminal amidine groups are susceptible to hydrolysis when the N-terminal residue is serine or threonine. The consequence of this side reaction is that y_{n-1} and b_1 ions are not formed by CID.

We have recently demonstrated that differential 2MEGA[33] (dimethylation after guanidinylation) labelling of N-termini of peptides with $^{12}CH_2O$- and $^{13}CD_2O$-formaldehyde, after blocking the amino groups on the side chains of lysines by guanidinylation with *O*-methylisourea, is a promising strategy for global quantitative and qualitative proteome analysis using auto-offline LC-MALDI MS and MS/MS because of the following reasons: (1) the uniform 6.032 Da mass difference between each derivatized peptide pair eliminates the significant overlapping of isotope envelopes, even for a peptide pair of around 3000 Da, and simplifies the quantification data analysis process; (2) the reaction itself is simple, fast, and complete and also can be done with commercially available and inexpensive reagents; (3) the presence of universal a_1 ions in the MALDI MS/MS spectra and the overlaid fragment ion spectra generated from a pair of differentially labelled peptides can be used to confirm peptide sequences obtained from MS/MS database searching, or to carry out de novo sequencing of peptides on the basis of their MS/MS spectra.

In this study, the effect of 2MEGA labelling on the large-scale membrane proteome analysis is evaluated using LC-ESI quadrupole time-of-flight (QTOF) MS. By comparison with the large-scale membrane proteome analysis of a native digest from the same sample, it is demonstrated that 2MEGA labelling not only increases the number of peptides and proteins identified but also provides the enhanced a_1 ions or a_1-related ions as a constraint to reduce the number of false positive identifications.

2.2 Experimental

2.2.1 Chemicals and Reagents

Formaldehyde (37% (w/w) in H_2O), *O*-methylisourea, sodium hydroxide, sodium bicarbonate, sodium cyanoborohydride, bovine trypsin, trifluoroacetic acid (TFA), and Leucine enkephalin (Leu-enk) were purchased from Sigma-Aldrich (Oakville, ON, Canada). HPLC grade acetonitrile was purchased from Fisher Scientific Canada (Edmonton, AB, Canada). Water used in these experiments was obtained from a Milli-Q Plus purification system (Millipore, Bedford, MA). Formic acid and bovine gamma globulin were purchased from Pierce (Rockford, IL). The other chemicals were from Sigma (St. Louis, MO) and were analytical grade.

2.2.2 Cell Culture and Membrane Preparation

Escherichia coli K-12 (*E. coli*, ATCC 47076) was from the American Type Culture Collection. A single *E. coli* K12 colony was used to inoculate 10 mL of LB broth (BBL, Becton Dickinson). The culture was incubated overnight with shaking at 37 °C. This saturated culture (1.5 mL) was added to 90 mL of growth medium in a 500-mL baffled Erlenmeyer flask. Cells were harvested in the mid-log phase by centrifugation at 3200 *g* for 10 min at 4 °C, resuspended, washed in 50 mM MOPS buffer, pH 7.3, and collected by centrifugation at 3200 *g* for 10 min at 4 °C. A 7-mL aliquot of the *E. coli* cell suspension was thawed in cold water, and the volume was brought to 15 mL with 50 mM MOPS buffer, pH 7.3. Then 1.4 mg of DNaseI was added. The suspension was passed twice through a French press (Aminco Rochester, NY) using rapid fill kit at 14000 psi. The final volume was about 20 mL after adding more 50 mM MOPS (pH 7.3) to rinse the tube. The lysate was centrifuged in a Beckman SX4250 rotor at 4500 rpm

(about 2300 *g*) for 10 min to pellet unbroken cells. The supernatant was collected, and the protein concentration was estimated by performing a BioRad protein assay using bovine gamma globulin as the standard. The membrane proteins were isolated using a carbonate fractionation procedure[34, 35] with some modifications. About 2 mL of lysate (containing approximately 20 mg of cellular proteins) was added to 10 mL of ice-cold MOPS buffer. Then, in a 250 mL beaker, 110 mL of 0.1 M sodium carbonate (pH 11.0) was slowly added. The solution was stirred slowly in an ice bath for 1 h to extract membranes. The extract was divided equally into two tubes, filled with about 5 mL more 0.1 M sodium carbonate each, and centrifuged in a Beckman Type 45Ti rotor for 65 min at 38400 rpm (115000 g_{av}). The supernatant was aspirated and the pellet was gently rinsed with 5 mL of water. Each pellet was suspended in 2 mL of 50 mM MOPS buffer (pH 7.3) and transferred to an 8-mL tube. About 5 mL more buffer was added to each tube to bring the volume to 7 mL. The tubes were centrifuged in a Beckman Type 70.1Ti rotor at 40 000 rpm (115 000 g_{av}) for 25 min.

2.2.3 Protein Digestion

To achieve maximum digestion efficiency, two consecutive digestion steps were performed in this study using a combination of organic-assisted[34] and SDS-assisted[36] solubilization and proteolysis. First, proteins in the membrane fraction were resuspended in 50 mM ammonium bicarbonate, pH 8.0, via intermittent vortexing and sonication using a sonicating bath (Branson model 1510, Danbury, CT). The proteins were thermally denatured by incubating the sample in airtight tubes at 90 °C for 20 min and then cooled in ice-cold water. The membrane protein concentration was estimated by a BioRad protein assay using bovine gamma globulin as the standard. About 1 mg of protein from the membrane fraction was then diluted with methanol to produce a composition of 60% organic solvent, resulting in a final protein concentration of 1 mg/mL. Tryptic digestion was immediately carried out by adding 20 µg of 1µg/µL trypsin and incubation at 37 °C for 5 h. Second, after methanol-assisted digestion, undissolved sample was pelleted and resuspended in 400 µL of 0.05% SDS with the addition of 15 µg trypsin. The sample was incubated at 37 °C overnight. Methanol in the supernatant from the methanol-assisted digestion was evaporated by SpeedVac. The solution from the SDS-assisted digestion of the leftover sample was

pooled with that from the first digestion. The digestion solution was stored at −80 °C. We note that we did not reduce and alkylate proteins prior to trypsin digestion. Theoretically, it would reduce the efficiency of the trypsin digestion. This could result in two opposite effects. Although this results in a smaller number of peptides being generated from each protein for protein identification, the peptide mixture is less complex for analysis.

2.2.4 2MEGA Labelling

2MEGA labelling of half of the pooled tryptic digest from methanol-assisted and SDS-assisted digestion was carried out as reported previously.[33] In brief, guanidinylation of lysine residues was performed as described previously[37-40] with some modifications. Trypsin in the 500-µL tryptic digest solution (about 1 µg/µL) was irreversibly inactivated by adding 50 µL of 2 M sodium hydroxide. The ε-amino groups of all lysines were blocked by adding 200 µL of 2 M *O*-methylisourea in 100 mM $NaHCO_3$, adjusting to pH 11 with 2 M sodium hydroxide and incubating the resulting mixture at 65 °C for 10 min. Then the reaction was stopped and the pH was adjusted to 8 by adding 10% TFA. Reductive methylation with $^{12}CH_2O$-formaldehyde was also carried out as described previously[28, 30-32] with some modifications. The above guanidinylated peptide solution was mixed with 30 µL of 2 M sodium cyanoborohydride. The mixture was then vortexed and mixed with $^{12}CH_2O$-formaldehyde (4% (w/w) in water, 6 µL). The mixture was vortexed and incubated at 37 °C for 1 h. If necessary, ammonium bicarbonate (1 M, 6 µL) was added to consume the excess formaldehyde. After labelling, the pH of the solution was adjusted to 2.5 using 10% TFA. Caution: sodium cyanoborohydride is a highly toxic compound that releases hydrogen cyanide gas upon exposure to strong acid, and formaldehyde is known to have carcinogenic effects, including cancer risk from inhalation exposure. Therefore, the 2MEGA labelling process must be performed in a fume hood.

2.2.5 Desalting Using Solid-Phase Extraction Cartridge

The 2MEGA-labelled peptide mixture solutions and the unlabelled half of the pooled tryptic digests from the methanol-assisted and SDS-assisted digestions were desalted by SPE using bonded phase octadecyl (C_{18}) cartridges. Each cartridge was

equilibrated with three bed volumes of acetonitrile and washed with three volumes of 0.1% TFA. The peptide mixture was applied to the cartridge and the cartridge was washed with three volumes of aqueous 0.1% TFA. Finally, the peptides were eluted, initially with 500 µL of acetonitrile/H_2O/TFA (50:49.9:0.1, v/v/v) and then with 1 mL of acetonitrile/H_2O/TFA (75:24.9:0.1, v/v/v). The eluate was concentrated to 300 µL by using a SpeedVac. The peptide mixture was stored at −20 °C.

2.2.6 Cation Exchange Chromatography

The desalted peptide mixture was separated by strong cation exchange (SCX) chromatography on an Agilent 1100 HPLC system (Palo Alto, CA) using a 2.1 × 150 mm Hydrocell SP 1500 column (5 µm, Catalog No.: 24-34 SP, BioChrom Labs, Inc., Terre Haute, IN). The solvent solutions used were 20% v/v acetonitrile in 0.1% TFA (solvent A) and 20% v/v acetonitrile in 0.1% TFA, 1 M NaCl (solvent B). About 400 µL (500 µg) of protein digest was loaded onto the SCX column, and peptides were eluted with linear gradients of 0−10% B in 2 min, 10−30% B in 10 min, and 30−50% B in 2 min at 0.25 mL/min, with collection of 1 min fractions. In total, 8 fractions were collected based on the chromatography signal recorded at 214 nm. The first two fractions were pooled and the last three fractions were pooled into another fraction because of their low UV absorbance signals. Finally, five fractions were obtained, and each was concentrated to 10 µL by using a SpeedVac. For the sample containing peptides pooled from the last three fractions, only the supernatant was analyzed by LC-ESI-QTOF MS. Salts at the bottom of the vial were discarded.

2.2.7 LC-ESI-QTOF Mass Spectrometric Analysis

The peptides in each SCX fraction were analyzed using a QTOF Premier Mass Spectrometer (Waters, Manchester, UK) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, 2 µL of peptide solution from each SCX fraction was injected onto a 75 µm × 100 mm Atlantis dC_{18} column (Waters, Milford, MA). Solvent A consisted of 0.1% formic acid in water, and solvent B consisted of 0.1% formic acid in acetonitrile. Peptides were separated using gradients of 5−30% solvent B in 80 min, 30−90% solvent B in 10 min, and 90−5% solvent B in 10 min and electrosprayed into the QTOF mass spectrometer, fitted with a nanoLockSpray source,

at a flow rate of 250 nL/min. Mass spectra were acquired from *m*/*z* 300 to 1600 for 1 s followed by 3 data-dependent MS/MS scans from *m*/*z* 50 to 1900 for 1 s each. The collision energy used to perform MS/MS was varied according to the mass and charge state of the eluting peptide. Leu-enk, a mass calibrant, or Lock-mass was infused at a rate of 250 nL/min and was acquired for 1 s every 2 min throughout the run. The exclusion list was generated on the basis of MASCOT searching results in which peptides with a score above the identity threshold were selected.

2.2.8 Protein Identification from MS/MS Data

Raw search data was lock-mass corrected, de-isotoped, and converted to peak list files by ProteinLynx Global Server 2.1.5 (Waters). Peptide sequences were identified via automated database searching of peak list files using the MASCOT search program (Matrix Science, London, United Kingdom). Database searching was restricted to *Escherichia coli* in Swiss-Prot database (UniProtKB/Swiss-Prot Release 47.7 of 16-Aug-2005). The following search parameters were selected for all database searching: enzyme, trypsin; missed cleavages, 3; peptide tolerance, ±30 ppm; MS/MS tolerance: ±0.2 Da; peptide charge, (1+, 2+, and 3+); variable modification, oxidation (M). In all cases, peak list files were searched twice, in one case with the instrument setting as ESI-QUAD-TOF, the other being constrained to a modified ESI-QUAD-TOF setting, in which a_1 ions and immonium ions were added as possible fragment ions. For the database search of MS/MS data generated from unlabelled pooled tryptic digests from methanol-assisted and SDS-assisted digestion, no additional fixed modifications were selected. However, for database searching of MS/MS data generated from 2MEGA-labelled tryptic peptides from pooled tryptic digests from methanol-assisted and subsequent SDS-assisted digestion, the following modifications were selected as fixed modifications: guanidinylation (K) and dimethylation-L (N-term). Where peptides matched more than one database entry due to redundant protein sequence submissions, assignments to the duplicated sequence were removed.

2.2.9 Hydropathy Calculation

All identified proteins were analyzed using the ProtParam tool (available at http://ca.expasy.org/tools/protparam.html), which allows the calculation of the grand average of hydropathy (GRAVY) value for a given protein.[41] As the hydropathy of the entire proteins gives a general measurement of protein hydrophobicity, The proteins exhibiting positive GRAVY values were recognized as hydrophobic.

2. 3 Results and Discussion

The membrane fraction of *E. coli* cell extract was chosen as the study model for two reasons. First, integral membrane proteins that are inserted into phospholipid bilayers are important biological and pharmacological targets. Second, qualitative and quantitative large scale proteome analysis of integral membrane proteins remains a challenge. This work focuses on investigating the effect of 2MEGA labelling on the large-scale proteome analysis of membrane proteins to address issues related to proteome coverage and the reliability of protein identification. This present work should pave the way for future quantitative analysis of membrane proteomes using 2MEGA isotopic labeling (Figure 2.1). To evaluate the effect of 2MEGA labelling on membrane proteome analysis, large-scale LC MS/MS datasets for native and 2MEGA-labelled tryptic digests from methanol-assisted and subsequent SDS-assisted solubilization and digestion were generated. The work flow for proteome analysis is shown in Figure 2.1. After proteins in the *E. coli* membrane fraction were digested with trypsin using methanol-assisted and subsequent SDS-assisted solubilization (Section 2.2: Experimental), half of the digest was labelled using the 2MEGA labelling strategy.[36] The native and 2MEGA-labelled digests then underwent SPE desalting, SCX separation, RP-LC MS/MS analysis, and database searching (Figure 2.2).

2.3.1 Fragmentation of 2MEGA-Labelled Peptides Produced by ESI

A previous study[33] demonstrated that a_1 ions are greatly enhanced after 2MEGA labelling in MALDI MS/MS analysis. In this study, the effect of 2MEGA labelling on the fragmentation of peptide ions produced by ESI on a large scale is studied. For

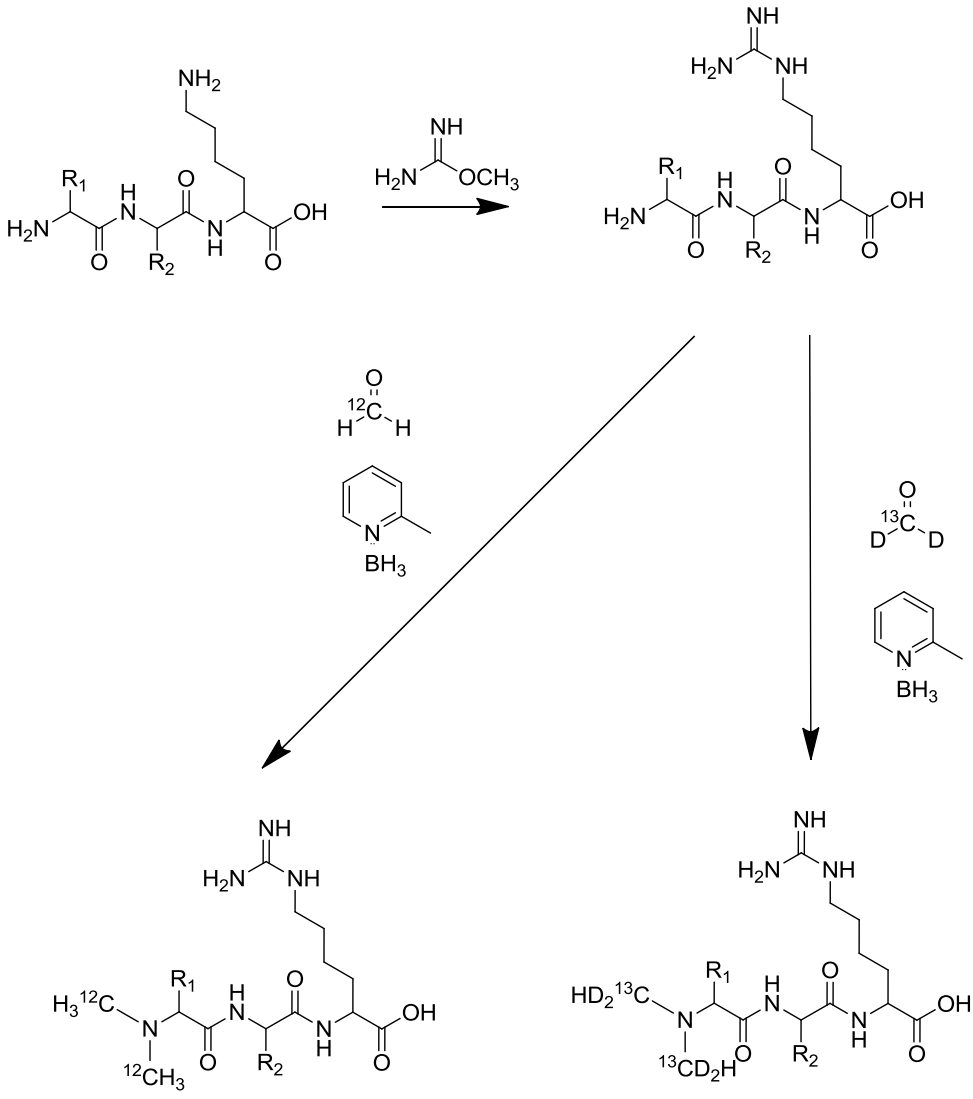


Figure 2.1 2MEGA Reaction Scheme. For the membrane proteome profiling work, only $^{12}CH_2O$ was required. For quantitative applications, both $^{12}CH_2O$ and $^{13}CD_2O$ are used.

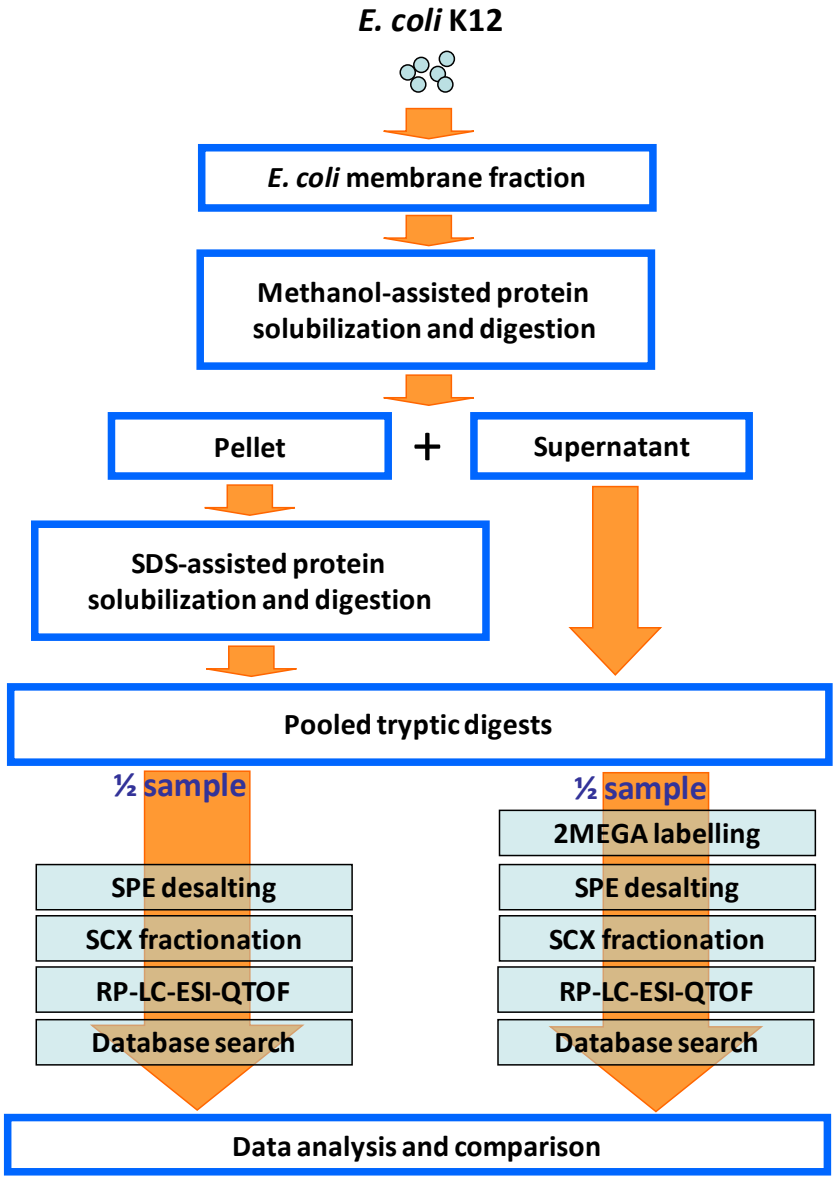


Figure 2.2 Experimental workflow for comparison of unlabelled to 2MEGA labelled samples

this purpose, both unlabelled and 2MEGA-labelled tryptic peptides were analyzed by 2-dimensional (2D) LC-ESI MS/MS. In total, 1104 unlabelled and 1484 2MEGA-labelled peptides were identified with MASCOT scores above the identity threshold.

Examination of several hundred ESI MS/MS spectra of unlabelled and 2MEGA-labelled peptide pairs showed that a_1 ion peaks were significantly enhanced in the ESI MS/MS spectra of 2MEGA-labelled tryptic peptides. Figure 2.3 displays a pair of representative tandem mass spectra of labelled and unlabelled SDVLFNFNK. The a_1 ion peak (60.04 Da) is absent in the spectrum of unlabelled SDVLFNFNK (Figure 2.3A), whereas the labelled a_1 ion peak (88.07 Da) is clearly present in the MS/MS spectrum of the 2MEGA-labelled SDVLFNFNK (Figure 2.3B). In addition, in this case, the whole peptide sequence can be easily deduced from the MS/MS spectrum of the 2MEGA-labelled peptide (Figure 2.3B). However, it should be noted that full sequence information can be deduced only if the spectrum is of good quality, such as having reasonably high signal-to-noise ratios for all major fragment ions.

One advantage of this 2MEGA labelling strategy over the previously reported dimethyl labelling strategy[28] for *de novo* peptide sequencing is that all N-terminal amino acids are easily distinguished, except L (leucine) and I (isoleucine), whose masses are identical. In addition to L and I, the previously reported dimethyl labelling strategy cannot distinguish between R and K (arginine and lysine), whose small mass difference (0.025 Da) makes them difficult to resolve, even when using reasonably high mass accuracy instruments, such as QTOF. In almost all cases, except when the N-terminal amino acid of a peptide is G, K, or R, the a_1 ion peak is the strongest peak in the low mass region of the MS/MS spectra of the 2MEGA-labelled peptides. Interestingly, instead of observing enhanced a_1 ion peaks, it was found that a_1-45 or a_1-17 peaks are enhanced in ESI MS/MS spectra of the 2MEGA-labelled tryptic peptides with K or R as the N-terminal amino acid (Figure 2.4), whereas a_1 ion peaks are often absent or very weak. The observed a_1-17 can be rationalized by the neutral loss of ammonia from the side chain of 2MEGA-labelled homoarginine or arginine. The tendency to form a_1-45 in the tandem spectrum of the 2MEGA-labelled peptides with N-terminal K or R arises from the neutral loss of $(CH_3)_2NH$. This may be due to the fact

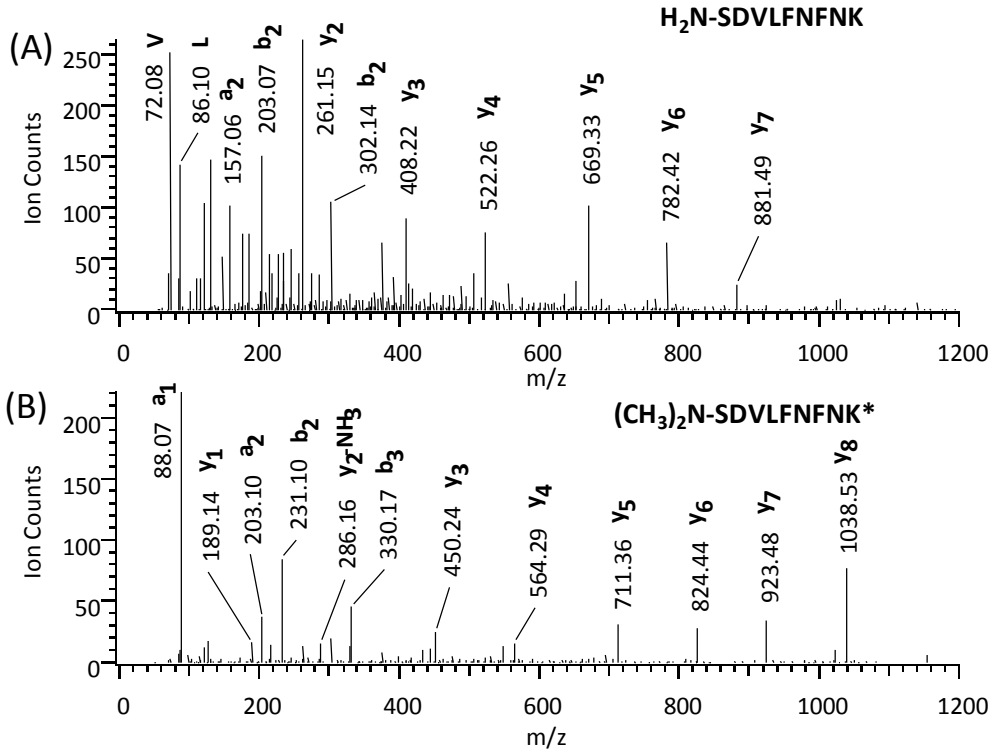


Figure 2.3 MS/MS spectra of a) unlabelled and b) 2MEGA labelled peptide SDVLFNFNK. The asterisk denotes guanidinylation at the lysine residue.

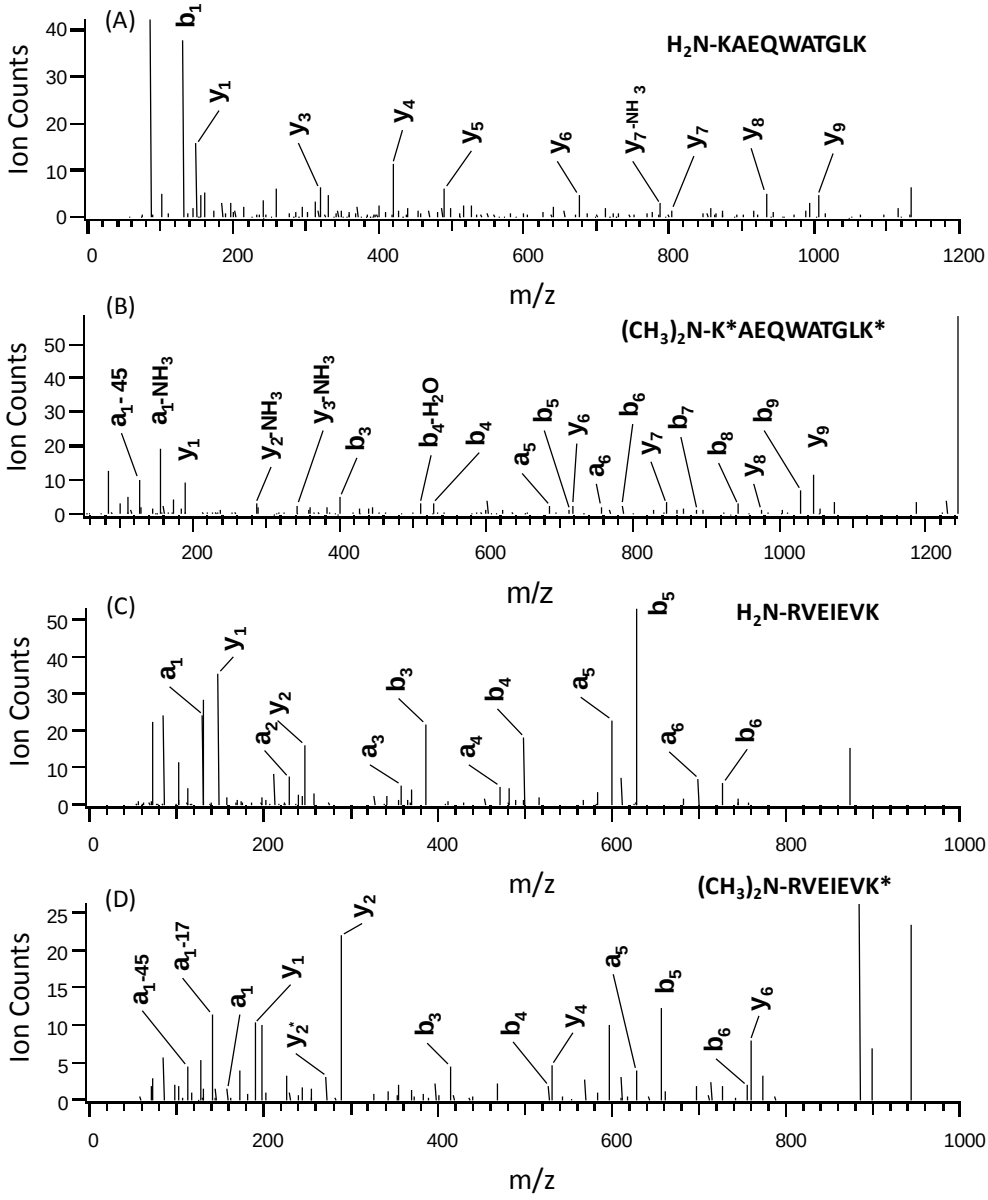


Figure 2.4 MS/MS spectra of a) unlabelled and b) 2MEGA labelled KAEQWATGLK and c) unlabelled and d) 2-MEGA labelled RVEIEVK. The asterisk denotes guanidinylation at the lysine residue.

that the originally formed a_1 positive ions (see Figure 2.5) are quickly attacked by one of the lone pairs of electrons on any of three nitrogen atoms on the side chain of the 2MEGA-labelled homoarginine or arginine to form either a five or seven-member ring for N-terminal R peptides, or either a six or eight-member ring for N-terminal K peptides. For 61 identified peptides with N-terminal K and 48 identified peptides with N-terminal R after 2MEGA labelling, at least one of the three ions (a_1, a_1-17 and a_1-45) was observed in the corresponding MS/MS spectrum. In most cases, two or all three ions were observed.

Table 2.1 lists the summary of the theoretical masses of a_1 or a_1-related ions from the twenty amino acids that are commonly observed in the ESI MS/MS spectra of the 2MEGA-labelled peptides. After examining 1484 MS/MS spectra of identified peptides from the analysis of a 2MEGA-labelled tryptic peptide mixture, using Table 2.1 as the reference mass table, a_1 or a_1-related ion peaks were observed in 1460 of them. It was found that a_1 ions (58.07 Da) were not observed in 11 tandem spectra of identified peptides with N-terminal G (glycine). This is not surprising because the a_1 ion peak has a relatively low intensity in the low mass range of the MS/MS spectra of the 2MEGA-labelled peptides with N-terminal G in which a_1 was observed. Therefore, a correlation between a weak a_1 signal and a peptide with N-terminal G can still be used to confirm the data search result. Not counting the identified peptides with N-terminal G, 1395 out of 1408 (99.08%) identified peptides with scores above the MASCOT identity threshold have a_1 or a_1-related ions in their MS/MS spectra after 2MEGA labelling. Only 13 of 1408 (0.92%) identified peptides with scores above the MASCOT identity threshold that did not have a_1 or a_1-related ions in their MS/MS spectra after 2MEGA labelling were discarded as false positive identifications. The low false positive identification rate, calculated on the basis of manually checking the tandem mass spectra using the a_1 or a_1-related ion table (Table 2.1), is consistent with the results reported by Balgley and co-workers.[16] Therefore, a_1 or a_1-related ions can be used as additional information to eliminate false positive identifications for large-scale proteome analysis.

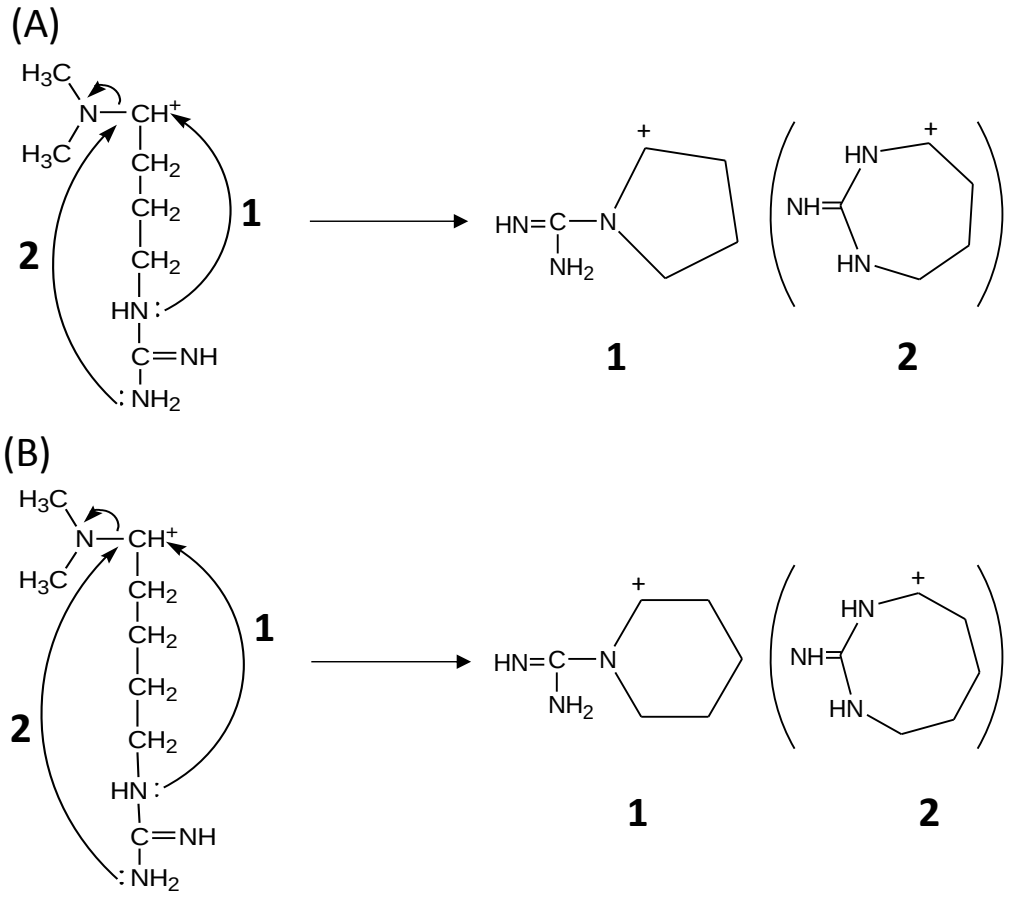


Figure 2.5 Proposed mechanism for the formation of a_1-related ions

Table 2.1. Theoretical masses of the a_1, a_1-NH_3 and a_1-$HN(CH_3)_2$ ions derived from the twenty amino acid residues after 2MEGA labelling

	Theoretical a_1 mass (Da)	Theoretical a_1-NH_3 mass** (Da)	Theoretical a_1-$HN(CH_3)_2$ mass*** (Da)
N-terminal amino acid residues		2MEGA labelled	
Alanine (A)	72.081		
Arginine (R)	157.145	140.118	112.087
Asparagine (N)	115.037		
Aspartic acid (D)	116.071		
Cysteine (C)*	161.087		
Glutamic acid (E)	130.087		
Glutamine (Q)	129.103		
Glycine (G)	58.066		
Histidine (H)	138.103		
Isoleucine (I)	114.128		
Leucine (L)	114.128		
Lysine (K)	171.161	154.134	126.103
Methionine (M)	132.085		
Phenylalanine (F)	148.113		
Proline (P)	84.081		
Serine (S)	88.076		
Threonine (T)	102.092		
Tryptophan (W)	187.124		
Tyrosine (Y)	164.108		
Valine (V)	100.113		

* Side chain of cysteine was blocked by iodoacetamide.

** a_1-NH_3 ion peaks only observed for peptides with N-terminal K or R.

*** a_1-$HN(CH_3)_2$ ion peaks only observed for peptides with N-terminal K or R.

Figure 2.6 shows an example of using a_1 or a_1-related ions as a criterion to eliminate the false positive identifications. An MS/MS spectrum (Figure 2.6A) was searched against the database using MASCOT. Peptide sequence NYQQSYAFVEK was identified as the only significant match with a score well above the identity threshold (Figure 2.6B). Most of the predicted fragment ion peaks of the identified peptide are matched well with those in the experimental MS/MS spectrum (Figure 2.6A). After manually checking the spectrum using the theoretical masses of a_1 or a_1-related ions, the a_1 ion peak with *m*/*z* close to 115.037, corresponding to the N-terminal N (asparagine), is not observed. However, an intense peak at *m*/*z* 114.13 strongly suggests that the potential true peptide should be the one with N-terminal L/I (leucine/isoleucine), which has a theoretical value of 114.128 for the a_1 ion. Therefore, the first match provided by the MASCOT searching result is most likely a false positive identification. Interestingly, in this case, the second matched peptide (IECPYGPLVEEK) with N-terminal I could be the correct match even though its calculated score was very low. However, because the overall matching score was low, the second match was not considered as the correct identification.

It should be noted that we did not consider N-terminal pyroglutamic acid as a possible modification in the database search. Pyroglutamic acid is not reductively methylated because the free amine group on the N-terminus of the peptide forms the amide bond with side chain of glutamic acid to form a pyrrolidone. N-terminal pyroglutamic acid does not result in the presence of a_1 ions in the MS/MS spectra. We have examined the MS/MS spectra that resulted in significant peptide matches and found none of the a_1 ions belongs to this category. Finally, because no reduction and alkylation reactions were carried out on the samples, the use of methylated cysteine as a modification in database search did not lead a significant increase in the number of peptides identified.

2.3.2 Effect of Instrument Settings on Database Searching

Immonium ion peaks and a-series ion peaks are often observed in the MS/MS spectra of labelled and unlabelled tryptic peptides, generated by LC-ESI-QTOF MS. In

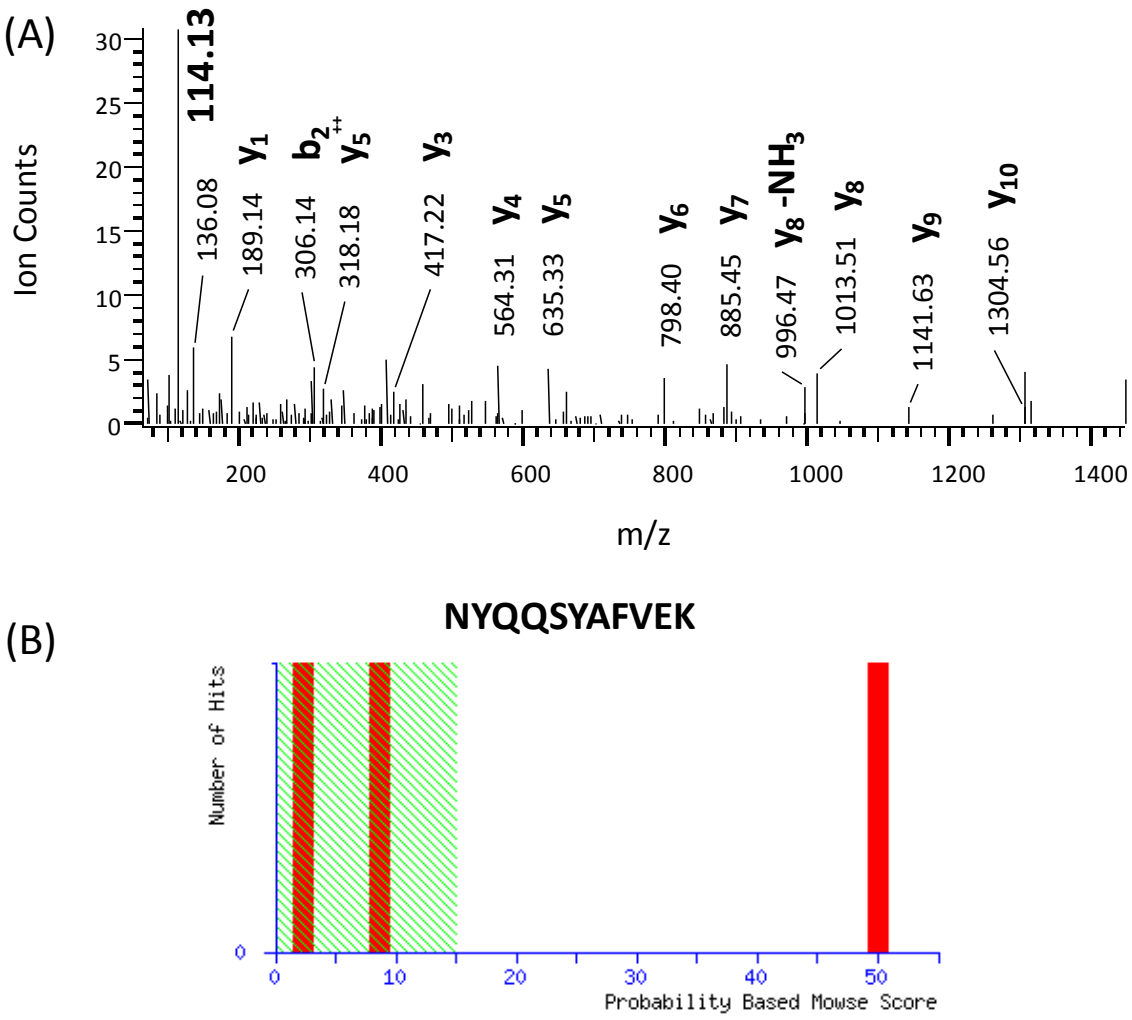


Figure 2.6 a) MS/MS spectrum assigned to NYQQSYAFVEK and the MASCOT score histogram for the spectrum

particular, a_1 ion peaks are enhanced after 2MEGA labelling. However, immonium and a-series ions are not considered as fragment ions in the default ESI-QTOF instrument setting in the MASCOT search parameters. To test the effect of including immonium and a_1 ions as search conditions, those ions were added as possible fragment ions for parent peptides when an ESI-QTOF instrument was used to generate CID spectra and generated an end-user defined ESI-QTOF instrument. For the purpose of comparison, both the raw spectral data of labelled and unlabelled tryptic peptides were searched against the database using MASCOT twice. The first search was constrained to the default ESI-QTOF defined in MASCOT software, whereas the second search was constrained to the end-user defined ESI-QTOF. It was generally found that, after counting immonium ions and a-series ions, MASCOT scores for the identified unlabelled and labelled peptides either remain the same or increase, whereas the identity threshold remains unchanged. Only two exceptions with very minor score decreases (in those cases, scores decreased by 1 and 3) were observed for unlabelled peptides. The score increase for labelled and unlabelled peptides ranges anywhere from 0 to 25, with an average increase of 7.10 for labelled peptides and 6.74 for unlabelled peptides.

Although counting immonium ions and a-series ions did not greatly increase the average MASCOT scores for the labelled or unlabelled peptides, it did lead to more positive identifications. An additional 70 and 113 unique peptides that were initially scored below threshold had their scores increase above threshold for the unlabelled samples and labelled samples, respectively. This represents a percentage increase of 6.76% for the unlabelled samples and 8.23% for the labelled samples. The CID spectra of 113 newly identified unique peptides for the labelled samples were manually checked using the presence of a_1 or a_1-related ions as the criteria to eliminate the false positive identifications. Of the 113 new peptides, only 2 were discarded as false positive identifications because of the absence of a_1 or a_1-related ions. Therefore, in this study, all the reported identified protein numbers and scores are based on database searching using a custom modified ESI-QTOF instrument, unless otherwise noted.

2.3.3. Effect of 2MEGA Labelling on Proteome Analysis

Table 2.2 shows a summary of the comparison for all peptides identified by the unlabelled and 2MEGA-labelled experiments. There is a significant increase (33.2%) in the total number of peptides identified in the labelled sample (1471 peptides) versus the unlabelled sample (1104 peptides) and a dramatic increase (85.3%) in the total number of peptides with C-terminal K identified in the labelled sample (645 peptides) versus the unlabelled sample (348 peptides). Both trends can be rationalized by the increased basicity of peptides after guanidinylation, which selectively converts the amino group on the lysine side chain into a guanidino moiety (identical to the functional group on the arginine side chain). After this conversion, the basicity of homoarginine residues becomes similar to that of arginine. As a consequence, the ionization efficiency of peptides with C-terminal K after guanidinylation may be increased and the chromatographic retention characteristics may be improved (i.e., more C-terminal K peptides may be retained on the C_{18} column). This explanation also supports the dramatic increase (101.9%) in the total number of peptides containing K but no R that were identified in the labelled sample (529 peptides) versus the unlabelled sample (262 peptides). These observations from LC-ESI MS/MS are consistent with previous reports that guanidinylation beneficially increases detection of lysine-terminal peptides in tryptic digest mixtures in MALDI analysis.[42] In addition, there is a relatively small increase (10.4%) in the total number of peptides identified in the labelled sample (814 peptides) versus the unlabelled sample (737 peptides) and a larger increase (31.2%) in the total number of peptides containing R but no K that were identified in the labelled sample (610 peptides) versus the unlabelled sample (465 peptides). One possible explanation is that two extra methyl groups added to the N-termini peptides can significantly alter the ESI response by increasing the gas phase basicity of the N-terminal group.[43]

2.3.4 Identification of Membrane Proteins in an E. coli Membrane Fraction

To maximize the number of unique peptides that can be identified by RP-LC MS/MS, a second injection of the same SCX fraction was carried out after the first

Table 2.2 Classification of the identified peptides from the labelled and unlabelled samples according to their terminal amino acids.

Peptide	Unlabelled		2MEGA labelled	
N-terminus starts with	no. of peptides	Percentage	no. of peptides	Percentage
A	102	9.2	140	9.5
C	0	0	0	0.0
D	63	5.7	97	6.6
E	53	4.8	85	5.8
F	56	5.1	74	5.0
G	79	7.2	76	5.2
H	37	3.4	45	3.1
I	72	6.5	115	7.8
K	50	4.5	61	4.1
L	105	9.5	163	11.1
M	50	4.5	57	3.9
N	42	3.8	60	4.1
P	1	0.1	1	0.1
Q	41	3.7	52	3.5
R	52	4.7	48	3.3
S	76	6.9	85	5.8
T	63	5.7	83	5.6
V	71	6.4	116	7.9
W	24	2.2	41	2.8
Y	67	6.1	72	4.9
Total No. of Peptides	1104		1471	
C-terminal end with K	348	31.5	645	43.8
C-terminal end with R	737	66.8	814	55.3
containing K but no R	262	23.7	529	36.0
containing R but no K	465	42.1	610	41.5

injection. The second RP-LC MS/MS run using the exclusion list of peptides identified in the first run resulted in the identification of an average of 20% more unique peptides for all SCX fractions. For one selected SCX fraction, a third run was done, only resulting in the identification of about 6% additional unique peptides. Therefore, to save instrument time, all other SCX fractions were run twice with the use of the exclusion list for the second run in RP-LC MS/MS.

Figure 2.7 shows the number of peptides that were used to identify proteins in 2D-LC QTOF analysis of labelled and unlabelled samples. In total, 1471 unique peptides corresponding to 498 unique proteins were unambiguously identified from 2MEGA-labelled tryptic peptides of proteins from the membrane fraction of the *E. coli* cell extract, of which 275 proteins (55.2%) were identified on the basis of two or more peptides (Figure 2.7A). From the unlabelled tryptic peptides, 1104 unique peptides, corresponding to 410 unique proteins, were identified from the same membrane fraction, of which 219 proteins (53.4%) were identified on the basis of two or more peptides (Figure 2.7B). We did not see an increase of multi-peptide identifications. The reason is unknown but may be due to the fact that we did not reduce and alkylate proteins prior to trypsin digestion, which would result in less peptides generated from each protein for the LC MS/MS analysis. Figure 2.7C illustrates the overlap in proteins identified from unlabelled and 2MEGA-labelled samples. Out of 640 proteins identified, 268 proteins are common to both the labelled and unlabelled experiments. These complementary results indicate that more comprehensive proteome coverage may be achieved by analyzing both the unlabelled and 2MEGA-labelled samples using the same 2D-LC MS/MS conditions.

As Figure 2.7C shows, 498 out of the 640 identified proteins (77.8%) could be identified by analyzing the labelled sample alone. Although more proteins (498 vs. 410 or 21.5% more) were identified in the labelled sample versus unlabelled sample, about one-quarter of the identified proteins were not detected in the labelled sample. This is understandable considering that the membrane fraction digest was a complicated peptide mixture and, despite the use of 2D-LC, co-elution of peptides in RP-LC was unavoidable, causing an ion suppression effect in ESI MS/MS. Future work will focus on

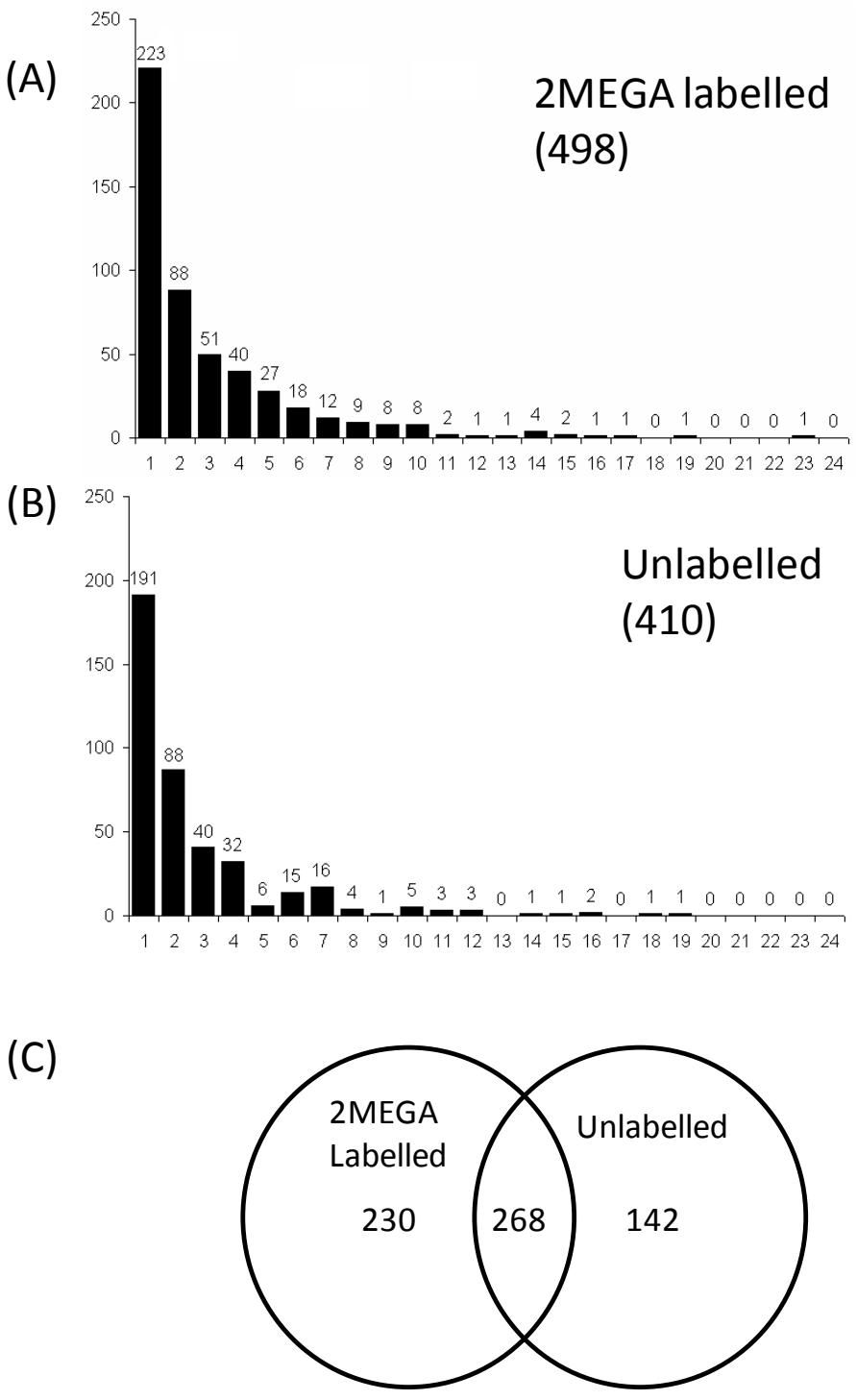


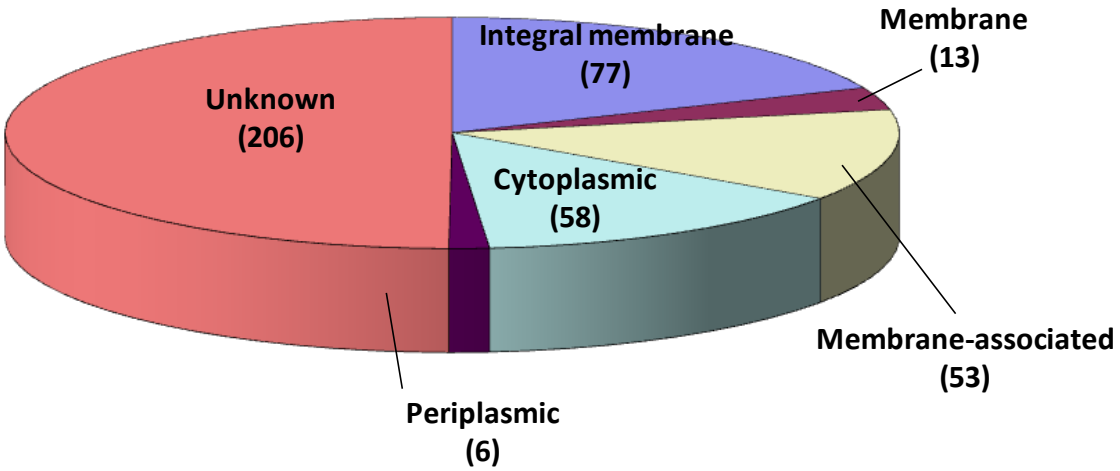
Figure 2.7 Number of identified peptides per proteins in a) 2MEGA labelled samples and b) unlabelled samples. The overlap between identified proteins is shown in c).

improving peptide separation in analyzing the labelled sample, which should increase the number of proteins identified. This is important because the labelled peptides can potentially be quantified when differential labelling is applied. The subcellular locations of the identified proteins in the labelled and unlabelled samples, as indicated in the Swiss-Prot database, were investigated further (Figure 2.8). Many of the identified proteins could not be classified (i.e., 206 out of 410 from the unlabelled sample and 212 out of 498 from the labelled sample). Among the classified proteins, there is a dramatic increase in the total number of integral membrane proteins identified in the 2MEGA-labelled sample (153 proteins) versus the unlabelled sample (77 proteins). There are also significant increases in the number and percentage of membrane and membrane-associated proteins identified in the 2MEGA-labelled sample (243 and 48.8%, respectively) when compared to the unlabelled sample (143 and 34.9%). These results demonstrate that this labelling strategy is an efficient way to identify membrane or membrane-associated proteins. Overall, 258 out of 336 classified proteins (76.8%) or 640 total identified proteins (40.3%) in this study are membrane or membrane-associated proteins. Positive GRAVY values have been considered a reliable marker for indicating the hydrophobicity of a protein and a valid indicator of its membrane involvement.[41, 44-46] Figure 2.9 shows the distribution of the number of proteins identified based on their calculated GRAVY values. Of 498 proteins identified in the 2MEGA-labelled sample, 137 (27.5%) are hydrophobic with positive GRAVY values ranging from +0.001 to +1.271, whereas 68 of 410 (16.6%) proteins identified in the labelled sample are hydrophobic with positive GRAVY values ranging from +0.001 to +1.121. These results further support the above statement that the 2MEGA-labelling strategy is an efficient way to identify membrane or membrane-associated proteins. Overall, 153 out of 640 (23.9%) proteins identified in this study are hydrophobic with positive GRAVY values.

2.4 Conclusions

The effect of 2MEGA labelling was evaluated on a large-scale proteome analysis of a membrane fraction of an *E. coli* cell extract by a shotgun proteomic

(A) Unlabelled (410)



(B) 2MEGA Labelled (498)

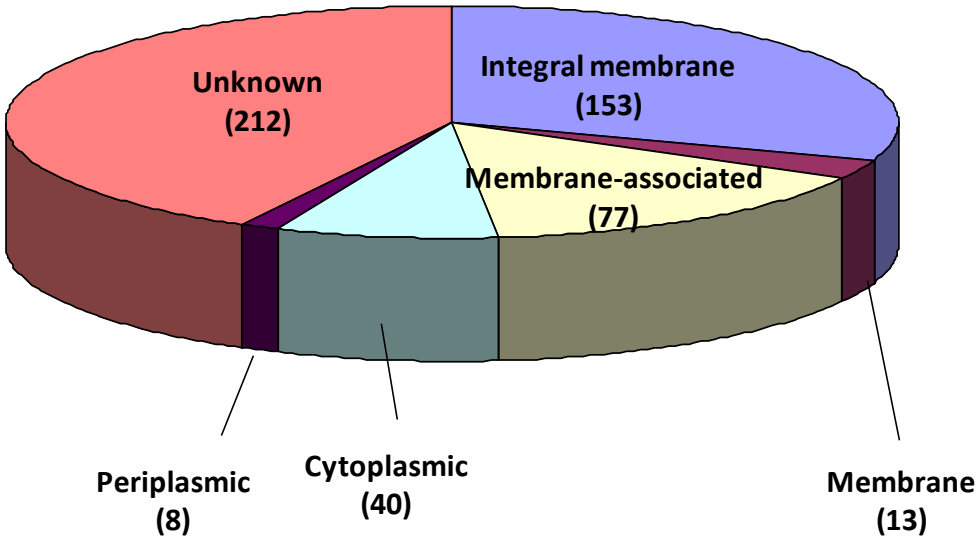


Figure 2.8 Subcellular localization of identified proteins from the a) labelled and b) 2MEGA labelled datasets

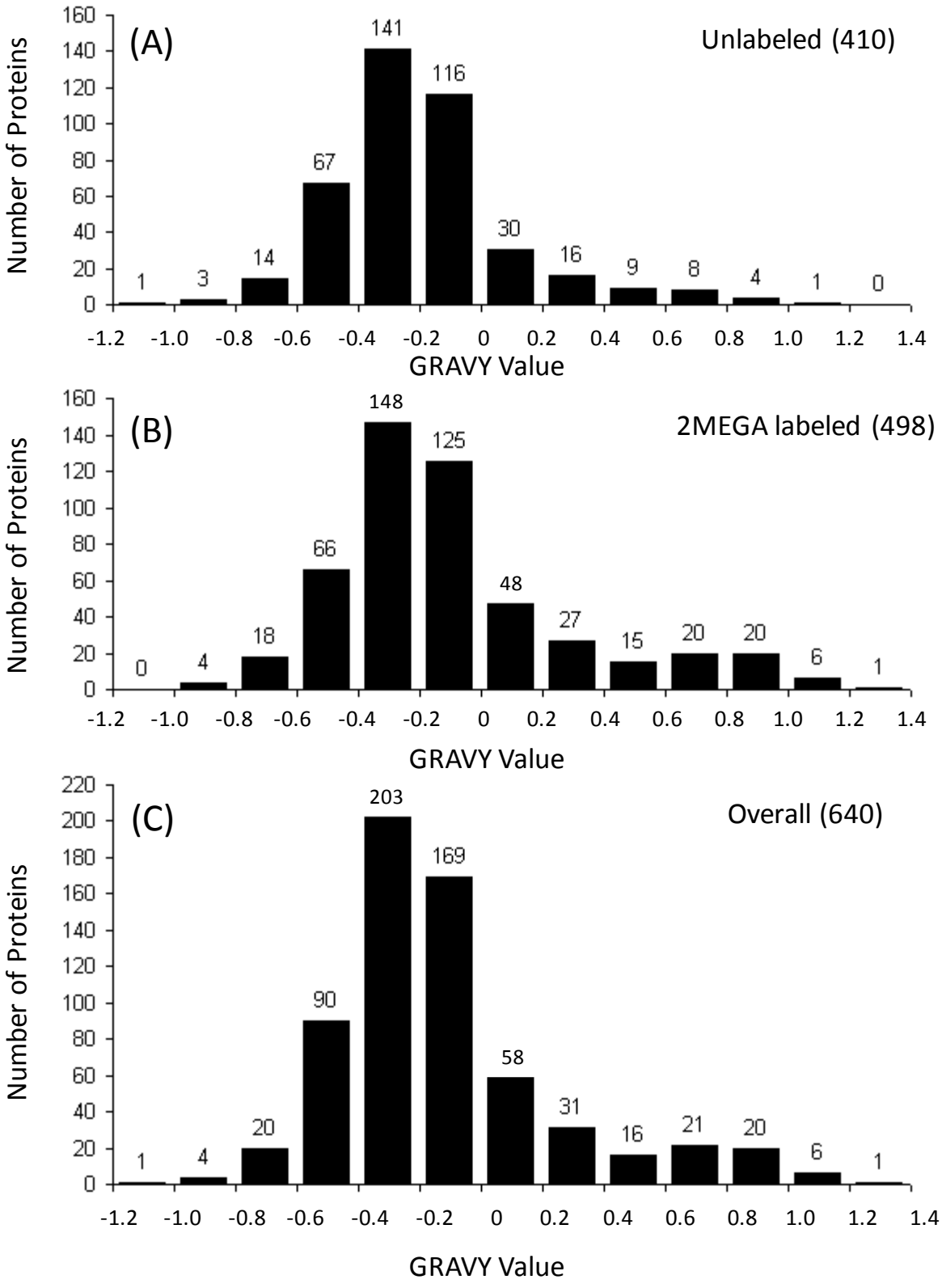


Figure 2.9 Histograms of GRAVY values for identified proteins from the a) labelled, b) 2-MEGA labelled, and c) combined overall datasets

strategy using a QTOF instrument with a mass accuracy of low parts-per-million. In this study, it was found that either a_1 ions for peptides having amino acid sequences starting with all amino acids except K and R, or a_1-17 or a_1-45 ions for peptides starting with K or R are greatly enhanced when analyzed by ESI MS/MS; these ions are usually difficult to detect in the MS/MS spectra of unlabelled peptides. The 2MEGA labelling strategy alleviated the biased detection of arginine-terminated peptides that is often observed in MALDI and ESI MS experiments. The enhanced a_1 or a_1-related ions in MS/MS spectra of the 2MEGA-labelled peptides provide additional information to re-examine the spectra and reduce the number of false positive identifications. Although spectra were manually examined in this study, it could be done automatically using a computer program. On the basis of the data evaluated, about 99% of peptides identified, using MASCOT identity as the threshold, were found to be true identifications. In addition, the addition of immonium ions and a-series ions in database searching could increase the number of positive identifications when QTOF is used to generate CID spectra. Overall, 640 unique proteins were identified from the *E. coli* membrane fraction and 336 proteins could be classified according to their known subcellular locations, including 171 membrane proteins and 86 membrane-associated proteins.

2.5 Literature Cited

(1) Hager, J. W. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 512-526.

(2) Morris, H. R.; Paxton, T.; Dell, A.; Langhorne, J.; Berg, M.; Bordoli, R. S.; Hoyes, J.; Bateman, R. H. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 889-896.

(3) Medzihradszky, K. F.; Campbell, J. M.; Baldwin, M. A.; Falick, A. M.; Juhasz, P.; Vestal, M. L.; Burlingame, A. L. *Anal. Chem.* **2000**, *72*, 552-558.

(4) Loboda, A. V.; Krutchinsky, A. N.; Bromirski, M.; Ens, W.; Standing, K. G. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1047-1057.

(5) Field, H. I.; Fenyo, D.; Beavis, R. C. *Proteomics* **2002**, *2*, 36-47.

(6) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976-989.

(7) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.

(8) Shen, Y.; Smith, R. D. *Electrophoresis* **2002**, *23*, 3106-3124.

(9) Washburn, M. P.; Wolters, D.; Yates, J. R. *Nat. Biotechnol.* **2001**, *19*, 242-247.

(10) Dai, J.; Shieh, C. H.; Sheng, Q.-H.; Zhou, H.; Zeng, R. *Anal. Chem.* **2005**, *77*, 5793-5799.

(11) Cargile, B. J.; Talley, D. L.; Stephenson, J. L., Jr. *Electrophoresis* **2004**, *25*, 936-945.

(12) Wolters, D. A.; Washburn, M. P.; Yates, J. R. *Anal. Chem.* **2001**, *73*, 5683-5690.

(13) MacCoss, M. J.; Wu, C. C.; Yates, J. R. *Anal. Chem.* **2002**, *74*, 5593-5599.

(14) Steen, H.; Mann, M. *Nat Rev Mol Cell Biol* **2004**, *5*, 699-711.

(15) Cargile, B. J.; Bundy, J. L.; Stephenson, J. L., Jr. *J. Proteome Res.* **2004**, *3*, 1082-1085.

(16) Rudnick, P. A.; Wang, Y.; Evans, E.; Lee, C. S.; Balgley, B. M. *J. Proteome Res.* **2005**, *4*, 1353-1360.

(17) Qian, W.-J.; Liu, T.; Monroe, M. E.; Strittmatter, E. F.; Jacobs, J. M.; Kangas, L. J.; Petritis, K.; Camp, D. G., II; Smith, R. D. *J. Proteome Res.* **2005**, *4*, 53-62.

(18) Olsen, J. V.; Ong, S.-E.; Mann, M. *Mol. Cell. Proteomics* **2004**, *3*, 608-614.

(19) Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L., Jr. *J. Proteome Res.* **2004**, *3*, 112-119.

(20) 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. *Anal. Chem.* **2003**, *75*, 1039-1048.

(21) Cagney, G.; Emili, A. *Nat. Biotechnol.* **2002**, *20*, 163-170.

(22) Keough, T.; Youngquist, R. S.; Lacey, M. P. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 7131-7136.

(23) Keough, T.; Youngquist, R. S.; Lacey, M. P. *Anal. Chem.* **2003**, *75*, 156A-165A.

(24) Lee, Y. H.; Kim, M.-S.; Choie, W.-S.; Min, H.-K.; Lee, S.-W. *Proteomics* **2004**, *4*, 1684-1694.

(25) Schnolzer, M.; Jedrzejewski, P.; Lehmann, W. D. *Electrophoresis* **1996**, *17*, 945-953.

(26) Shevchenko, A.; Chernushevich, I.; Ens, W.; Standing, K. G.; Thomson, B.; Wilm, M.; Mann, M. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1015-1024.

(27) Beardsley, R. L.; Sharon, L. A.; Reilly, J. P. *Anal. Chem.* **2005**, *77*, 6300-6309.

(28) Hsu, J.-L.; Huang, S.-Y.; Shiea, J.-T.; Huang, W.-Y.; Chen, S.-H. *J. Proteome Res.* **2005**, *4*, 101-108.

(29) Boja, E. S.; Sokoloski, E. A.; Fales, H. M. *Anal. Chem.* **2004**, *76*, 3958-3970.

(30) Ji, C.; Li, L.; Gebre, M.; Pasdar, M.; Li, L. *J. Proteome Res.* **2005**, *4*, 1419-1426.

(31) Ji, C.; Li, L. *J. Proteome Res.* **2005**, *4*, 734-742.

(32) Hsu, J.-L.; Huang, S.-Y.; Chow, N.-H.; Chen, S.-H. *Anal. Chem.* **2003**, *75*, 6843-6852.

(33) Ji, C.; Guo, N.; Li, L. *J. Proteome Res.* **2005**, *4*, 2099-2108.

(34) Fujiki, Y.; Hubbard, A. L.; Fowler, S.; Lazarow, P. B. *J. Cell Biol.* **1982**, *93*, 97-102.

(35) Blonder, J.; Goshe, M. B.; Moore, R. J.; Pasa-Tolic, L.; Masselon, C. D.; Lipton, M. S.; Smith, R. D. *J. Proteome Res.* **2002**, *1*, 351-360.

(36) Zhang, N.; Li, N.; Li, L. *J. Proteome Res.* **2004**, *3*, 719-727.

(37) Kimmel, J. R. *Methods Enzymol.* **1967**, *11*, 584-589.

(38) Beardsley, R. L.; Reilly, J. P. *Anal. Chem.* **2002**, *74*, 1884-1890.

(39) Zappacosta, F.; Annan, R. S. *Anal. Chem.* **2004**, *76*, 6618-6627.

(40) Brancia, F. L.; Montgomery, H.; Tanaka, K.; Kumashiro, S. *Anal. Chem.* **2004**, *76*, 2748-2755.

(41) Kyte, J.; Doolittle, R. F. *J. Mol. Biol.* **1982**, *157*, 105-132.

(42) Brancia, F. L.; Oliver, S. G.; Gaskell, S. J. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 2070-2073.

(43) Brancia, F. L.; Openshaw, M. E.; Kumashiro, S. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 2255-2259.

(44) Nouwens, A. S.; Cordwell, S. J.; Larsen, M. R.; Molloy, M. P.; Gillings, M.; Willcox, M. D. P.; Walsh, B. J. *Electrophoresis* **2000**, *21*, 3797-3809.

(45) Molloy, M. P.; Herbert, B. R.; Slade, M. B.; Rabilloud, T.; Nouwens, A. S.; Williams, K. L.; Gooley, A. A. *Eur. J. Biochem.* **2000**, *267*, 2871-2881.

(46) Wilkins, M. R.; Gasteiger, E.; Sanchez, J.-C.; Bairoch, A.; Hochstrasser, D. F. *Electrophoresis* **1998**, *19*, 1501-1505.

Chapter 3 - Targeted Quantitative Mass Spectrometric Identification of Differentially Expressed Proteins between Bax-Expressing and Deficient Colorectal Carcinoma Cells*

3.1 Introduction

Colon carcinoma is a genetic disease in which genetic defects result in the deregulation of apoptotic pathways, leading to the formation, progression, and resistance of cancer to treatment.[1] There are two pathways that control the initiation of apoptosis: death receptors-mediated extrinsic[2] and mitochondria-involved intrinsic pathways.[3] Current conventional chemotherapy targets the intrinsic mitochondrial pathway[4], whereas death receptors can be activated through interaction with ligands of the tumor necrosis factor (TNF) family. Tumor necrosis factor-related apoptosis inducing ligand (TRAIL) is a member of the TNF family of ligands[5, 6] and can bind on death receptors, DR4 and DR5, on the cell surface, thus triggering apoptotic cell death in cancer cells[7, 8]. Recombinant human TRAIL and its agonistic antibodies are currently in clinical trials for cancer treatment[9].

TRAIL-induced apoptosis occurs through the activation of both extrinsic and intrinsic pathways. Upon TRAIL ligation, DR4/DR5 recruits Fas-associated death domain (FADD) and caspase-8 for the formation of a death-inducing signaling complex (DISC)[10, 11]. In the DISC, caspase-8 is cleaved and initiates apoptosis through cleavage of caspase-3[12] and Bcl-2-inhibitory BH3 domain protein (Bid)[13]. The truncated Bid

---

A version of this chapter was published as: Wang, P.; Lo, A.; Young, J. B.; Song, J. H.; Lai, R.; Kneteman, N. M.; Hao, C.; Li, L. *Journal of Proteome Research* **2009**, *8*, 3403-3414. Dr. P. Wang, Mr. A. Lo, Dr. R. Lai, Dr. C. Hao, and Dr. L. Li were responsible for experimental design and manuscript preparation. Dr. P. Wang cultured cells, ran cell viability assays, and performed Western blot analyses. Dr. J. H. Song and Dr. N. Kneteman performed the microscopy work for cell morphology. Dr. P. Wang and Mr. A. Lo prepared samples, performed SCX separation, LC-MS analysis, and data analysis. Dr. J. B. Young developed and optimized the LC-MALDI MS interface.

interacts with pro-apoptotic Bcl-2 family proteins, Bax and Bak, to induce their oligomerization on the mitochondrial membrane. This leads to mitochondrial membrane permeability and the release of cytochrome *c*[14] and a second mitochondria-derived activator of caspase (Smac)[15] or direct inhibitor of apoptosis binding protein with low pI (DIABLO)[16]. In the cytosol, Smac interacts with X-linked inhibitor of apoptosis (XIAP) and releases its inhibition of caspase-3 cleavage,[17] whereas cytochrome *c* facilitates caspase-9 cleavage through the apoptosome[18]. Bax is required for both the extrinsic and intrinsic apoptotic pathways; however, the molecular mechanism by which Bax controls the mitochondrial membrane permeability remains to be defined.[3] Bax may bind voltage-dependent anion channel 1 (VDAC1) for the formation of mitochondrial permeability transition (MPT) pore and thus release of cytochrome *c.*[4] In contrast, others suggest that VDAC is dispensable for Bax-induced cytochrome *c* release.[19, 20] Bax apoptotic activity is negatively regulated through its interaction with anti-apoptotic Bcl-2 family proteins, Bcl-2 and Bcl-XL [3] and regulatory proteins such as Ku70[21] and 14-3-3[22, 23], although the protein-protein interaction models remain largely unknown. Bax is often mutated and inactivated in DNA mismatch repair (MMR)-deficient colon carcinomas[24], in which 94% of HCT116 human colon carcinoma cells have one wild type intact Bax allele [25]. Two clones of HCT116 cells have been generated with different Bax loci: a Bax-expressing ($Bax^{+/-}$) HCT116 clone with one intact wild type Bax allele and a Bax-deficient ($Bax^{-/-}$) clone in which one wild type Bax allele is genetically inactivated[25]. The $Bax^{-/-}$ HCT116 clone is resistant to chemotherapy and TRAIL-induced apoptosis[25-27]. Here, we report a proteomic analysis of these two cell lines to investigate the Bax associated protein network.

In earlier studies, the technique of N-terminal dimethylation after lysine guanidinylation (2MEGA) for introducing different isotope tags to two comparative proteome samples after protein digestion for identification of differentially expressed proteins was demonstrated.[28, 29] In this study, we developed a proteome analysis strategy by combining 2MEGA labelling with two-dimensional liquid chromatography (2D-LC) for peptide separation and matrix-assisted laser desorption ionization (MALDI) tandem mass spectrometry (MS/MS)[30] for targeted peptide quantification and identification. Using this quantitative proteome analysis method, we examined $Bax^{+/-}$

and $Bax^{-/-}$ HCT116 clone and identified 200 proteins differentially expressed in the two clones. Of the 200 proteins, we showed that Bax regulators and some Bax associated proteins such as VDAC1, VDAC2, Ku70, and 14-3-3 theta are differentially expressed in the $Bax^{-/-}$ clones and suggest that the potential protein-protein interaction networks are required for the expression and function of these proteins in regulation of apoptotic pathways.[31]

3.2 Experimental Section

3.2.1 Chemicals and Reagents

Recombinant human TRAIL (amino acids 114–281) was purchased from PeproTech, Inc. (Rocky Hill, USA). The mouse antibodies used included: anti-caspase-8 (MBL, Nagoya, Japan), anti-VDAC1, anti-VDAC2, anti-heat shock protein (HSP) 70, and anti-HSP90β (Santa Cruz Biotechnology, Santa Cruz, CA). The rabbit antibodies used included anti-caspase-9 (Cell Signal, Danvers, MA), anti-caspase-3, anti-HSP90α (StressGen, Victoria, BC), anti-Macrophage migration inhibitory factor (MIF), and anti-14-3-3 theta (Santa Cruz Biotechnology). The goat antibodies included anti-HSP60 (StressGen), anti-peroxiredoxin, and anti-leucine-rich PPR motif-containing protein (LRPPRC) (Santa Cruz Biotechnology). Horseradish peroxidase conjugated goat anti-mouse, donkey anti-goat, and goat anti-rabbit antibodies were from Jackson IR Labs (West Grove, USA). LC-MS grade water, acetonitrile, and methanol were purchased from Fisher Scientific (Edmonton, AB). LC-MS grade trifluoroacetic acid was purchased from Sigma-Aldrich (Oakville, ON, Canada). Isotopically enriched formaldehyde ($^{13}CD_2O$) was purchased from Cambridge Isotope Laboratories (Andover, MA). Sequencing grade modified trypsin was purchased from Promega (Madison, WI). All other chemicals used were of analytical grade and purchased from Sigma Aldrich (Oakville, ON).

3.2.2 Cell Cultures, Cell Viability and Morphological Observation of Apoptotic Cell Death

Human HCT116 $Bax^{+/-}$ and $Bax^{-/-}$ clones were kindly provided by Dr. Bert Vogelstein (Johns Hopkins University, Baltimore, MD) and cultured in McCoy 5α

medium (Invitrogen) supplemented with 10% FBS and 1% antibiotics. For the cell viability assay, cells (2×10^4 cells/100 µL) were planted in each well of a 96-well plate, cultured overnight and then treated with TRAIL in the doses as indicated in Figure 3.1. After incubation, cells were washed once with 100 µL PBS. Cell death was determined by an acid phosphatase assay. Briefly, 100 µL buffer containing 0.2 M sodium acetate (pH 5.5), 0.2% (v/v) Triton X-100, and 20 mM *p*-nitrophenyl phosphate were added to each well. The plates were placed in a water-jacketed incubator at 37 °C for 2 h. The reaction was stopped by the addition of 10 µL 1 M NaOH to each well and the color developed was measured at 405 nm using a microplate reader (Bio-Rad).

3.2.3. Western Blot

Cells in culture, treated or untreated, were harvested and lysed in 20 mM Tris-HCl (pH 7.4) containing 150 mM NaCl, 2 mM EDTA, 10% glycerol, 1% Triton X-100, 1 mM phenylmethylsulfonyl fluoride, and protease inhibitor cocktail (Sigma). The lysed cells were centrifuged at 20 000 *g* for 15 min and the supernatant was collected. Protein concentrations in the supernatant were determined by the BCA assay following the manufacturer's protocol (Bio-Rad). Equal amounts of protein were separated on SDS-PAGE gels and transferred onto Immunoblot membranes (Bio-Rad). The membranes were incubated overnight at 4 ˚C first with various primary antibodies, then for 1 hour with the horseradish peroxidase-conjugated-secondary antibodies, and examined with enhanced chemiluminescence (ECL) reagents (Amersham Biosciences, Piscataway, USA).

3.2.4 Cell Lysis and Protein Digestion

Cells were lysed using a French press at 35 000 psi with two passes into PBS. Proteins were precipitated by adding four parts acetone (v/v) and chilled overnight at -80 °C. Samples were centrifuged at 3 900 *g* for 60 minutes at 4 °C. The supernatant was removed and discarded. Protein pellets (~7 mg) were solubilized using 1% SDS before dilution to 0.1% SDS. Ammonium bicarbonate was added to a final concentration of 100 mM, followed by addition of dithiothreitol and iodoacetamide to reduce and alkylate disulfide bonds and digestion by trypsin in a 1:30 (w/w, enzyme:protein) ratio.

3.2.5 Peptide Desalting and Quantification

Samples were desalted and quantified using reversed-phase liquid chromatography using 0.1% TFA and 4% acetonitrile in water as solvent A and 0.1% TFA in acetonitrile as solvent B. The following gradient program was used (time in min, % B): 0.00, 0%; 5.00, 0%; 5.01, 90%; 10.00, 90%; 15.00, 0%; 25.00, 0 % B. The flow rate used was 1.0 mL/min using a 4.6 mm i.d. × 50 mm C_{18} 3µm particle size column (Varian). Samples were quantified based on their absorbance at 214 nm.

3.2.6 2MEGA Isotopic Labelling

Peptides were labelled using the 2MEGA labelling method[29]. In brief, samples were adjusted to pH 11 using 2 M NaOH and 6 M *O*-methylisourea was added. Samples were heated to 65 °C for 25 minutes to guanidinylate the lysines. The pH of the solution was adjusted with 10% TFA to approximately pH 7. Formaldehyde (4%, v/v; $^{12}CH_2O$ for light chain labelling or $^{13}CD_2O$ for heavy chain labelling) and sodium cyanoborohydride (1 M) were added to dimethylate the N-termini of the peptides. Samples were then desalted and quantified as described above. Heavy chain labelled $Bax^{+/-}$ was mixed with light chain labelled $Bax^{-/-}$ (denoted as A_HB_L for the convenience of discussion) in a 1:1 ratio based on the total peptide content by weight. Similarly, light chain labelled $Bax^{+/-}$ was mixed with heavy chain labelled $Bax^{-/-}$ (denoted as A_LB_H).

3.2.7 Strong Cation Exchange Chromatography

Labelled peptide mixtures were separated by strong cation exchange (SCX) chromatography using 50 mM KH_2PO_4 (pH 2.8) in 30% acetonitrile as solvent A and 1 M KCl in 50 mM KH_2PO_4 (pH 2.8) in 30% acetonitrile as solvent B. The following gradient program was used (time in min, % B): 0, 0%; 7, 0%; 8, 3%; 36, 14%; 44, 20 %; 49, 30%; 53, 50%; 58, 50 %; 60, 0%; 70, 0%. Fractions were collected in one minute fractions from 17 to 71 minutes, then desalted and quantified. Less abundant fractions were pooled together to form ~10 µg samples. A total of 19 SCX fractions were produced for subsequent reversed phase (RP) LC-MALDI MS.

3.2.8 Offline LC-MALDI MS

Individual SCX fractions were separated and analyzed by RPLC-MALDI MS using an LC-MALDI interface constructed in house[30, 32]. Approximately 10 µg of peptides were injected into a 1.0 mm i.d. × 150 mm C_8 column and directly spotted onto a custom-made 400-well MALDI plate. 0.1% TFA with 4% ACN in water was used as solvent A and 0.1% TFA in ACN was used as solvent B. The following gradient program was used (time in min, % B): 0, 0%; 19, 0 %; 20, 5%; 145, 30%; 156, 40%; 175, 45%; 180, 0%; 185, 0%. Fractions in 20-s intervals were collected from 40 to 174 min of the gradient run. After sample collection, individual MALDI wells were spotted with 0.6 µL of 0.6 µg/µL 2,5-dihydroxybenzoic acid solution in 50:50 methanol/water and allowed to dry.

3.2.9 MS Analysis and Targeted MS/MS Analysis

MALDI MS spectra were acquired with an Applied Biosystems/MDS Sciex QSTAR XL quadrupole time-of-flight mass spectrometer (Concord, ON, Canada). The peptide ion intensities in individual MS spectra were determined using peak picking and peak area calculation software ProTSData (Efeckta, Denver, CO). Peptide peaks with significant changes in relative intensities in two comparative samples (see section 3.3 for detailed explanation) were selected for automated MALDI MS/MS data acquisition via a precursor ion inclusion list entered into the instrument control software.

3.2.10 MASCOT Database Search and Data Analysis

MS/MS spectral data were searched using MASCOT with the following parameters: taxonomy: *Homo sapiens* (human); enzyme: trypsin; missed cleavages: 2; fixed modifications: guanidinyl (K) and carbamidomethyl (C); MS tolerance: 0.2 Da; MS/MS tolerance: 0.1 Da. Additional parameters included a modified ESI-QUAD-TOF ion fragmentation series that permitted a-type ions. Data were searched twice; first selecting the light isotope modification (+C_2H_4, +28.0313 Da, N-term) and then selecting the heavy isotope modification (+$^{13}C_2D_4$, +34.0631 Da, N-term) to identify peaks of peptides with either labelling tag. Peptide and protein identification data was extracted from the MASCOT files using in-house software (ProteinExtractor). Relative ratios for the identified peptide pairs were taken from the ProTSData output. In cases where the peptide eluted over multiple reversed-phase fractions, the fraction

containing the highest intensity of the peptide peak was used for calculating the relative ratio. When a peak identified did not have the corresponding isotope counterpart (i.e., single peak, instead of a pair), its relative ratio was taken as its signal-to-noise ratio (S/N) divided by eight, which was the maximum S/N ratio possible for the unobserved peak. From our working experience with the use of isotope labelling LC-MALDI for quantitative analysis, it was found that peaks with S/N of greater than 8 were highly reproducible in replicate experiments, compared to the low intensity peaks with S/N < 8.

3.2.11 Protein-Protein Interaction Analysis

HiMAP and Metacore[33, 34] were used to map the differentially expressed proteins into the protein-protein interaction networks. Differentially expressed proteins were converted into appropriate gene symbols and uploaded into both HiMAP and Metacore for analysis. The protein-protein interaction analysis[34] was based on literature-confirmed interactions from the Human Protein Reference Database, yeast-two-hybrid-defined interactions, and predicted interactions generated by a Bayesian analysis. For network analysis, two algorithms were used: 1) the sparse interaction algorithm to map direct protein-protein interactions among differentially expressed proteins and 2) the bridge interaction algorithm to map the shortest path for interactions.

3. 3 Results

3.3.1 Blockage of TRAIL-induced Apoptosis in $Bax^{-/-}$ Clone

Bax is required for both the extrinsic and intrinsic apoptosis pathway[25-27]. To examine these pathways, $Bax^{+/-}$ and $Bax^{-/-}$ cells were treated with TRAIL in a series of dilutions and the cell death and the cleavage of caspase-8 in the extrinsic pathway and the cleavage of Bid and caspase-9 in the intrinsic pathway were examined. TRAIL induced a significant cell death in $Bax^{+/-}$ but not $Bax^{-/-}$ cells as demonstrated by cell viability assay (Figure 3.1A) and observed morphologically under phase contract microscopy (Figure 3.1B). Western blotting further revealed the cleavage of caspase-8, Bid, caspase-9 and caspase-3, a downstream caspase of caspase-8 and caspase-9 in

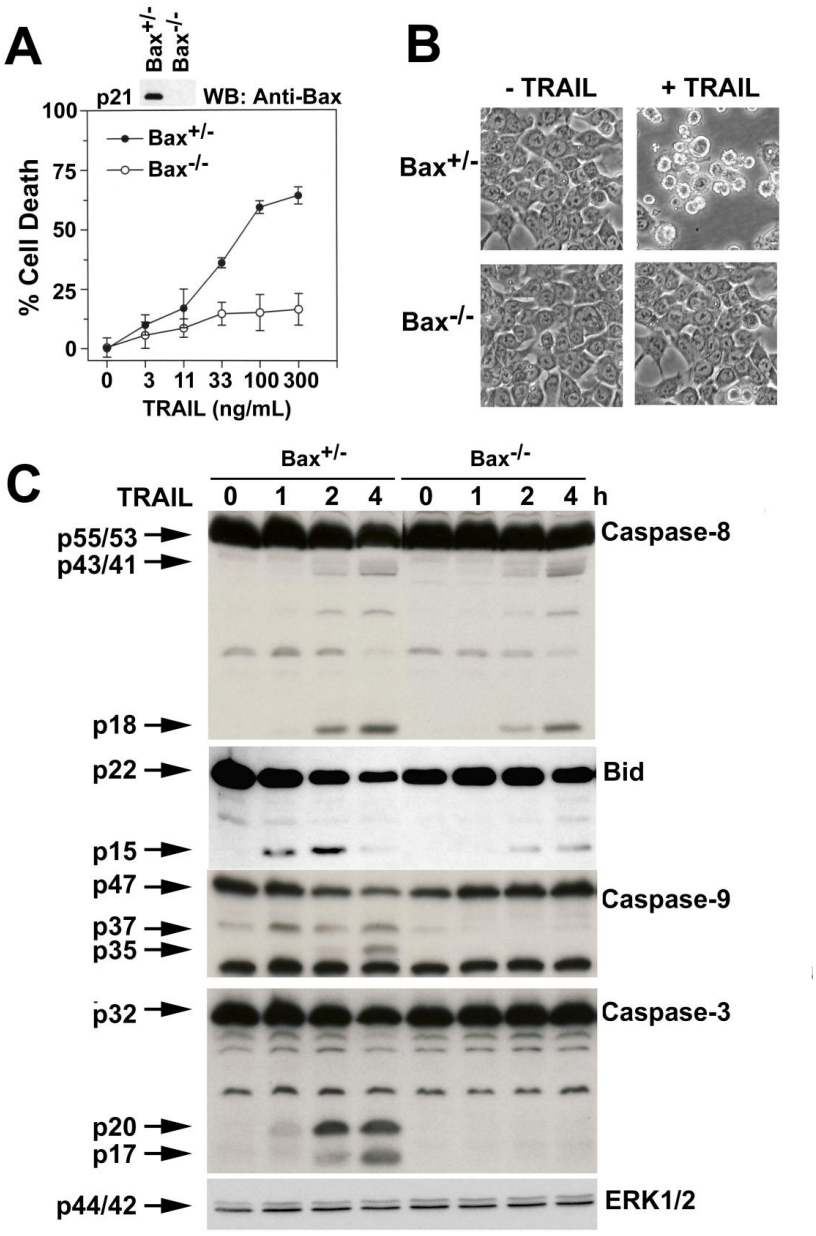


Figure 3.1 Bax plays a crucial role in the TRAIL-induced apoptosis pathway. (A) HCT 116 $Bax^{+/-}$ and $Bax^{-/-}$ cells were analyzed via Western blot for Bax expression levels and treated with recombinant TRAIL for 24 h for cell viability analysis. (B) Two HCT 116 clones were treated with 100ng/ml TRAIL for 4 hours for microscope imaging and detection of apoptotic bodies. (C) $Bax^{+/-}$ and $Bax^{-/-}$ clones were treated with 100ng/ml TRAIL for Western blot analysis of the cleavage of caspase-8, caspase-9, Bid, and caspase-3 with ERK1/2 as the loading control.

$Bax^{+/-}$ cells (Figure 3.1C). In contrast, the cleavage of caspase-8 and Bid, but not caspase-9 and caspase-3 were seen in $Bax^{-/-}$ cells. These results indicate that Bax inactivation blocks TRAIL-induced apoptosis at the Bid downstream mitochondrial pathway. Some crucial pro-apoptotic factors of Bcl-2 family including Bak and Puma are expressed in HCT116 cells[25, 35, 36] but fails to restore the apoptotic pathway in $Bax^{-/-}$ cells. Thus, Bax is indispensable in TRAIL-induced extrinsic and intrinsic apoptotic pathways and the $Bax^{+/-}$ and $Bax^{-/-}$ cells therefore provide comparative cell models in proteome analysis of the potential Bax-associated protein-protein interaction network in the absence and presence of Bax.

3.3.2 Method Validation of 2MEGA Quantitative MS

Before we applied the quantitative LC-MALDI method for proteomic comparison of the $Bax^{+/-}$ and $Bax^{-/-}$ cells, we had to determine the relative intensity threshold of the isotope differentially labelled peak pairs in MS analysis above which the changes were deemed to be statistically significant. In this work, a 1:1 mixture of two standard proteins, cytochrome *c* and myoglobin, was used as a test system to evaluate the confidence intervals. The protein solution was digested with trypsin and half of the sample was labelled with the light chain reagent, $^{12}CH_2O$-formaldehyde, and the other half with the heavy chain reagent, $^{13}CD_2O$-formaldehyde. The digest samples were desalted, quantified, and mixed in a light:heavy ratio of 2:1 and 1:2, followed by LC-MALDI MS. A total of 21 peptide pairs were identified from each mixture. Ratios of 2.00 and 0.51 were observed in the two mixtures with an average coefficient of variation (CV) of 0.25. Based on the determined CV, the threshold for differential expression in our subsequent quantification work was set at 1.50-fold, which represents two standard deviations and 95% confidence in reporting the relative peptide quantification results.

3.3.3. Forward and Reverse Labelling Strategy and MS Analysis

Replicate determinations of the two samples were conducted as indicated by the quantitative MS workflow shown Figure 3.2 using the forward (A_HB_L) and reverse (A_LB_H) labelling strategy. $Bax^{+/-}$and $Bax^{-/-}$ cells were lysed and proteins were extracted,

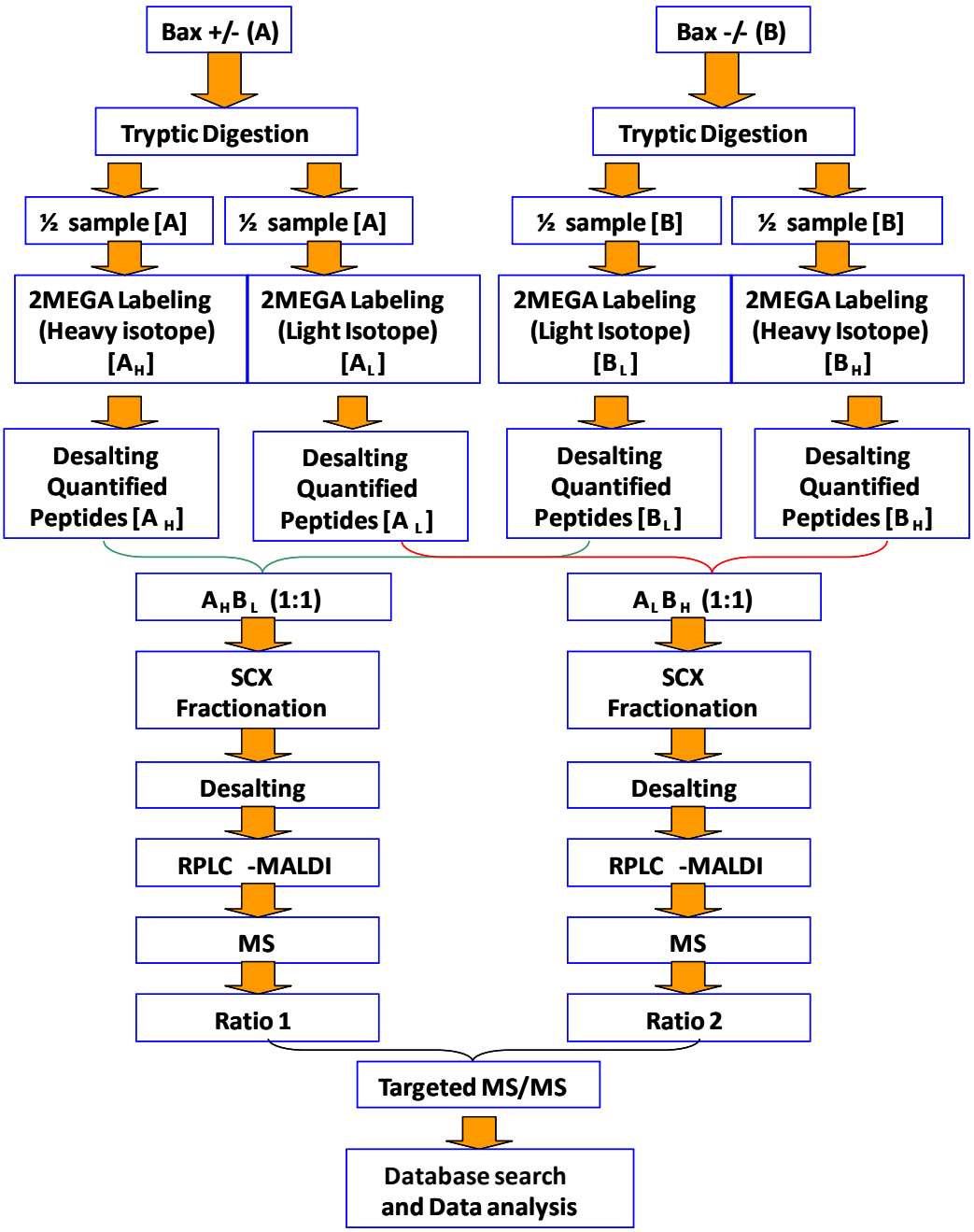


Figure 3.2 Experimental workflow for the comparison of HCT116 Bax$^{+/-}$ and Bax$^{-/-}$ cells

re-solubilized, and then digested with trypsin. Both $Bax^{+/-}$ and $Bax^{-/-}$ digest samples were divided into half and labelled using the 2MEGA protocol. All of the labelled digests were quantified and desalted to facilitate accurate 1:1 mixing. The heavy chain labelled $Bax^{+/-}$ sample (A_H) was mixed with the light chain labelled $Bax^{-/-}$ sample (B_L) to form a mixture of A_HB_L and the reverse labelling produced the A_LB_H mixture (see Figure 3.2). The resulting two mixtures were individually subjected to SCX chromatography fractionation. The SCX fractions were then separated by RPLC onto a homemade 400-well MALDI plate using an online LC-MALDI interface, followed by an MS survey scan[30].

Processing the MS survey spectra is crucial in determining the peptide ions with significant relative abundance changes in the $Bax^{+/-}$ and $Bax^{-/-}$ digest samples. In our work, if the same peptide was present in both samples, it would be identified as a peak pair in a MS spectrum with a mass difference of 6.032 Da. Pairs with a relative abundance difference of greater than 1.50 or less than 0.67 were placed into a "quantitative pairs list". If the peptide was only present in one of the two samples, a single, unpaired peak would be observed in the MS spectrum. Unpaired peaks with a signal-to-noise ratio of greater than 15 were placed into a "quantitative singles list". For the forward and reverse labelled peptide samples (i.e., A_HB_L and A_HB_L mixtures), two different SCX fractions were run. To compare the quantitative results from the two runs, the individual quantitative peak lists (either in pairs or singles) were matched against one another to identify peaks found under similar chromatographic conditions (±3 min in RPLC retention time, ±2 SCX fractions, and ±0.2 Da in *m/z* ratio). If a pair was found in both quantitative lists within the above criteria, it was selected for targeted MS/MS analysis. The analysis of the unpaired peaks was performed in a similar fashion, requiring the presence of an unpaired peak with mass difference ± (6.023±0.200 Da) and similar chromatographic behavior in the complementary mixture. Since a heavy chain labelled peptide found in one mixture would be found as a light chain labelled peptide in the complementary mixture, this *m/z* restriction was used to increase confidence in the peak matches. The peaks were analyzed automatically by MALDI MS/MS via the use of an inclusion list. In the end, the identified peptides from MASCOT database search of the MS/MS spectra were combined with quantification

data of these peptides from ProTSData. From the list of identified peptides, a list of corresponding proteins was generated.

3.3.4 Peptide Quantification and Identification

The 2MEGA labelling provides a differential mass tag of approximately 6 Da for the comparison of peptide pairs. The 6 Da mass difference is not significantly affected by the natural abundance isotope envelopes of the light and heavy chain labelled peptides, facilitating relative quantification of peptide pairs. As an example, Figure 3.3A shows the overlaid MS spectra identified as HELQANCYEEVK from cofilin-1 in the A_HB_L and A_LB_H mixtures. The relative ratio of the peptide pair is 2.65 (Light/Heavy) in the A_HB_L mixture and the corresponding change in the complementary A_LB_H mixture is 0.48 (Light/Heavy). In both cases, the $Bax^{-/-}$ clone (i.e., the B sample) shows relative over-expression by approximately 2.4-fold. The MS/MS spectrum used to identify the heavy-chain labelled HELQANCYEEVK is shown in Figure 3.3B. Since only relatively high abundance peptides (S/N >8) were quantified and then selected for MS/MS, the quality of MS/MS spectra were generally good and often led to peptide identification via database searching. In addition, the a_1 ion is clearly visible in the MS/MS spectrum which can be used to confirm the identity of the N-terminal amino acid, further increasing the confidence level of peptide identification.

Since the mass-to-charge ratio difference and chromatographic behaviors of matched pairs were sufficiently stringent conditions, peptide pairs were considered identified if any of the four constituent peaks were identified. Approximately 69% (124 out of 180) of peptides identified from the pairs data were identified by at least two MS/MS scans scoring above the identity threshold. However, for quantitative analysis, manual inspection of the data was found to be essential. While most pairs showed the correct ratios in the two runs of forward and reverse labelled mixtures (e.g., increasing in A_HB_L and decreasing in the complement A_LB_H), four pairs were found with the same direction in the two runs. Two of the pairs were found to be overlapping with other peaks and were discarded after manual data analysis. There was no clear rationale as to why the other two pairs were incorrect and are likely due to random error or chemical noise.

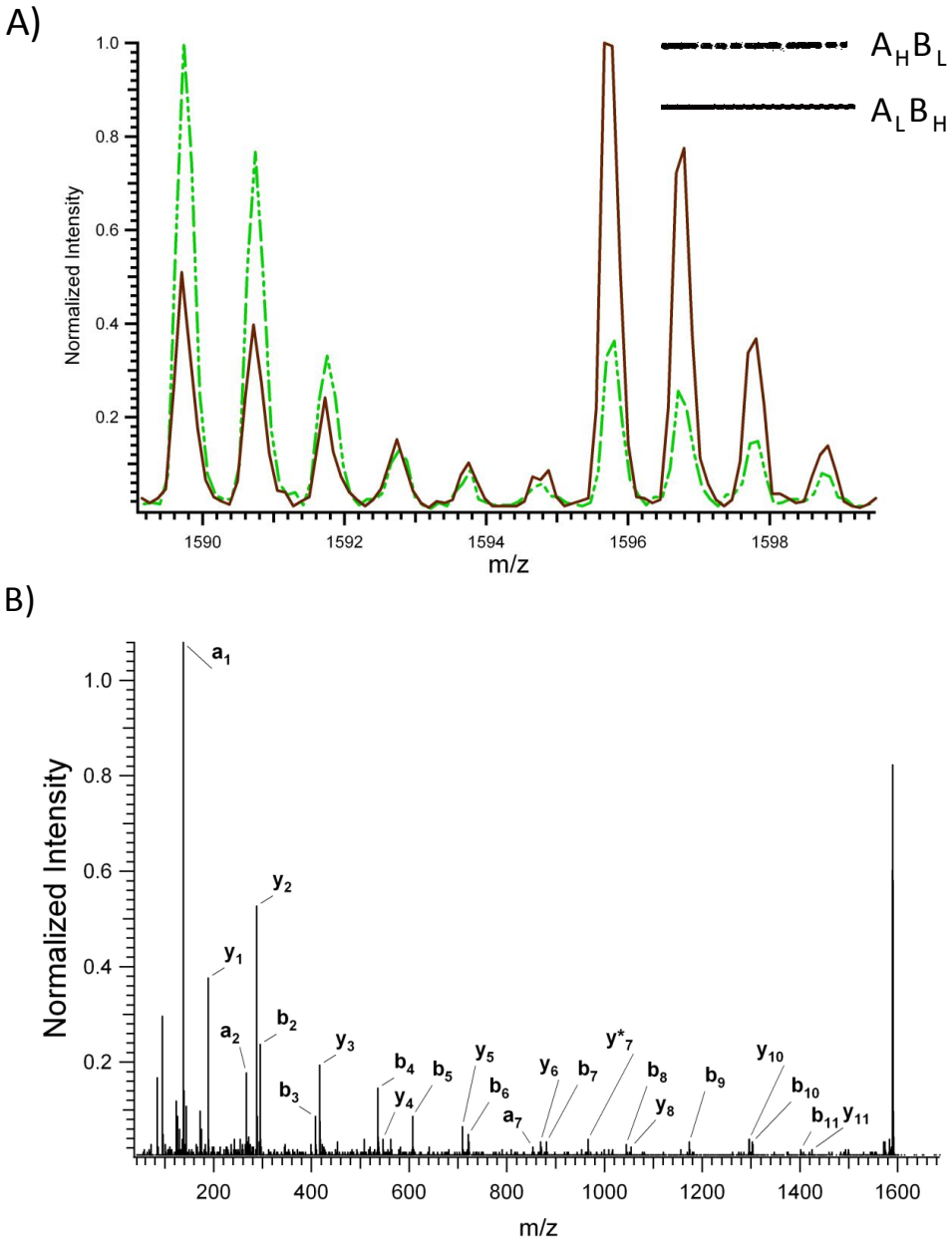


Figure 3.3 a) Overlaid MS spectra of HELQANCYEEVK from forward and reverse labelling experiments and b) MS/MS spectrum of heavy labelled HELQANCYEEVK

Similarly to the pairs, single peaks were considered to be positively identified if the corresponding peak was detected in either A_HB_L or A_LB_H and the MASCOT search of the MS/MS spectrum generated a score above identity. After manual data checking, twelve peptides with scores above the identity threshold from the single peaks data were found to be from misidentified pairs and were discarded. This is likely due to an error with the software used to determine the peaks and pairs which occasionally included observed peaks as both a pair and an unpaired peak. Of the remaining positively identified 133 peptides, 85 (64%) were identified in only one mixture with 48 (36%) positively identified in both mixtures. A total of 313 unique peptides were identified from the pairs (180) and single-peaks data (133) to generate a list of 200 unique proteins. Among those 200 proteins, 20 proteins were identified to be up-regulated, 153 proteins were identified to be down-regulated in the $Bax^{-/-}$ cells with 27 proteins having uncertain protein level ratios. Table 3.1 lists the proteins ranked by the decreasing ratio and their relative abundance differences in the two clones.

3.3.5 Analysis of Interaction Network of Differentially Expressed Proteins

To explore possible interactions among differentially expressed proteins identified by the 2MEGA quantitative MS method, the 200 proteins listed in Table 3.1 were analyzed by searching literature and the Swiss-Prot database for various protein functions and protein-protein interactions. Among these proteins, 57 differentially expressed proteins involved in the processes of apoptosis, cell cycle regulation, DNA repair, stress regulation, detoxification, drug resistance, protein modification, and cellular signaling were selected to be analyzed with HiMAP and Metacore[34]. Thirty of these proteins were connected using the sparse interaction algorithm, which connects proteins based on direct known or predicted interactions (Figure 3.4). Using the bridge path algorithm, which allows intermediary proteins to map short paths of interaction, 46 of the 57 differentially expressed proteins including both up-regulated and down-regulated proteins were brought together in the network (Figure 3.5). Among these proteins, four groups of proteins, including MPT channel proteins, heat shock proteins, Bax-regulator proteins, and oxidative stress triggered proteins, were found differentially expressed between $Bax^{+/-}$ and $Bax^{-/-}$ clones. This protein-protein

Table 3.1 List of Differentially Expression Proteins from Bax$^{+/-}$ and Bax$^{-/-}$ clones

Down-regulated Proteins

Swiss-Prot ID	Protein name	(# of Peptides, Ratio)	Swiss-Prot ID	Protein name	(# of Peptides, Ratio)
Q12906	Interleukin enhancer-binding factor 3	(1, 16.84)	P12268	Inosine-5'-monophosphate dehydrogenase 2	(1, 4.43)
P23284	Peptidyl-prolyl cis-trans isomerase B precursor	(1, 12.49)	P35579	Myosin-9	(1, 4.43)
P06744	Glucose-6-phosphate isomerase	(1, 10.28)	P40429	60S ribosomal protein L13a	(2, 4.41)
P50914	60S ribosomal protein L14	(1, 9.25)	Q9Y617	Phosphoserine aminotransferase	(1, 4.40)
P62081	40S ribosomal protein S7.	(1, 8.80)	P50395	Rab GDP dissociation inhibitor beta	(1, 4.37)
P62249	40S ribosomal protein S16.	(1, 8.41)	P55884	Eukaryotic translation initiation factor 3 subunit 9	(2, 4.26)
P49411	Elongation factor Tu	(2, 7.99)	P15880	40S ribosomal protein S2	(1, 4.24)
O75390	Citrate synthase	(1, 7.45)	P00558	Phosphoglycerate kinase 1	(1, 4.08)
P12004	Proliferating cell nuclear antigen	(1, 7.43)	P60660	Myosin light polypeptide 6	(2, 4.08)
P25786	Proteasome subunit alpha type 1	(1, 7.26)	P23396	40S ribosomal protein S3.	(2, 4.06)
P61254	60S ribosomal protein L26	(1, 7.22)	P45880	Voltage-dependent anion-selective channel protein 2	(1, 4.04)
Q460N5	Poly [ADP-ribose] polymerase 14	(1, 6.75)	Q07020	60S ribosomal protein L18	(1, 4.02)
P48643	T-complex protein 1 subunit epsilon	(1, 6.62)	P17174	Aspartate aminotransferase	(1, 4.01)
P22102	Trifunctional purine biosynthetic protein adenosine-3	(2, 6.59)	P12429	Annexin A3	(1, 3.90)
Q8WXH0	Nesprin-2	(1, 6.39)	P18077	60S ribosomal protein L35a	(1, 3.89)
Q9Y5B9	FACT complex subunit SPT16	(1, 6.33)	Q9NYF0	Dapper homolog 1	(1, 3.87)
Q7Z628	Neuroepithelial cell-transforming gene 1 protein	(1, 6.27)	O00231	26S proteasome non-ATPase regulatory subunit 11	(1, 3.86)
P09429	High mobility group protein B1	(1, 6.25)	P49736	DNA replication licensing factor MCM2	(1, 3.84)
P30044	Peroxiredoxin-5	(1, 6.10)	P60866	40S ribosomal protein S20.	(1, 3.78)
P61247	40S ribosomal protein S3a.	(1, 5.83)	P15259	Phosphoglycerate mutase 2	(1, 3.76)
P27348	14-3-3 protein theta	(2, 5.60)	P04792	Heat shock protein beta-1 (HspB1)	(1, 3.74)
P62899	60S ribosomal protein L31	(2, 5.54)	P13010	ATP-dependent DNA helicase 2 subunit 2	(2, 3.73)
P84103	Splicing factor, arginine/serine-rich 3	(1, 5.54)	Q06830	Peroxiredoxin-1	(1, 3.69)
P62805	Histone H4.	(1, 5.46)	O15259	Nephrocystin-1	(1, 3.64)
Q9UPV7	Protein KIAA1045	(1, 5.20)	P56705	Protein Wnt-4 precursor	(1, 3.64)
Q15102	Platelet-activating factor acetylhydrolase IB subunit gamma	(1, 5.11)	P99999	Cytochrome c	(1, 3.63)
Q7Z4S6	Kinesin-like protein KIF21A	(1, 4.93)	P55084	Trifunctional enzyme subunit beta	(1, 3.61)
P50991	T-complex protein 1 subunit delta	(1, 4.75)	P60900	Proteasome subunit alpha type 6	(1, 3.60)
P08107	Heat shock 70 kDa protein 1	(2, 4.72)	P62937	Peptidyl-prolyl cis-trans isomerase A	(1, 3.54)
P17987	T-complex protein 1 subunit alpha	(2, 4.71)	O14949	Ubiquinol-cytochrome c reductase complex	(1, 3.53)
Q13748	Tubulin alpha-3C/D chain	(3, 4.70)	P05141	ADP/ATP translocase 2	(2, 3.50)
P14868	Aspartyl-tRNA synthetase	(2, 4.69)	Q03252	Lamin-B2.	(1, 3.42)
P83731	60S ribosomal protein L24	(1, 4.68)	P10809	60 kDa heat shock protein	(1, 3.34)
P62424	60S ribosomal protein L7a	(1, 4.65)	P78527	DNA-dependent protein kinase catalytic subunit	(2, 3.34)
Q9UQ80	Proliferation-associated protein 2G4	(1, 4.48)	P62269	40S ribosomal protein S18	(2, 3.33)
P09960	Leukotriene A-4 hydrolase	(1, 4.43)	Q16629	Splicing factor, arginine/serine-rich 7	(1, 3.26)

Table 3.1 continued

Swiss-Prot ID	Protein name	(# of Peptides, Ratio)	Swiss-Prot ID	Protein name	(# of Peptides, Ratio)
Q9Y5Z6	Beta-1,3-galactosyltransferase 1	(1, 3.25)	P18669	Phosphoglycerate mutase 1	(1, 2.51)
Q5TZA2	Rootletin	(1, 3.21)	P23246	Splicing factor, proline- and glutamine-rich	(1, 2.51)
Q99623	Prohibitin-2	(2, 3.12)	P61978	Heterogeneous nuclear ribonucleoprotein K	(1, 2.51)
P62826	GTP-binding nuclear protein Ran	(2, 3.11)	Q8IZK6	Mucolipin-2.	(1, 2.45)
Q14697	Neutral alpha-glucosidase AB precursor	(2, 3.11)	P57088	Transmembrane protein 33	(1, 2.44)
P62266	40S ribosomal protein S23	(1, 3.09)	Q9Y266	Nuclear migration protein nudC	(1, 2.42)
P15311	Ezrin	(5, 3.05)	Q9H4Z2	Zinc finger protein 335	(1, 2.38)
Q9H361	Polyadenylate-binding protein 3	(2, 3.04)	P27482	Calmodulin-like protein 3	(1, 2.29)
P08195	4F2 cell-surface antigen heavy chain	(1, 3.02)	P21796	Voltage-dependent anion-selective channel protein 1	(3, 2.27)
P36542	ATP synthase gamma chain	(1, 3.01)	O75829	Chondromodulin-1 precursor	(1, 2.26)
O14980	Exportin-1	(1, 2.97)	P39019	40S ribosomal protein S19.	(1, 2.18)
P26639	Threonyl-tRNA synthetase	(1, 2.96)	P17096	High mobility group protein HMG-I/HMG-Y	(1, 2.17)
P26373	60S ribosomal protein L13	(1, 2.95)	Q13200	26S proteasome non-ATPase regulatory subunit 2	(1, 2.17)
P62701	40S ribosomal protein S4, X isoform	(2, 2.95)	P14174	Macrophage migration inhibitory factor	(1, 2.16)
P49327	Fatty acid synthase (EC 2.3.1.85)	(2, 2.94)	Q9NVA2	Septin-11.	(1, 2.15)
P08758	Annexin A5	(1, 2.93)	Q9Y662	Heparan sulfate glucosamine 3-O-sulfotransferase 3B1	(1, 2.14)
P39023	60S ribosomal protein L3	(1, 2.91)	P55769	NHP2-like protein 1	(1, 2.10)
Q00839	Heterogeneous nuclear ribonucleoprotein U	(2, 2.91)	P46781	40S ribosomal protein S9.	(2, 2.08)
P12956	ATP-dependent DNA helicase 2 subunit 1	(2, 2.88)	P06576	ATP synthase subunit beta	(2, 2.07)
P47895	Aldehyde dehydrogenase 1A3	(1, 2.88)	Q6P5R6	Ribosomal protein L22-like 1	(1, 2.07)
A0AVF1	Tetratricopeptide repeat protein 26	(1, 2.84)	P09211	Glutathione S-transferase P	(2, 2.06)
P38606	Vacuolar ATP synthase catalytic subunit A	(1, 2.84)	P62273	40S ribosomal protein S29	(1, 2.06)
P50502	Hsc70-interacting protein (Hip)	(1, 2.83)	P49207	60S ribosomal protein L34	(1, 2.03)
P22087	rRNA 2'-O-methyltransferase fibrillarin	(1, 2.78)	Q15233	Non-POU domain-containing octamer-binding protein	(1, 2.01)
P68036	Ubiquitin-conjugating enzyme E2 L3	(1, 2.68)	P61313	60S ribosomal protein L15	(1, 2.00)
P05388	60S acidic ribosomal protein P0	(1, 2.67)	Q99832	T-complex protein 1 subunit eta	(2, 1.98)
P12277	Creatine kinase B-type	(1, 2.65)	P46777	60S ribosomal protein L5	(1, 1.94)
P52565	Rho GDP-dissociation inhibitor 1	(1, 2.64)	Q9NTK5	Putative GTP-binding protein 9	(1, 1.94)
P35637	RNA-binding protein FUS	(1, 2.62)	P62851	40S ribosomal protein S25.	(1, 1.93)
Q01844	RNA-binding protein EWS (EWS oncogene)	(1, 2.62)	Q02878	60S ribosomal protein L6	(2, 1.91)
P13639	Elongation factor 2	(4, 2.58)	P42704	Leucine-rich PPR motif-containing protein	(1, 1.86)
Q12931	Tumornecrosis factor type 1 receptor-associated protein	(1, 2.58)	Q01105	Protein SET	(1, 1.86)
Q8IZY2	ATP-binding cassette sub-family A member 7	(1, 2.57)	P27635	60S ribosomal protein L10	(1, 1.85)
O00299	Chloride intracellular channel protein 1	(1, 2.51)	P29401	Transketolase (EC 2.2.1.1)	(1, 1.85)

Table 3.1 continued

Swiss-Prot ID	Protein name	(# of Peptides, Ratio)	Swiss-Prot ID	Protein name	(# of Peptides, Ratio)
Q13151	Heterogeneous nuclear ribonucleoprotein A0	(1, 1.85)	P07437	Tubulin beta chain	(1, 1.70)
Q9NP73	Probable glycosyltransferase GLT28D1	(1, 1.85)	P36551	Coproporphyrinogen III oxidase	(1, 1.67)
Q07666	Src-associated in mitosis 68kDa protein	(1, 1.84)	P30050	60S ribosomal protein L12	(1, 1.64)
P09651	Heterogeneous nuclear ribonucleoprotein A1	(1, 1.82)	P62857	40S ribosomal protein S28.	(1, 1.62)
Q15369	Transcription elongation factor B polypeptide 1	(1, 1.80)	P62861	40S ribosomal protein S30	(1, 1.61)
P50990	T-complex protein 1 subunit theta	(1, 1.79)	P05783	Keratin, type I cytoskeletal 18	(1, 1.54)
P62888	60S ribosomal protein L30	(1, 1.72)			

Up-regulated Proteins

Swiss-Prot ID	Protein name	(# of Peptides, Ratio)	Swiss-Prot ID	Protein name	(# of Peptides, Ratio)
Q8TD47	40S ribosomal protein S4, Y isoform 2.	(1, 0.65)	Q96MA6	Putative adenylate kinase-like protein C9orf98	(1, 0.46)
Q13765	Nascent polypeptide-associated complex subunit alpha	(1, 0.64)	P33527	Multidrug resistance-associated protein 1	(1, 0.44)
P06703	Protein S100-A6	(1, 0.61)	P23528	Cofilin-1	(1, 0.43)
Q8IY18	Structural maintenance of chromosomes protein 5	(1, 0.61)	P09848	Lactase-phlorizin hydrolase precursor	(1, 0.41)
P18124	60S ribosomal protein L7	(2, 0.60)	P27695	DNA-(apurinic or apyrimidinic site) lyase	(1, 0.41)
P04406	Glyceraldehyde-3-phosphate dehydrogenase	(2, 0.54)	P60842	Eukaryotic initiation factor 4A-I	(1, 0.40)
Q8TF42	Suppressor of T-cell receptor signaling 1	(1, 0.51)	P04114	Apolipoprotein B-100 precursor	(1, 0.28)
P62913	60S ribosomal protein L11	(1, 0.49)	P84090	Enhancer of rudimentary homolog	(1, 0.25)
P06733	MBP-1	(3, 0.48)	P20700	Lamin-B1.	(1, 0.18)
P16403	Histone H1.2	(1, 0.46)	Q9NZC4	ETS homologous factor	(1, 0.18)

Uncertain

Swiss-Prot ID	Protein name	(# of Peptides, Ratio)	Swiss-Prot ID	Protein name	(# of Peptides, Ratio)
P07900	Heat shock protein HSP 90-alpha	(9, Uncertain)	P14625	Heat shock protein 90 kDa beta member 1	(3, Uncertain)
P05787	Keratin, type II cytoskeletal 8	(7, Uncertain)	P60174	Triosephosphate isomerase	(3, Uncertain)
P08727	Keratin, type I cytoskeletal 19	(7, Uncertain)	P62988	Ubiquitin	(3, Uncertain)
P11142	Heat shock 70 kDa protein 8	(6, Uncertain)	O43707	Alpha-actinin-4	(2, Uncertain)
P00338	L-lactate dehydrogenase A chain	(5, Uncertain)	P04075	Fructose-bisphosphate aldolase A	(2, Uncertain)
P60709	Actin, cytoplasmic 1	(5, Uncertain)	P04843	Dolichyl-oligosaccharide-protein glycosyltransferase	(2, Uncertain)
P68104	Elongation factor 1-alpha 1	(5, Uncertain)	P09382	Galectin-1	(2, Uncertain)
P07910	Nuclear ribonucleoprotein C1/C2	(4, Uncertain)	P20671	Histone H2A type 1-D	(2, Uncertain)
P08238	Heat shock protein HSP 90-beta	(4, Uncertain)	P22392	Nucleoside diphosphate kinase B	(2, Uncertain)
P14618	Pyruvate kinase isozymes M1/M2	(4, Uncertain)	P30041	Peroxiredoxin-6	(2, Uncertain)
P04083	Annexin A1	(3, Uncertain)	P62263	40S ribosomal protein S14	(2, Uncertain)
P07195	L-lactate dehydrogenase B chain	(3, Uncertain)	P62807	Histone H2B type 1-C/E/F/G/I	(2, Uncertain)
P07237	Disulfide-isomerase precursor	(3, Uncertain)	Q02539	Histone H1.1	(2, Uncertain)
P11021	Heat shock 70 kDa protein 5	(3, Uncertain)			

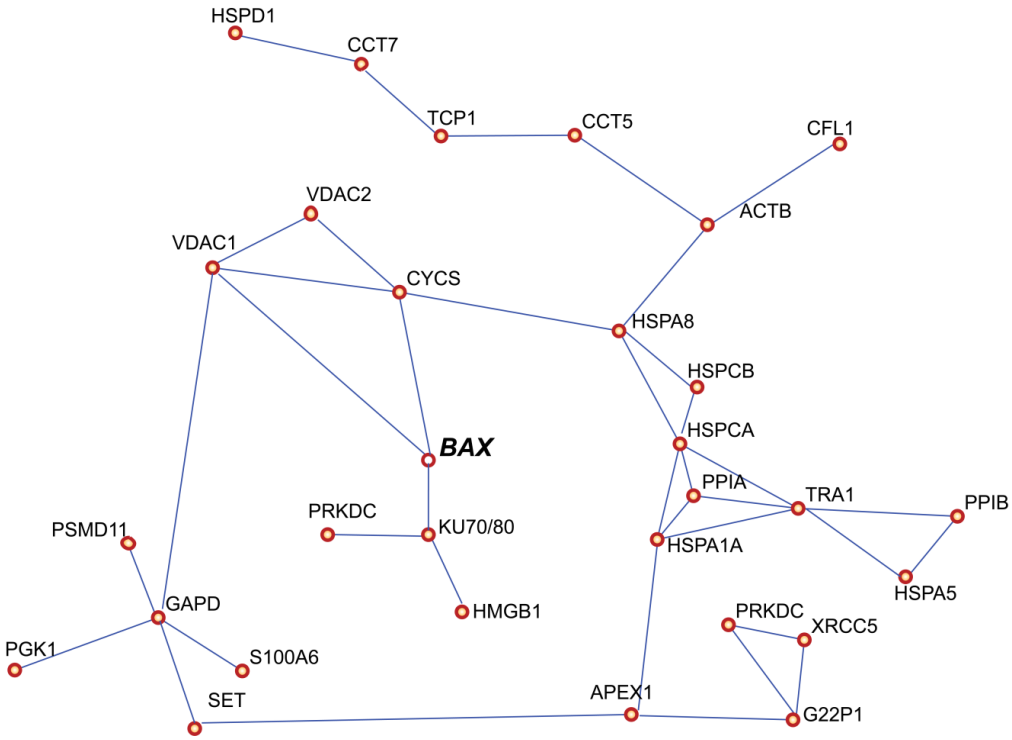


Figure 3.4 Biological network analysis of identified apoptosis-related proteins using the sparse (direct) interaction algorithm

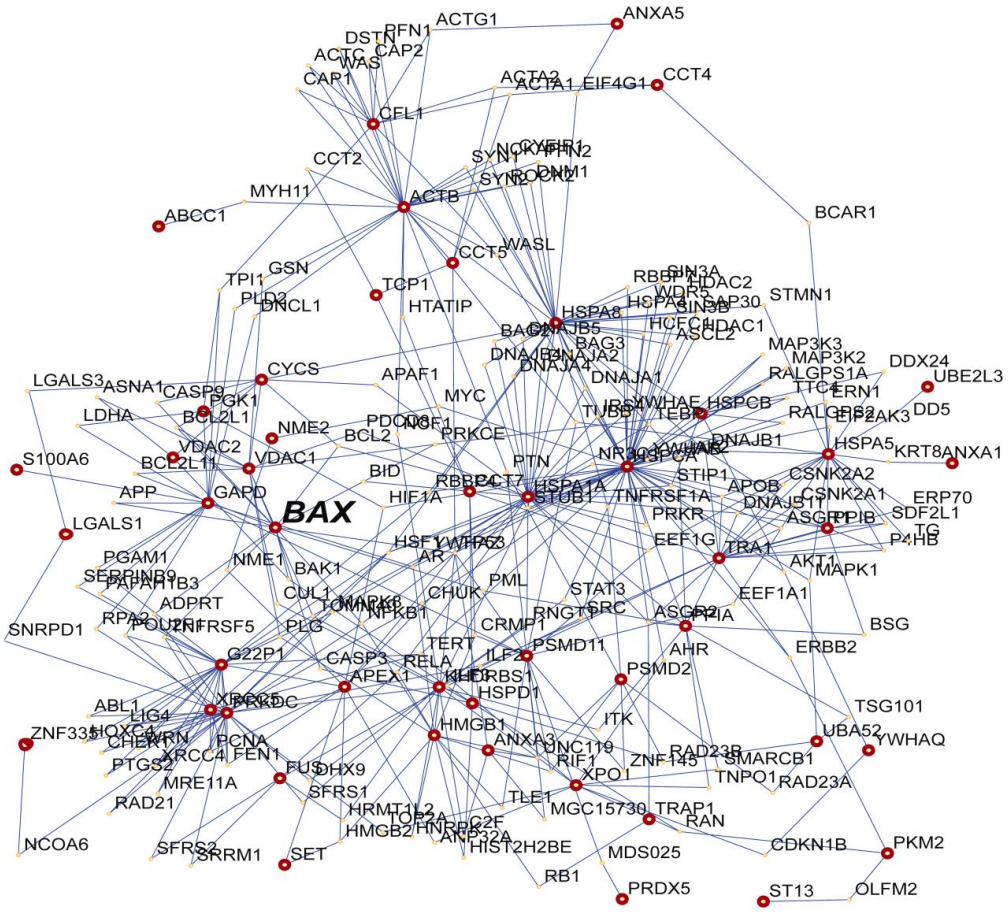


Figure 3.5 Biological network analysis of identified apoptosis-related proteins using the bridge (indirect) interaction algorithm

interaction analysis suggests for the first time that the absence of Bax results in changes in the expression levels of Bax signaling network proteins.

3.3.6 Biological validation of differentially expressed proteins between $Bax^{+/-}$ and $Bax^{-/-}$ clones by Western blot analysis

Eight of the proteins differentially expressed in the current study were subjected to biological validation by Western blot analysis: VDAC1, VDAC2, 14-3-3 theta, MIF, HSP70, HSP90β, HSP60, and LRPPRC. Since the major function of Bax is involved in apoptotic cell death, most of these proteins were selected because of their involvement in apoptotic signaling pathways. VDAC1, VDAC2, 14-3-3 theta, MIF, HSP70, HSP60, HSP90β and LPPRC were found to be down-regulated, which is qualitatively consistent with the mass spectrometric data. HSP90β was found to be down-regulated in the $Bax^{-/-}$ clone as well (Figure 3.6).

3.4 Discussion

Bax plays a crucial role in multiple processes, including cell cycle regulation, stress regulation, detoxification, and especially apoptosis. In previous reports, $Bax^{-/-}$ cells are resistant to multiple stimuli-induced apoptosis, including radiation and DNA damage drugs. In our research, even though Bak and other pro-apoptotic factors remain, the $Bax^{-/-}$ clone is still completely resistant to TRAIL-induced apoptosis. These studies indicate that Bax not only works as an apoptotic executor but also regulates different cellular processes by interacting with other proteins. To further investigate these potential interactions, we used a targeted quantitative MS method to identify the differentially expressed protein profile between $Bax^{+/-}$ and $Bax^{-/-}$ clones.

3.4.1 Targeted Quantitative MS

With the vast majority of proteins remaining generally unchanged, a targeted mass spectrometric approach selectively analyzes only differentially expressed proteins. The forward and reverse labelling strategy was chosen in this work to minimize quantification inaccuracies. Due to the overall complexity of most biological samples, even simple shotgun proteomics identification experiments can have a

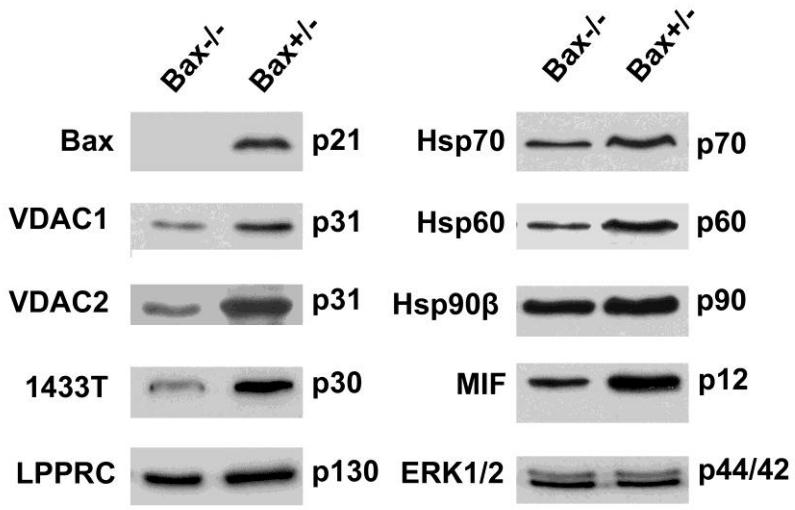


Figure 3.6 Western blot validation of selected differentially expressed proteins between $Bax^{+/-}$ and $Bax^{-/-}$ clones

significant degree of variability in terms of the peptides and proteins identified. Although this study design effectively doubles the instrument time used, the increased reliability in the identification and quantification data determines genuine changes in the $Bax^{+/-}$ and $Bax^{-/-}$ proteome profiles by applying stringent conditions for both identification and quantification. The 2MEGA protocol was selected since it: 1) is a global mass-tagging strategy for quantification by MS; 2) uses MS-based quantification that allows candidate selection prior to sequencing; 3) is readily amenable with LC-MALDI for cross-system comparison; and 4) the reagents used are inexpensive. Figure 3.3 shows the advantage of using the forward and reverse labelling experimental design to confirm quantification values. Although the candidate peaks were selected using data analysis software, overlaying MS spectra from complementary mixtures allows for facile manual confirmation of differential expression in peptide pairs. This provides a simple check to determine that the peptides have similar retention behavior and that the quantification accuracy has not been compromised by neighboring peaks.

3.4.2 Isotope Effect and LC-MALDI Fractionation

Reliable quantification from MS scans is complicated by differing elution profiles arising from the isotope effect and the online fractionation process of LC-MALDI. Even in cases where the isotope effect is small, high abundance peptides will elute over multiple consecutive fractions. For pairs and single-peaks that eluted over multiple fractions (SCX, RPLC, or both), the pair or single-peak with the highest intensity in the MS scans was selected as the ratio for that peptide. It was found that there was no advantage between using the aforementioned method versus weighting the ratios with the observed intensities over multiple fractions. Since the isotope effect from the dimethylation on RPLC retention time shift is between 1 and 3 seconds and the peak width for most peptides in RPLC is ~15 seconds, the overall elution profiles are quite similar. Any variation from the isotope effect is effectively reduced by the fractionation process in LC-MALDI which averages peptide elution over 20 seconds.

Figure 3.3 also highlights the unique feature of 2D-LC MALDI MS for targeted quantification of peptides using the forward and reverse labelling strategy, compared to 2D-LC electrospray ionization (ESI) MS. The isotope labelled peptide pair,

HELQANCYEEVK, while present in the same RPLC fraction from the A_HB_L and A_LB_H mixtures, was found in neighboring SCX fractions. This is not uncommon when using SCX LC for separation and fractionation of a proteome digest, as SCX usually does not provide good chromatographic resolution and small variations in retention time from running the A_HB_L and A_LB_H mixtures may result in the collection of the same peptide in different SCX fractions. Overlaying the MS spectra for these identified peaks shows the expected reversed ratios between the light and heavy isotope labelled samples in the complementary systems. By fractioning onto MALDI plates, samples are "stored" on plates while the MS data analysis can be performed to find peptides that may vary slightly in their *m/z* ratio or chromatographic behavior, but are similar enough to be considered matches. In contrast, samples run with ESI-MS are consumed during analysis, so identifying candidates subject to the same parameters would require re-running the same sample several times. Thus, in 2D-LC ESI-MS, the MS and MS/MS data acquisition is preferably done during the same run, while in 2D-LC MALDI MS, the MS data acquisition on a MALDI plate can be first carried out to determine relative abundance changes of peak pairs, followed by MS/MS of targeted peptide pairs with their abundance changes of greater than a certain threshold (i.e., 1.5-fold).

3.4.3 Reproducibility

Ideally the relative ratio of a peptide pair determined from the forward labelled mixture is the reciprocal of the ratio determined from the reverse labelled mixture. Any deviation from the reciprocal relation can be mainly attributed to experimental variations. Quantification reproducibility, measured as the relative deviation from the mean for a matched pair in the two mixtures, was found to deteriorate under certain conditions. In cases where one peak was significantly higher than the other, quantification values were skewed towards higher values. Since the peak selection algorithm calculates the baseline using a rolling average of data points, a relatively high abundance peak artificially increases the calculated baseline at the nearby corresponding pair, ultimately reducing the signal-to-noise ratio of the lower intensity peak. However, the overall qualitative nature of the data from both the pairs and peaks quantification data are generally internally consistent, with only two cases where the ratios in both experimental systems were contradictory after manual data

analysis. With improvements in peak picking software in the future, the manual data analysis step may be minimized. In both the pairs and peaks data, greater than threefold relative changes in expression levels can be difficult to quantify accurately, but are simple to identify with a high degree of discrimination.

While observed peptide ratios were generally consistent between the two runs, relating the peptide quantification values to overall protein abundance levels was complicated by a few factors. When peptides are clustered to form a list of identified proteins, contradicting ratios can be observed. While some peptides are specific for a particular homolog, others may represent an average of several related proteins. Protein homologues producing the same tryptic peptide will result in an abundance-weighted quantification average of all isoforms. Without additional protein level information, "shared" peptides cannot act as sole indicators of protein abundance. Heat shock protein beta (HSP90β) is one such example. The unique HSP90β peptide ALLFIPR suggests that it is relatively down-regulated in the $Bax^{-/-}$ clone, peptides common to the larger heat shock protein family provide evidence for both relative over- and under-expression. Since the majority of peptides strongly suggested HSP90β down-regulated, its relative expression level was examined by Western blot. In accordance with the MS data, it was found to be down-regulated in the $Bax^{-/-}$ clone.

A second consideration for peptide to protein level quantification is the effect of post-translational modifications. If a protein is present in equal amounts in two samples, differences in the extent of modification will ultimately result in abundance differences of the unmodified tryptic peptide. It is possible that the observed changes in the identified peptides may not be due to differential expression/degradation, but alterations in modification behavior. One potential indicator of this phenomenon was the identification of peptides from various histones, which are known to have both frequent and diverse post-translational modifications. Despite the general shortcomings of peptide to protein level quantification, identification of peptides with changing concentrations can highlight candidate proteins for further studies of regulation, degradation, and modification.

Overall, the data present a consistent profile of the differentially expressed proteins between the $Bax^{+/-}$ and $Bax^{-/-}$ clones. The confirmation of eight candidates by Western blot analysis increases confidence in the MS quantification results. Seven of the proteins (VDAC1, VDAC2, 14-3-3 theta, MIF, HSP70, HSP60, and LRPPRC) which were determined to be down-regulated by MS were clearly observed to be down-regulated in the Western blots. HSP90β was believed to be down-regulated in the $Bax^{-/-}$ clone from the MS results, and was shown to be down-regulated via the Western blot results. The contradictory values for HSP90β arise from sequence similarity between various proteins in the heat shock family. Since most of the peptide ratios were down-regulated, especially those unique to HSP90β, the MS-determined down-regulation is considered consistent with the Western blot result.

3.4.4 Bioinformatics Analysis

Our methods identified a total of 200 differentially expressed proteins. These proteins can be grouped using HiMAP and Metacore (see Figures 3.4, 3.5, and 3.7), which generate visual networks, based on the predicted and known protein-protein interactions. Now, it has been well recognized that Bax plays a crucial pro-apoptotic role by changing MMP in response to various stimuli, although it is still unknown how activated or oligomerized Bax induces MMP changes.[37] Our analysis primarily focussed on the apoptosis related proteins, and four major groups of proteins of biological significance are highlighted based on the analysis results.

MPT channel proteins, including VDAC-1, VDAC-2, and cytochrome C, are down-regulated more than 2 fold in the $Bax^{-/-}$ clone. Related to these MPT channel proteins, we found that glyceraldehyde-3-phosphate dehydrogenase (GADPH),[38] a VDAC1 interacting partner, was up-regulated in the $Bax^{-/-}$ clone, and its interacting proteins, including phosphoglycerate kinase 1 (predicted interaction), Protein SET [39], and 26S proteasome non-ATPase regulatory subunit 11 (yeast two-hybrid dataset) were down-regulated which may come from negative feedback regulation (Figure 3.4). These data suggest that $Bax^{-/-}$ cells become resistant to most apoptotic stimuli, not only because of the loss of Bax, but also due to down-regulation of MPT channel proteins that control their apoptotic potential. However, since Bax significantly

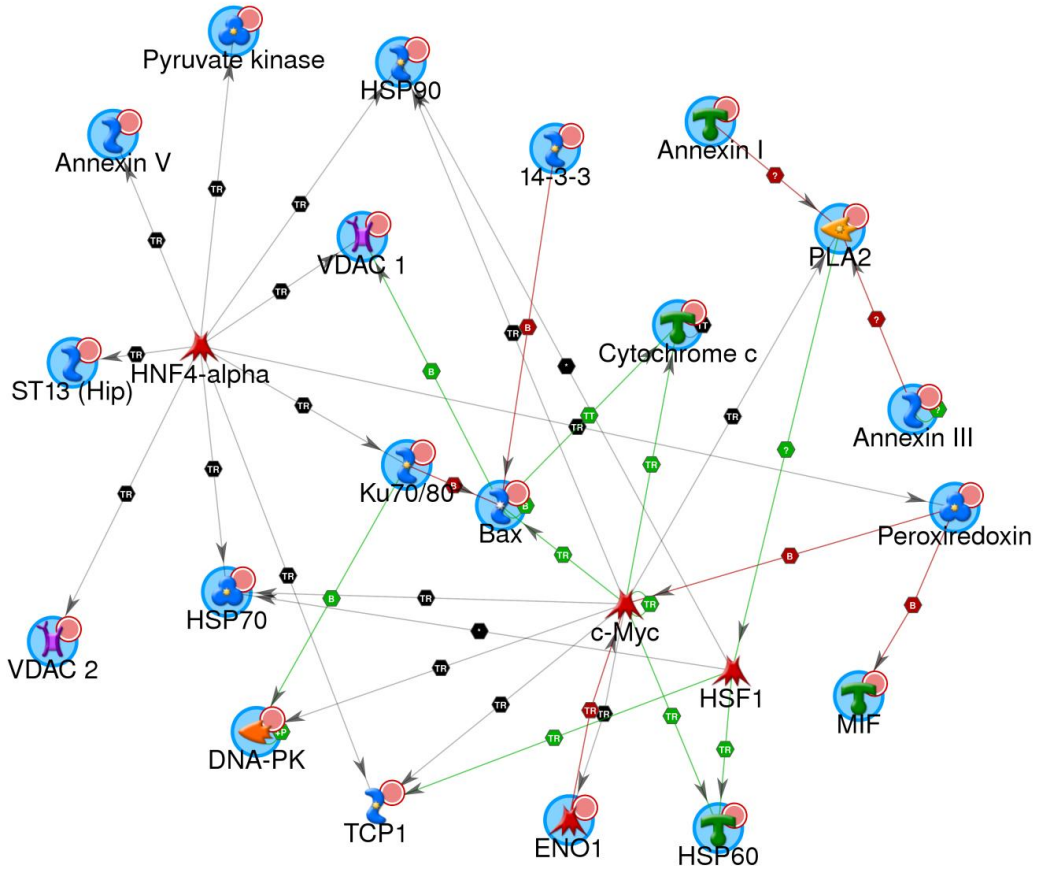


Figure 3.7 Bioinformatics analysis of the four crucial groups of differentially expressed proteins between Bax$^{+/-}$ and Bax$^{-/-}$ clones analyzed by using shortest pathway algorithm

interacts with MPT channel proteins under apoptotic stimuli, how these proteins were down-regulated in the Bax$^{-/-}$ clone still need to be explored further.

Bax activation has been reported to be regulated by various factors, and some of these regulators were identified down-regulated in the Bax$^{-/-}$ clone in our experiments including DNA-PK complex, MIF, and 14-3-3 theta**.** DNA-PK complex is composed of three proteins, including KU70, KU86, and PRKDC, all of which were found to be down-regulated more than 2.5-fold in the Bax$^{-/-}$ clone. Bax has been reported to associate with and promote KU70 to move into the nucleus, releasing Bax to be activated in the cytosol under various apoptotic stimuli[21, 40-44] (Figure 3.4A). Corresponding to the down-regulation of DNA-PK complex, its DNA-binding regulatory component, high mobility group protein B1 (HMGB1), was found down-regulated in the Bax$^{-/-}$ clone (Figure 4A) [45]. Other two Bax regulators, MIF[46] and 14-3-3 theta,[22, 47] were also found down-regulated in the Bax$^{-/-}$ clone. Both proteins have been reported to interact with Bax to inhibit Bax-driven cell death[22, 46, 47]. It is possible that these Bax regulators were down-regulated because of the loss of their regulating target, Bax.

In addition to the above proteins, two other groups of proteins, heat shock protein (Hsp) family members and oxidative stress-triggered proteins also drew our attention. Within these Hsp family members, HSP60,[48] T-complex protein-1 (TCP-1), HSP70, HSP90α, and HSP90β, were found down-regulated, and C-myc promoter-binding protein (MBP-1, ENO1)[49] was found up-regulated in the Bax$^{-/-}$ clone**.** Within oxidative stress-triggered proteins, peroxiredoxins[50], pyruvate kinase[51], and annexins[52, 53] were found down-regulated in the Bax$^{-/-}$ clone. At the same time, S100A6, a member of the S100 family of proteins expressed in colorectal carcinoma, was found up-regulated in the Bax$^{-/-}$ clone[54]. These two groups of proteins have been reported to be highly involved in cell proliferation and apoptosis pathways[55-61]. Although little is known so far regarding the potential interactions between Bax and these two group of proteins, it is likely that the differentially expression of these proteins make Bax$^{-/-}$ cells more difficult to kill under oxidative stress or other apoptotic stimuli.

3.5 Conclusions

Bax plays a central role in various stimuli-induced apoptosis pathways. Even though Bak and other apoptotic executors remain in the $Bax^{-/-}$ clone, it is not clear why the $Bax^{-/-}$ clone is resistant to most apoptosis stimuli, including DNA-damage drugs and death ligands. In this study, we used a newly developed targeted quantitative mass spectrometry analysis strategy which greatly decreases the percentage of false information, since we introduced an internal reference (A_LB_H vs. A_HB_L) to identify and analyze the differentially expressed proteins. Using this targeted quantitative MS proteomic analysis with 2MEGA isotope labelling and 2D-LC MALDI MS and MS/MS, we identified a total of 200 proteins expressed differentially between the wild type and Bax knockout HCT116 clone, including proteins involved in apoptosis, DNA repair, stress-induced proteins, cell cycle regulation proteins, cytoskeletal rearrangements, and signal transduction molecules. Within these proteins, four groups of proteins are highlighted here because of their important roles in the Bax-modulated apoptosis pathway. First, MPT channel proteins, including VDAC-1, VDAC-2, and cytochrome c, are significantly down-regulated in the $Bax^{-/-}$ clone. Second, Bax-regulator proteins, such as KU70, KU80, PRKDC, 14-3-3 theta, and MIF, which directly or indirectly interact with Bax to modulate Bax-mediated apoptosis, were found down-regulated in the $Bax^{-/-}$ clone. The third group of proteins was the heat shock protein family members, including HSP60, HSP70, TCP-1, HSP90α, and HSP90β. The fourth group was the oxidative stress-triggered proteins, peroxiredoxins, pyruvate kinase, S100A6, and annexins. These proteins play both pro- and anti-apoptosis roles. These data indicate that Bax functions not only as an apoptotic signal executor by modulating the damage of mitochondria membrane potential, but also as a regulator for the expression level of other proteins. By modulation of those crucial groups of proteins, the $Bax^{-/-}$ clone further lost its apoptotic potential. Our findings through the quantitative MS analysis enabled us to draw a detailed protein profile map with the Bax as the central protein. To our knowledge, the expression of most of these proteins has not been reported to be related to Bax and thus novel targets for further study of the Bax-modulated cell signaling have been provided.

3.6 Literature Cited

(1) Vogelstein, B.; Kinzler, K. W. *Nat Med* **2004**, *10*, 789-799.

(2) Ashkenazi, A. *Nat Rev Cancer* **2002**, *2*, 420-430.

(3) Kroemer, G.; Reed, J. C. *Nat Med* **2000**, *6*, 513-519.

(4) Debatin, K. M.; Poncet, D.; Kroemer, G. *Oncogene* **2002**, *21*, 8786-8803.

(5) Wiley, S. R.; Schooley, K.; Smolak, P. J.; Din, W. S.; Huang, C. P.; Nicholl, J. K.; Sutherland, G. R.; Smith, T. D.; Rauch, C.; Smith, C. A.; Goodwin, R. G. *Immunity* **1995**, *3*, 673-682.

(6) Pitti, R. M.; Marsters, S. A.; Ruppert, S.; Donahue, C. J.; Moore, A.; Ashkenazi, A. *J Biol Chem* **1996**, *271*, 12687-12690.

(7) Ashkenazi, A.; Pai, R. C.; Fong, S.; Leung, S.; Lawrence, D. A.; Marsters, S. A.; Blackie, C.; Chang, L.; McMurtrey, A. E.; Hebert, A.; DeForge, L.; Koumenis, I. L.; Lewis, D.; Harris, L.; Bussiere, J.; Koeppen, H.; Shahrokh, Z.; Schwall, R. H. *J Clin Invest* **1999**, *104*, 155-162.

(8) Walczak, H.; Miller, R. E.; Ariail, K.; Gliniak, B.; Griffith, T. S.; Kubin, M.; Chin, W.; Jones, J.; Woodward, A.; Le, T.; Smith, C.; Smolak, P.; Goodwin, R. G.; Rauch, C. T.; Schuh, J. C.; Lynch, D. H. *Nat Med* **1999**, *5*, 157-163.

(9) Gajewski, T. F. *J Clin Oncol* **2007**, *25*, 1305-1307.

(10) Bodmer, J. L.; Holler, N.; Reynard, S.; Vinciguerra, P.; Schneider, P.; Juo, P.; Blenis, J.; Tschopp, J. *Nat Cell Biol* **2000**, *2*, 241-243.

(11) Xiao, C.; Yang, B. F.; Asadi, N.; Beguinot, F.; Hao, C. *J Biol Chem* **2002**, *277*, 25020-25025.

(12) Medema, J. P.; Scaffidi, C.; Kischkel, F. C.; Shevchenko, A.; Mann, M.; Krammer, P. H.; Peter, M. E. *EMBO J* **1997**, *16*, 2794-2804.

(13) Li, H.; Zhu, H.; Xu, C. J.; Yuan, J. *Cell* **1998**, *94*, 491-501.

(14) Luo, X.; Budihardjo, I.; Zou, H.; Slaughter, C.; Wang, X. *Cell* **1998**, *94*, 481-490.

(15) Du, C.; Fang, M.; Li, Y.; Li, L.; Wang, X. *Cell* **2000**, *102*, 33-42.

(16) Verhagen, A. M.; Ekert, P. G.; Pakusch, M.; Silke, J.; Connolly, L. M.; Reid, G. E.; Moritz, R. L.; Simpson, R. J.; Vaux, D. L. *Cell* **2000**, *102*, 43-53.

(17) Deng, Y.; Lin, Y.; Wu, X. *Genes Dev* **2002**, *16*, 33-45.

(18) Li, P.; Nijhawan, D.; Budihardjo, I.; Srinivasula, S. M.; Ahmad, M.; Alnemri, E. S.; Wang, X. *Cell* **1997**, *91*, 479-489.

(19) Kuwana, T.; Mackey, M. R.; Perkins, G.; Ellisman, M. H.; Latterich, M.; Schneiter, R.; Green, D. R.; Newmeyer, D. D. *Cell* **2002**, *111*, 331-342.

(20) Baines, C. P.; Kaiser, R. A.; Sheiko, T.; Craigen, W. J.; Molkentin, J. D. *Nat Cell Biol* **2007**, *9*, 550-555.

(21) Mazumder, S.; Plesca, D.; Kinter, M.; Almasan, A. *Mol Cell Biol* **2007**, *27,* 3511-3520.

(22) Nomura, M.; Shimizu, S.; Sugiyama, T.; Narita, M.; Ito, T.; Matsuda, H.; Tsujimoto, Y. *J Biol Chem* **2003**, *278*, 2058-2065.

(23) Porter, G. W.; Khuri, F. R.; Fu, H. *Semin Cancer Biol* **2006**, *16*, 193-202.

(24) Rampino, N.; Yamamoto, H.; Ionov, Y.; Li, Y.; Sawai, H.; Reed, J. C.; Perucho, M. *Science* **1997**, *275*, 967-969.

(25) Zhang, L.; Yu, J.; Park, B. H.; Kinzler, K. W.; Vogelstein, B. *Science* **2000**, *290*, 989-992.

(26) LeBlanc, H.; Lawrence, D.; Varfolomeev, E.; Totpal, K.; Morlan, J.; Schow, P.; Fong, S.; Schwall, R.; Sinicropi, D.; Ashkenazi, A. *Nat Med* **2002**, *8*, 274-281.

(27) Wang, S.; El-Deiry, W. S. *Proc Natl Acad Sci U S A* **2003**, *100*, 15095-15100.

(28) Ji, C.; Li, L.; Gebre, M.; Pasdar, M. *J Proteome Res* **2005**, *4*, 1419-1426.

(29) Ji, C.; Guo, N.; Li, L. *J Proteome Res* **2005**, *4*, 2099-2108.

(30) Young, J. B.; Li, L. *J Am Soc Mass Spectrom* **2006**, *17*, 325-334.

(31) Arnoult, D.; Parone, P.; Martinou, J. C.; Antonsson, B.; Estaquier, J.; Ameisen, J. C. *J Cell Biol* **2002**, *159*, 923-929.

(32) Young, J. B.; Li, L. *Anal Chem* **2007**, *79*, 5927-5934.

(33) Ekins, S.; Nikolsky, Y.; Bugrim, A.; Kirillov, E.; Nikolskaya, T. *Methods Mol Biol* **2007**, *356*, 319-350.

(34) Rhodes, D. R.; Tomlins, S. A.; Varambally, S.; Mahavisno, V.; Barrette, T.; Kalyana-Sundaram, S.; Ghosh, D.; Pandey, A.; Chinnaiyan, A. M. *Nat Biotechnol* **2005**, *23*, 951-959.

(35) Hoffmann, J.; Vitale, I.; Buchmann, B.; Galluzzi, L.; Schwede, W.; Senovilla, L.; Skuballa, W.; Vivet, S.; Lichtner, R. B.; Vicencio, J. M.; Panaretakis, T.; Siemeister, G.; Lage, H.; Nanty, L.; Hammer, S.; Mittelstaedt, K.; Winsel, S.;

Eschenbrenner, J.; Castedo, M.; Demarche, C.; Klar, U.; Kroemer, G. *Cancer Res* **2008**, *68*, 5301-5308.

(36) Lim, D. Y.; Park, J. H. *Am J Physiol Gastrointest Liver Physiol* **2009**, *296*, G1060-1068.

(37) Wong, W. W.; Puthalakath, H. *IUBMB Life* **2008**, *60*, 390-397.

(38) Tarze, A.; Deniaud, A.; Le Bras, M.; Maillier, E.; Molle, D.; Larochette, N.; Zamzami, N.; Jan, G.; Kroemer, G.; Brenner, C. *Oncogene* **2007**, *26*, 2606-2620.

(39) Carujo, S.; Estanyol, J. M.; Ejarque, A.; Agell, N.; Bachs, O.; Pujol, M. J. *Oncogene* **2006**, *25*, 4033-4042.

(40) Sawada, M.; Hayes, P.; Matsuyama, S. *Nat Cell Biol* **2003**, *5*, 352-357.

(41) Sawada, M.; Sun, W.; Hayes, P.; Leskov, K.; Boothman, D. A.; Matsuyama, S. *Nat Cell Biol* **2003**, *5*, 320-329.

(42) Yoo, C. B.; Jones, P. A. *Nat Rev Drug Discov* **2006**, *5*, 37-50.

(43) Sawada, M.; Sun, W.; Hayes, P.; Leskov, K.; Boothman, D. A.; Matsuyama, S. *Nature cell biology.* **2003**, *5*, 320-329.

(44) Subramanian, C.; Opipari, A. W., Jr.; Bian, X.; Castle, V. P.; Kwok, R. P. *Proc Natl Acad Sci U S A* **2005**, *102*, 4842-4847.

(45) Yumoto, Y.; Shirakawa, H.; Yoshida, M.; Suwa, A.; Watanabe, F.; Teraoka, H. *J Biochem (Tokyo)* **1998**, *124*, 519-527.

(46) Baumann, R.; Casaulta, C.; Simon, D.; Conus, S.; Yousefi, S.; Simon, H. U. *Faseb J* **2003**, *17*, 2221-2230.

(47) Porter, G. W.; Khuri, F. R.; Fu, H. *Seminars in cancer biology.* **2006**, *16*, 193-192.

(48) Kirchhoff, S. R.; Gupta, S.; Knowlton, A. A. *Circulation* **2002**, *105*, 2899-2904.

(49) Perconti, G.; Ferro, A.; Amato, F.; Rubino, P.; Randazzo, D.; Wolff, T.; Feo, S.; Giallongo, A. *Biochim Biophys Acta* **2007**, *1773*, 1774-1785.

(50) Immenschuh, S.; Baumgart-Vogt, E. *Antioxid Redox Signal* **2005**, *7*, 768-777.

(51) Aisaki, K.; Aizawa, S.; Fujii, H.; Kanno, J.; Kanno, H. *Exp Hematol* **2007**, *35*, 1190-1200.

(52) Buckingham, J. C.; John, C. D.; Solito, E.; Tierney, T.; Flower, R. J.; Christian, H.; Morris, J. *Ann N Y Acad Sci* **2006**, *1088*, 396-409.

(53) Lim, L. H.; Pervaiz, S. *FASEB J* **2007**, *21*, 968-975.

(54) Komatsu, K.; Andoh, A.; Ishiguro, S.; Suzuki, N.; Hunai, H.; Kobune-Fujiwara, Y.; Kameyama, M.; Miyoshi, J.; Akedo, H.; Nakamura, H. *Clin Cancer Res* **2000**, *6*, 172-177.

(55) Samali, A.; Cai, J.; Zhivotovsky, B.; Jones, D. P.; Orrenius, S. *Embo J* **1999**, *18*, 2040-2048.

(56) Leroux, M. R.; Candido, E. P. *Biochem Biophys Res Commun* **1997**, *241*, 687-692.

(57) Gotoh, T.; Terada, K.; Oyadomari, S.; Mori, M. *Cell Death Differ* **2004**, *11*, 390-402.

(58) Pandey, P.; Saleh, A.; Nakazawa, A.; Kumar, S.; Srinivasula, S. M.; Kumar, V.; Weichselbaum, R.; Nalin, C.; Alnemri, E. S.; Kufe, D.; Kharbanda, S. *Embo J* **2000**, *19*, 4310-4322.

(59) Calderwood, S. K.; Khaleque, M. A.; Sawyer, D. B.; Ciocca, D. R. *Trends Biochem Sci* **2006**, *31*, 164-172.

(60) Ray, R. B. *Cell Growth Differ* **1995**, *6*, 1089-1096.

(61) Ghosh, A. K.; Majumder, M.; Steele, R.; Liu, T. J.; Ray, R. B. *Oncogene* **2002**, *21*, 2775-2784.

Chapter 4 – Automation of 2MEGA Labelling Chemistry for High Throughput Proteomics Applications

4.1 Introduction

Mass spectrometric proteomics approaches have experienced a marked shift from identifying proteins to quantifying thousands of peptides from complex matrices to generate detailed quantitative information about proteome changes. The ability of mass spectrometric based methods to both identify and quantify thousands of components in a single experiment positions it uniquely within the repertoire of techniques available to researchers interested in a variety of biological processes and phenomena. With careful experimental design, quantitative information about alterations in a proteome resulting from a given perturbation or organism state can be obtained for a variety of cellular processes, such as phosphorylation,[1] acetylation,[2] glycosylation,[3] and protein production/degradation.[4] Although MS-based approaches can provide substantial amounts of data, increasing demands on the overall productivity of liquid chromatography mass spectrometry-based (LC-MS) workflows for proteomics analysis remains, particularly for commercial or pharmaceutical applications. Methods amenable to automation or reducing overall analyst intervention may improve throughput by reducing the time required for sample processing.

Currently, multiple reaction monitoring (MRM) based methods are widely used for quantitative monitoring of previously identified peptide sequences in validation studies where the study system is relatively well defined.[5] However, discovery-based LC-MS platforms may become commonplace as the detection sensitivity of modern instrumentation is improved and both vendor-supplied and third-party software suites seamlessly integrate quantification based capabilities into their existing products. One of the main limitations of shotgun proteomics based LC-MS studies is the undersampling problem: the significant complexity of a protein digest mixture leads to incomplete identification of the constituent proteins. Various studies have shown that re-analysis of an identical sample generates about a 25 to 30% increase in identified

peptides,[6, 7] depending on the origin of the sample and its complexity. Among the identifiable peptides, which exclude those of low intensity or poor fragmentation behaviour, nearly complete identification is reached after triplicate analysis. To improve the overall data quality of discovery-based experiments, it may be advantageous to re-analyze the same sample multiple times (technical replicates) to improve coverage. Alternatively, sample replicates can be analyzed, not only to improve peptide identification coverage, but also to address sample handling and measurement variation affecting quantification accuracy and to improve counting statistics for redundant peptides. Common to both of these potential solutions is that additional or multiple samples need to be prepared for analysis.

While a range of approaches for MS-based quantification have been developed, they can be broadly categorized as either label-based or label-free, depending on whether or not isotopes are introduced for quantification. Regardless of the strategy employed, both general methods offer their own advantages. Label-free methods typically use additional information from identified peptides across multiple runs, such as ion current intensity or frequency of MS/MS sequencing, to determine relative changes between various samples. Label-based methods utilize relative signal intensities from isotopically-encoded references. Within the realm of label-based methods, various metabolic and chemical isotope incorporation methods exist alongside targeted approaches using standard addition of synthetically prepared isotopically labelled peptides. The introduction of isotopes by metabolic or chemical derivatization methods is an additional experimental procedure that can be a source of variation in the observed ratio between samples in LC-MS experiments.

For metabolic introduction methods, such as stable isotope labelling by cell culture (SILAC),[8] care is taken to ensure that the samples are grown on isotopically enriched media for a sufficient duration to minimize contributions from pre-existing unlabelled material in the sample and to correct for less than complete incorporation. Furthermore, conversion of arginine to proline has been addressed with different biological strategies.[9, 10] Metabolic labelling strategies offer the advantage that isotopes are introduced at the very beginning of the sample preparation workflow,

allowing samples to be mixed early in the sample preparation workflow. Early mixing ultimately minimizes variation from downstream sample preparation and allows protein-level separation methods, such as isoelectric focussing or SDS-PAGE.[11] Despite these attractive advantages, the cost of SILAC experiments can be prohibitive for high throughput applications and samples may not originate from a source for which metabolic labelling is feasible.

Chemical derivatization approaches are universally applicable to samples regardless of origin. However, considerable care must be taken to reduce variation in the experimental steps before samples are combined for analysis. Any variation prior to sample mixing introduces changes not reflective of the genuine differences between samples. Derivatization schemes typically react protein digests using isotopically-encoded variants of the same reagent. As with metabolic methods, differences in the isotopic purity of the labels can be corrected during data processing. However, variations in label incorporation can be nearly impossible to correct, since it is difficult to accurately estimate the conversion efficiency. While complete label incorporation is ideal, incomplete labelling can be tolerated so long as consistent reaction performance is achieved.

Here, we describe our efforts to standardize a previously reported labelling chemistry for quantitative proteomics experiments. Using the 2MEGA protocol[12] (dimethylation after guanidinylation), an automated differential isotopic labelling method utilizing a commercial liquid handler is described to minimize variability from sample handling during the labelling reaction for high throughput applications. The 2MEGA protocol produces peptides with a fixed mass shift when used for labelling experiments and has been previously shown to increase the percentage of lysine containing peptides observed.[13] Furthermore, the comparatively low cost of the isotopically labelled reagent allows the automated method to be used for processing multiple samples. The reaction conditions were optimized for labelling of simple protein mixtures and complex tryptic digests of *E. coli*. Both front-end sample preparation methods and post-labelling workup are discussed. Potential side reactions, functional sample concentration ranges, and method limitations are also considered.

4.2 Experimental

4.2.1 Chemicals and Reagents

LC-MS grade water, methanol, acetonitrile, and ProteaseMAX™ were obtained from Fisher Scientific (Edmonton, AB). *O*-methylisourea hemisulfate, formaldehyde, sodium cyanoborohydride, triacetoxyborohydride, pyridine-borane complex in THF, borane-THF complex, 2-picoline borane, urea, Cellytic™ M cell lysis buffer, LC-MS grade formic acid and trifluoroacetic acid were obtained from Sigma Aldrich (Oakville, ON). Rapigest™ was purchased from Waters (Milford, MA). Anionic acid labile surfactant was obtained from Canadian Life Sciences (Peterborough, ON). Total protein extraction kit was purchased from Biochain Institute (Hayward, CA).

4.2.2 Protein Sample Preparation

E. coli K12 digest was prepared by culturing cells until OD_{600} = 0.5, lysing cells using an Emulsiflex homogenizer, and precipitating proteins using acetone (1:5, v/v) at -80 °C overnight. Proteins were re-solubilized in 0.1% SDS in water and the protein concentration was estimated by BCA assay. Samples were reduced with dithiothreitol, alkylated with iodoacetamide and digested using a 50:1 ratio (protein:enzyme) of trypsin in 50 mM NH_4HCO_3. After digestion, samples were acidified with 10% formic acid to pH 2 and SDS was removed from the peptide digest by strong cation exchange chromatography. Desalting and peptide quantification was determined using a LC-UV method as previously described.[14] Digests were dried down in a vacuum centrifuge and reconstituted in the appropriate buffers for labelling optimization (200 mM NH_4HCO_3 or 200 mM KH_2PO_4)

4.2.3 Labelling Optimization

The 2MEGA manual labelling method was previously described and used as the basis for labelling optimization.[12] The detailed optimized protocol can be found in the Appendix 1. A Gilson 215 liquid handler with standard racks was used. The only modification was a homebuilt aluminum heating block with a thermocouple for temperature control with a temperature controller (Barnart Scientific). Initial

experiments used tryptic horse myoglobin digest and *E. coli* digest was used for validation of general tryptic mixtures. In brief, peptide solutions (0.5 µg/µL) were adjusted to pH 11 using 2 M NaOH. *O*-methylisourea hemisulfate solution (~3M) was prepared in a 1:1 (v/v) of 2 M NaOH and 1 M Na_2CO_3 (pH 12). The guanidinylation reaction was allowed to proceed before adjustment to pH 7 using 6 M HCl and further adjustment to pH 5 using 1 M acetate buffer. Formaldehyde (4% in H_2O, (w/w)) was added and followed by subsequent addition of 2-picoline borane (1 M in methanol). The dimethylation reaction was allowed to complete before adjusting to pH <2 using 10% TFA. Samples were desalted and quantified by LC-UV prior to mass spectrometric analysis. Experiments simulating various protein solubilization agents were performed by preparing high concentration solutions to spike into samples to the appropriate final concentrations. Samples were prepared and analyzed in at least duplicate, typically triplicate.

4.2.4 Mass Spectrometry and Data Analysis

Samples were separated on a 300 µm i.d. x 150 mm Discovery C_{18} column using a Waters nanoAcquity LC followed by analysis on a Waters ESI-QTOF Premier mass spectrometer. Peaks lists were processed by ProteinLynx and searched using MASCOT (enzyme: trypsin; missed cleavages: 2; fixed modifications: Carbamidomethyl (C); MS tolerance: 30 ppm; MS/MS tolerance: 0.2 Da). Variable modifications for searches included both expected modifications from labelling (Guanidinyl (K, $+CN_2H_2$); Dimethylation (N-term; $+C_2H_4$)) as well as known side reaction products (Guanidinyl_NTerm (N-term, $+CN_2H_2$); Dimethylation_K (K; $+C_2H_4$)). A modified instrument-type setting using standard ESI-QTOF fragmentations further allowing a-ions was used. A sequence database containing only *E. coli* K12 proteins was used for database searching. Calculations to determine the extent and efficiency of labelling were taken as the number of correct peptide identifications divided by the total number of peptide identifications. Incorrect modifications were classified as peptides that had unlabelled groups, N-terminal guanidinylation, or dimethylation at lysine.

4.3 Results and Discussion

Although MS-based proteomics approaches can generate detailed quantitative information from complex proteomic samples, high-throughput processing of samples remains an underserved area of the research pipeline. Here, we describe our efforts to optimize a labelling chemistry method amenable for quantitative MS analysis and consider both upstream and downstream sample processing considerations to produce a workflow with potential applications for automated processing of multiple samples. To allow for the parallel processing of samples with minimal analyst intervention, a commercially available liquid handler was used (Figure 4.1). Dispensing with the liquid handler was performed with a single dispensing head that exchanged pipette tips between solution additions; it is expected that the described method should be applicable for multi-head liquid handling systems. The only non-standard modification used was a homemade thermocouple-controlled aluminum heating block used to accommodate standard microcentrifuge vials and capable of heating to 95 °C for labelling optimization. However, it was found that heating to 37 °C was sufficient and that sample mixing and agitation during the reaction were not required. The minimal equipment setup was selected to allow the method to be easily ported to address the specific demands of alternate dispensing configurations.

The 2MEGA labelling method is a two step labelling procedure using successive selective covalent modifications of the side chain amine of lysine followed by reaction at the primary amine at the N-termini of peptides (Figure 2.1). By limiting introduction of the isotopically coded groups exclusively to free peptide N-termini, a fixed mass shift is observed for all correctly modified peptides. Furthermore, conversion of lysine residues to homoarginine increases the relative proportion of lysine containing peptides identified from LC-MS experiments,[13] likely by increasing the ESI response of lysine containing peptides.[15] As the side chain amine and peptide N-termini are similar, careful and complete modification of the side chain amines is required before proceeding to the second reaction. Labelling method optimization was performed by considering reagent amounts, pH, temperature, and reaction time.

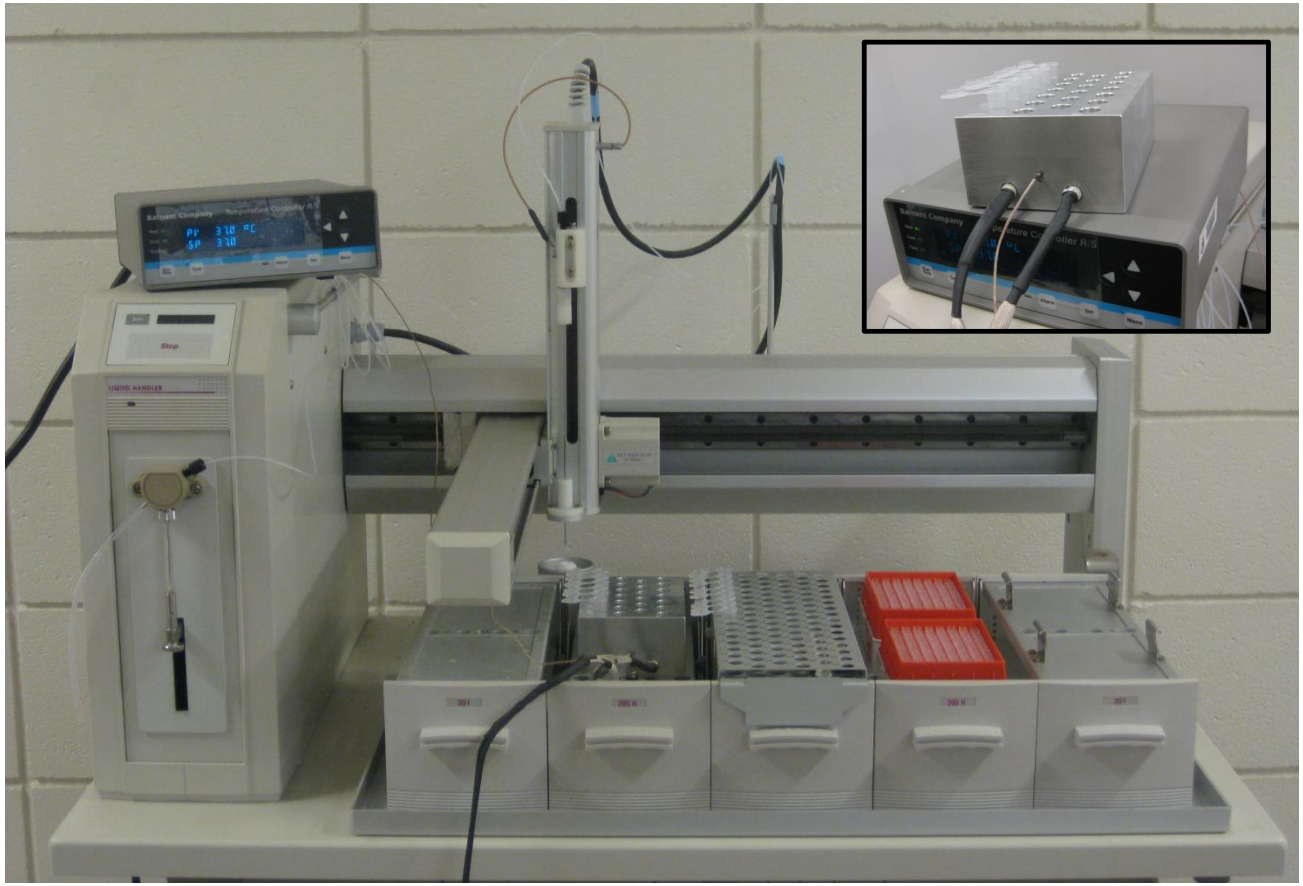


Figure 4.1 Liquid handler for automated 2-MEGA labelling. A commercially available Gilson 215 liquid handler was used for the labelling work. The only modification was a homemade aluminum heating block used to heat the samples (shown in the inset).

4.3.1 Guanidinylation

The first reaction involves conversion of the primary amine side chain of lysine into homoarginine, while leaving the N-terminus unreacted.[16, 17] Although the N-terminus and lysine chain amine are both primary amines, differences in reactivity are observed due to differences in the local steric environment and the presence of the amide group beta to the N-terminus of a peptide. The amide carbonyl reduces the nucleophilicity of the N-terminal amine, which imparts differences in reactivity between the N-terminal amine and lysine side chain. The difference in their chemical behaviour is also evident from differences in their pKas: peptide N-termini have pKas around 8, whereas the lysine side chain is around 10.5. Since reaction of the lysine side chain would require the amine to be deprotonated, pHs around pH 10, 10.5, 11, 11.5, and 12 for the guanidinylation reaction were considered with pH 11.5 found to be optimal. For pHs higher than 12, an increased proportion of guanidinylation at peptide N-termini were observed. At pHs lower than 10.5, the reaction was very slow and would not reach completion even after two hours at temperatures as high as 65 °C.

Two different buffer systems were considered for the digestion and initial reaction step. Ammonium bicarbonate (pKa ~9) was adjusted with NaOH to the bicarbonate/carbonate buffer pair (pKa ~12) for guanidinylation. Similarly, sodium phosphate was also considered since the $H_2PO_4^-/HPO_4^{2-}$ pair buffers around the desired range for digestion (pKa ~8) and the HPO_4^{2-}/PO_4^{3-} pair can be used for guanidinylation (pKa ~11). Guanidinylation in the presence of phosphate often gave incomplete yields, even when increased amounts of *O*-methylisourea or elevated temperatures between 45 to 75 °C were used. The reason for incomplete yields is unclear, but may be due to the interaction of the phosphate groups with the primary amines or charged *O*-methylisourea cation in solution. In order for the guanidinylation reaction to reach completion, a significant molar excess of *O*-methylisourea hemisulfate is required. Given the sample solution buffer concentration (200 mM) and the nearly equal volume of *O*-methylisourea hemisulfate solution (~3 M) used, the overall buffer pH is controlled by the reagent solution. When troubleshooting unacceptable results during the guanidinylation reaction, adjustment of the reagent solution pH using 2 M NaOH was the most effective method for controlling the reaction outcome.

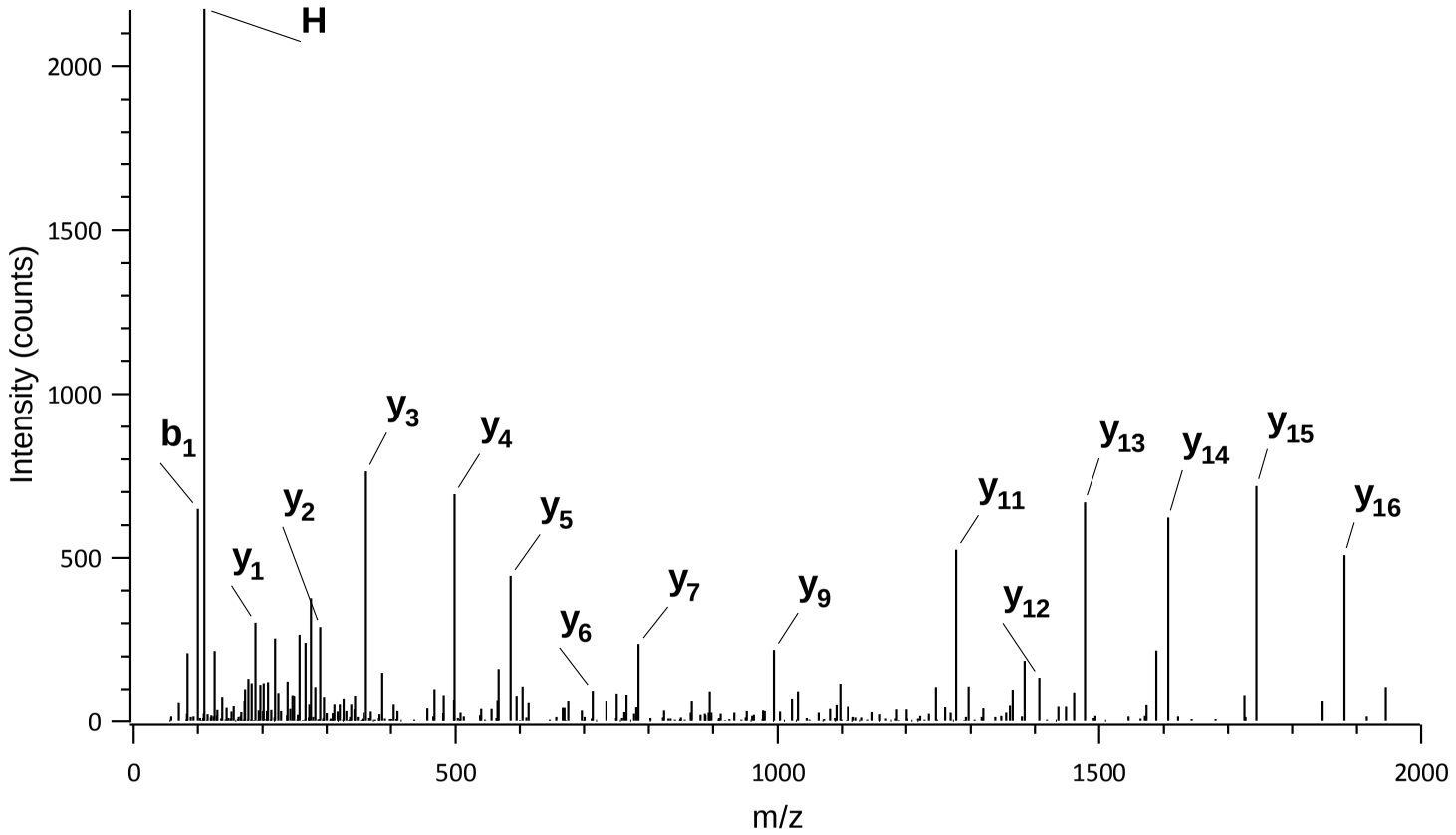


Figure 4.2 MS/MS Spectrum of GHHEAELKPLAQSHATK. The precursor mass of the peptide suggests three guanidino groups attached to the peptide. The MS/MS spectrum provides assignment of one of modifications to the N-terminus, due to the location of the b_1 ion. The nearly completely y-ion ladder also suggests modification at the N-terminus. The strong peak in the low mass region is the immonium ion of histidine.

Even under the optimized conditions, approximately 1-2% of peptides observed will have guanidinylation at the peptide N-terminus, primarily on glycine and alanine N-terminal peptides (Figure 4.2). Excessive reagent amounts (threefold increase) will lead to a slight increase of guanidinylation of the N-terminus (~5% of total identified peptides). Although most proteomics protocols reduce disulfide bonds and alkylate using thiol active reagents such as dithiothreitol or N-ethylmaleimide, if free cysteines are present in the sample, they will become methylated quantitatively ($+CH_2$, +14 Da).[18]

4.3.2 Dimethylation

The dimethylation reaction was found to be robust and only required limited modification from previously reported protocols.[12, 19, 20] Side products were not found and insufficient reagent often resulted in properly labelled and fully unlabelled peptides. When insufficient reagent amounts were used, it was noted that the abundance of monomethylated peptides was often less than unlabelled and dimethylated peptides (Figure 4.3). This observation may be due to the increased nucleophilicity of the monomethylated amine, which allows the second methylation to proceed more readily than the initial methylation. The primary objective was substitution of the toxic sodium cyanoborohydride used for the reduction of the imine formed from the condensation of formaldehyde with the free N-termini of peptides and remaining unreacted lysine side chains. Sodium cyanoborohydride is particularly useful for the reductive methylation of peptides due to its reasonably strong reducing potential and high stability under aqueous conditions.[21] The commercial liquid handler apparatus used was open to the lab atmosphere and slow outgassing of hydrogen cyanide, even under basic conditions, remained a key safety consideration. Alternative reducing agents were considered to overcome this issue. The initial reducing agents tested (triacetoxyborohyride,[22] borane-THF, and borane-pyridine complex[23]) are known to be water sensitive, but have been previously used for reductive aminations. Since no literature on the hydrolysis half-lives of these compounds was available, they were evaluated for their suitability in aqueous solutions. Even with several molar equivalents of reducing agent, it was found that the borane-THF and borane-pyridine complex resulted in non-quantitative conversion (~90%) and sodium

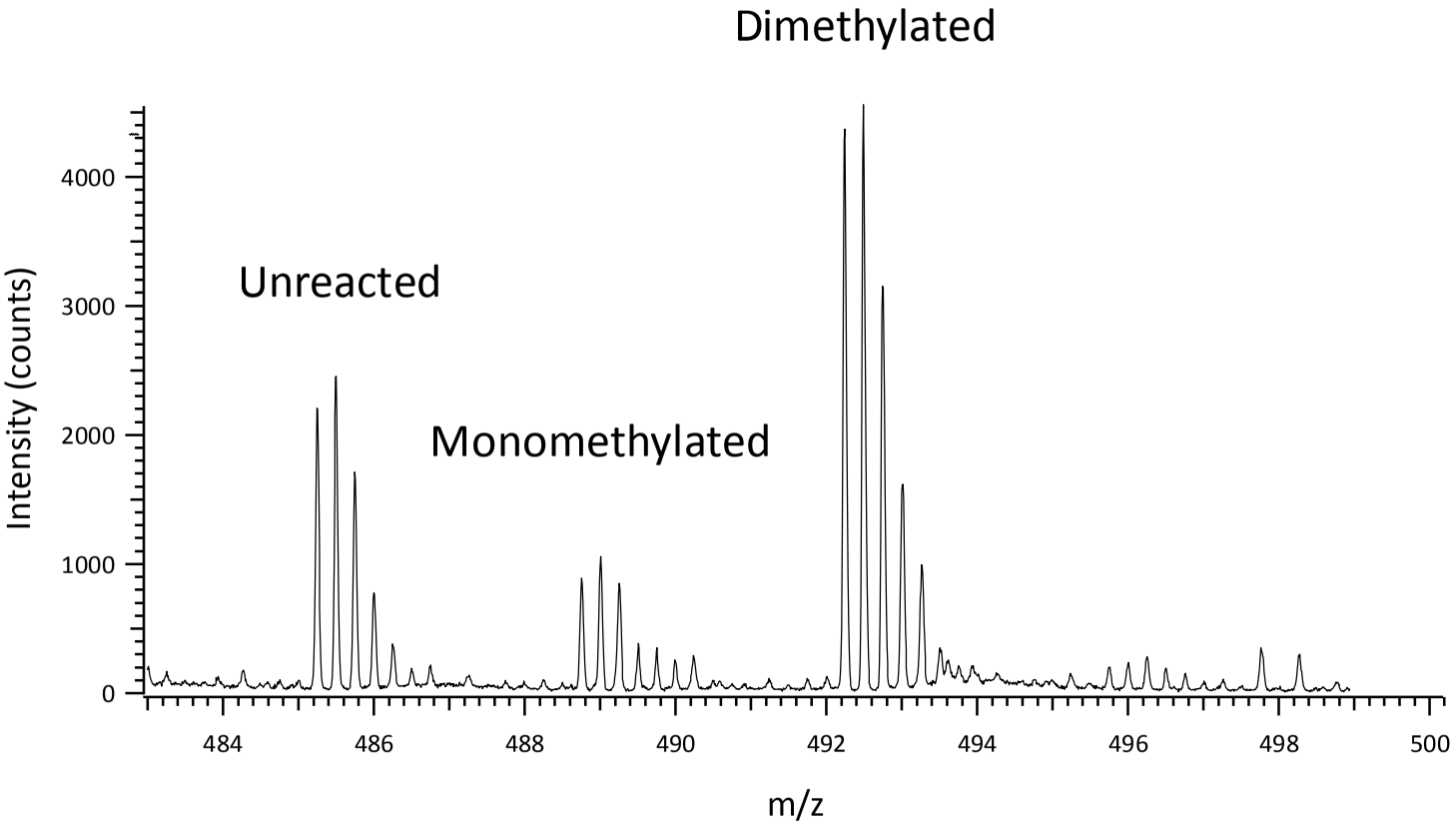


Figure 4.3 Incomplete Dimethylation. When insufficient amounts of the reagents are used for dimethylation, the following pattern of reactivity is observed, where the monomethylated peak is usually weaker than the unreacted peak or the dimethylated peak.

triacetoxyborohydride had nearly no conversion (<10%). A similar reagent, 2-picoline borane complex,[24] was found to be a useful alternative since it is relatively air stable, non-toxic, and can be prepared in methanol to useable concentrations (~2M). The main drawback encountered in using 2-picoline borane instead of sodium cyanoborohydride is its limited solubility in aqueous solutions. Upon standing in aqueous solution for an extended period of time, a precipitate formed which needs to be removed by centrifugation before the sample is subjected to additional downstream processing. Since the dimethylation reaction proceeded quantitatively under a wide range of pHs (4-8), pH optimization was not required. It was found that various buffers, such as ammonium bicarbonate (pH 8), triethylammonium bicarbonate (pH 8) and sodium acetate/acetic acid (pH 5) were all suitable for the dimethylation reaction.

With the general reaction conditions established, the labelling of complex digests was used to adjust and verify method performance. Overall method performance was determined by subjecting the sample to LC-MS analysis and identifying the various peptide products. The total count of correctly labelled peptides (guanidinylation at lysine with dimethylation at N-terminus) was compared against the total count of unreacted groups (i.e., unmodified N-termini/lysine or dimethylated lysine) or incorrectly labelled groups (i.e., guanidinylation at N-terminus, Figure 4.3). Since the dimethylation reaction was found to be relatively complete, a balance between peptides lacking the guanidinyl group at lysine (unmodified or dimethylated lysines) and peptides showing excessive guanidinylation (guanidinylation at the N-terminus) was desired. Peptides with ambiguous assignments (e.g., peptides that are lysine N-terminated) were discarded and not counted. With the optimized conditions, it was found that the most frequent undesired reaction product was guanidinylation at glycine or alanine N-terminated peptides. Other potential rare modifications, such as methylation (serine, threonine, histidine, aspartic acid, or glutamic acid) or guanidinylation (histidine) were not observed based on MS/MS sequencing and database searching of peptides.

4.3.3 Method Validation

Tryptic digests between 2 to 200 µg per sample vial at a concentration of 0.5 µg/µL were successfully labelled with >95% complete labelling, with the optimal reaction efficiency around 20 to 150 µg (>97%). It was found that the liquid handler was imprecise for the delivery of volumes less than 2 µL, which ultimately limited the lower limit that could be reached. Concentration ranges between 0.1-2 µg/µL were also tested and found to also give similar performance characteristics. For each sample, the amount of *O*-methylisourea and formaldehyde/2-picoline borane complex added was adjusted for the total peptide amount.

Since the stepwise reaction scheme necessitates that conversion of lysine groups is complete before addition of the reductive methylation reagents, a reasonably close estimate of the guanidinylation reagent amount is required. Initially, an LC-UV peptide quantification method was used in order to determine the optimal reagent amount for guanidinylation. Ideally, it would be preferential to go directly from digestion to the labelling step without an intermediate quantification step. Assuming that a protein concentration was determined prior to trypsin digestion, such as by Bradford or BCA assay, we investigated how deviations from the ideal reagent amount would affect labelling efficiency. For a 50 µg sample, reagent amounts corresponding to 20 to 500% of the ideal were tested. To maintain over 95% labelling efficiency, reagent ranges from 25% to 200% of ideal were required. Given the relative accuracy of these protein quantification methods, it should be feasible to go directly from estimation of the protein concentration by the aforementioned methods directly to the finished labelled peptide products for over 95% correct labelling.

4.3.4 Front End Sample Preparation Methods

In order to test the optimized labelling chemistry with a variety of front end sample preparation methods, different protein solubilization and cell lysis methods were considered. Since detergents and buffers are not always removed prior to sample workup after digestion, they were evaluated for their potential to interfere with the automated 2MEGA labelling protocol. Labelling strategies targeting the amine functionalities of proteins are among, if not, the most common for proteomics

applications,[25] and the findings here should be generally applicable to other similar methods such as the commercial iTRAQ and TMT reagents. Among the cleavable detergents used in this study, the structures of some detergents are available in the research literature (Rapigest™ [26, 27]) or the associated commercial literature (ProteaseMAX™, sodium 3-((1-(furan-2-yl)undecyloxy)carbonylamino)propane-1-sulfonate), whereas others detergents do not have their structures disclosed (AALS). Cleavable detergents often use acetals[26] and carbamates as the linker functionality between the hydrophobic and hydrophilic portions of the detergent. The cleavage protocol described for the reported acetal-containing detergents is typically treatment with acid (pH < 2), which is similar to the suggested treatment for cleavage of AALS. Since the cleavage products may yield functional groups that interfere with the reaction by potentially consuming reagents, such as the amine group produced from the hydrolysis of the carbamate linker in ProteaseMAX™, they should be evaluated for their potential effects. Similarly, commercial extraction buffers often contain proprietary mixtures of detergents for protein extraction which may complicate labelling or downstream processing in the presence of reagents or additional salts.

High concentration spikes of urea, SDS, Rapigest™, ProteaseMAX™, and anionic acid labile surfactant (AALS from Progenta) were added to tryptic digests at standard working conditions in order to simulate the potential impact on labelling efficiency. Two commercial protein extraction buffers (CellLytic™ M and TM Buffer from the Total Protein Extraction Kit) were also tested. The final concentrations of the sample preparation buffers and overall labelling efficiencies are described in Table 4.1. Samples with SDS were subjected to cleanup either using a combination of SCX/RPLC to sequentially remove SDS and salts or RPLC desalting alone. Rapigest, ProteaseMAX, and anionic acid labile surfactant were hydrolyzed as per the manufacturers' instructions. All samples were desalted using the desalting/quantification LC-UV method as previously described prior to MS analysis. Samples were acidified with trifluoroacetic acid, then desalted and quantified using an RPLC-UV prior to LC-MS analysis. Desalting was found to be necessary, due to the relatively high concentration of salts and buffers used in the labelling scheme. The typical hold with the equilibration

Table 4.1. Percentage of correct labelled peptides using different solvent conditions

Sample Solution	Replicate	Correct IDs	Total IDs	% Labelling	Average	St. Dev
200 mM NH_4HCO_3	1	676	731	92.5%	93.8%	1.1%
	2	702	746	94.1%		
	3	903	965	93.6%		
	4	960	1008	95.2%		
60% Methanol	1	819	860	95.2%	96.2%	0.8%
	2	941	973	96.7%		
	3	1045	1089	96.0%		
	4	1167	1202	97.1%		
4M Urea	1	897	962	93.2%	93.5%	0.3%
	2	983	1054	93.3%		
	3	1192	1272	93.7%		
	4	1347	1436	93.8%		
0.1% SDS	1	671	716	93.7%	94.2%	0.6%
	2	666	712	93.5%		
	3	915	965	94.8%		
	4	893	944	94.6%		
0.1% Rapigest SF™	1	814	857	95.0%	94.7%	0.9%
	2	905	967	93.6%		
	3	1090	1140	95.6%		
	4	1250	1322	94.6%		
0.1% ProteaseMax™	1	920	965	95.3%	95.1%	0.5%
	2	868	919	94.5%		
	3	1232	1290	95.5%		
	4	1103	1162	94.9%		
0.1% AALS	1	1090	1149	94.9%	94.9%	0.6%
	2	1007	1070	94.1%		
	3	1492	1561	95.6%		
	4	1363	1434	95.0%		

mobile phase was extended from five to eight minutes in order to remove the weakly retained salts, which appeared as a strong tailing peak in the chromatogram. (Figure 4.4).

Overall, it was found that all sample preparation methods, with the exception of SDS, did not give a significant change in the percentage of correctly labelled peptides. SDS was found to be problematic, since the high concentration of salt from the labelling chemistry severely reduced retention on the SCX column, leading to peptide recoveries around 10% and very low peptide identification numbers. Urea was found to induce carbamylation on lysines for of a small percentage of peptides (~1%). Since the carbamyl group (+43 Da) is a non-labile modification that prevents conversion to the homoarginine group (+42 Da), database searches also considering deamidation as a peptide modification (+1 Da) can lead to an isobaric modification pattern where a peptide with guanidinylation and deamidation is assigned as having carbamylation.

4.4 Conclusion and Future Work

With MS-based methods for peptide identification now commonplace, widespread adoption and integration of quantitative workflows appears to be next major step in addressing the analytical challenges coming to the forefront in many bioanalytical laboratories. The automation of a labelling chemistry suitable for isotopic labelling experiments using a commercially available liquid handler is reported. The method was found was to be compatible for simulated versions of various front end protein preparation methods. Future work will include application of the described labelling chemistry for the preparation and analysis of samples for quantitative workflows. More recently, there have been other reports of automated sample preparation methods using online sample preparation or with samples that have been loaded onto SPE cartridges.[20]

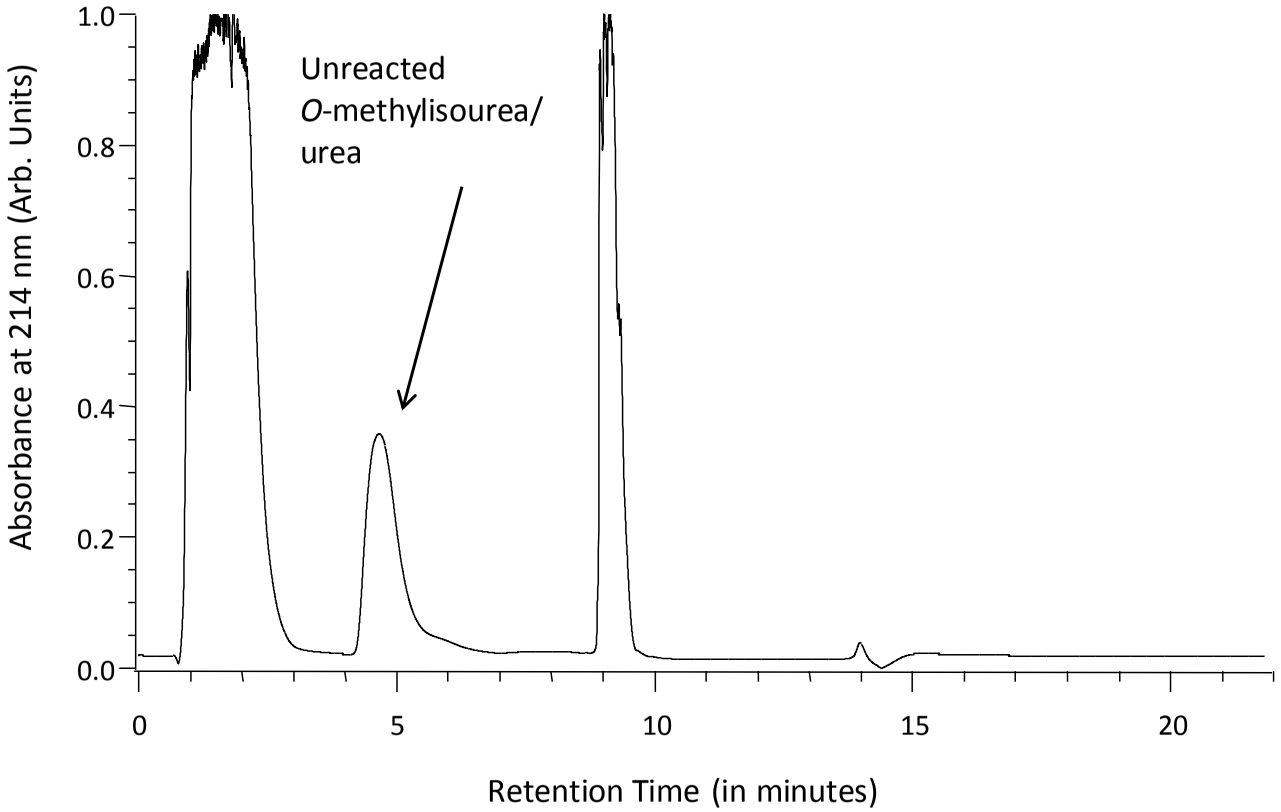


Figure 4.4 Desalting Chromatogram for 2MEGA labelled sample. The previously published desalting and peptide quantification method elutes peptides around $t_R = 5$ mins. However, the presence of *O*-methylisourea leads to a broad, weakly retained peak at $t_R = 4.8$ minutes, which requires a longer hold time.

4.5 Literature Cited

(1) Lemeer, S.; Jopling, C.; Gouw, J.; Mohammed, S.; Heck, A. J. R.; Slijper, M.; den Hertog, J. *Mol. Cell. Proteomics* **2008**, *7*, 2176-2187.

(2) Choudhary, C.; Kumar, C.; Gnad, F.; Nielsen, M. L.; Rehman, M.; Walther, T. C.; Olsen, J. V.; Mann, M. *Science* **2009**, *325*, 834-840.

(3) Shakey, Q.; Bates, B.; Wu, J. *Anal. Chem.* **2010**, *82*, 7722-7728.

(4) Jayapal, K. P.; Sui, S.; Philp, R. J.; Kok, Y.-J.; Yap, M. G. S.; Griffin, T. J.; Hu, W.-S. *J. Proteome Res.* **2010**, *9*, 2087-2097.

(5) Lange, V.; Malmstrom, J. A.; Didion, J.; King, N. L.; Johansson, B. P.; Schafer, J.; Rameseder, J.; Wong, C.-H.; Deutsch, E. W.; Brusniak, M.-Y.; Buhlmann, P.; Bjorck, L.; Domon, B.; Aebersold, R. *Mol. Cell. Proteomics* **2008**, *7*, 1489-1500.

(6) Wang, H.; Chang-Wong, T.; Tang, H.-Y.; Speicher, D. W. *J. Proteome Res.* **2010**, *9*, 1032-1040.

(7) Wang, N.; Li, L. *Anal. Chem.* **2008**, *80*, 4696-4710.

(8) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen Dan, B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell. Proteomics* **2002**, *1*, 376-386.

(9) Bendall, S. C.; Hughes, C.; Stewart, M. H.; Doble, B.; Bhatia, M.; Lajoie, G. A. *Mol. Cell. Proteomics* **2008**, *7*, 1587-1597.

(10) Bicho, C. C.; de Lima Alves, F.; Chen, Z. A.; Rappsilber, J.; Sawin, K. E. *Mol. Cell. Proteomics* **2010**, *9*, 1567-1577.

(11) de Godoy, L. M. F.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Froehlich, F.; Walther, T. C.; Mann, M. *Nature* **2008**, *455*, 1251-1254.

(12) Ji, C.; Guo, N.; Li, L. *J. Proteome Res.* **2005**, *4*, 2099-2108.

(13) Ji, C.; Lo, A.; Marcus, S.; Li, L. *J. Proteome Res.* **2006**, *5*, 2567-2576.

(14) Wang, N.; Xie, C.; Young, J. B.; Li, L. *Anal. Chem.* **2009**, *81*, 1049-1060.

(15) Brancia, F. L.; Openshaw, M. E.; Kumashiro, S. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 2255-2259.

(16) Brancia, F. L.; Oliver, S. G.; Gaskell, S. J. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 2070-2073.

(17) Beardsley, R. L.; Reilly, J. P. *Anal. Chem.* **2002**, *74*, 1884-1890.

(18) Banks, T. E.; Shafer, J. A. *Biochemistry* **1970**, *9*, 3343-3348.

(19) Hsu, J.-L.; Huang, S.-Y.; Chow, N.-H.; Chen, S.-H. *Anal. Chem.* **2003**, *75*, 6843-6852.

(20) Boersema, P. J.; Raijmakers, R.; Lemeer, S.; Mohammed, S.; Heck, A. J. R. *Nat. Protoc.* **2009**, *4*, 484-494.

(21) Borch, R. F.; Bernstein, M. D.; Durst, H. D. *J. Amer. Chem. Soc.* **1971**, *93*, 2897-2904.

(22) Abdel-Magid, A. F.; Carson, K. G.; Harris, B. D.; Maryanoff, C. A.; Shah, R. D. *J. Org. Chem.* **1996**, *61*, 3849-3862.

(23) Bomann, M. D.; Guch, I. C.; DiMare, M. *J. Org. Chem.* **1995**, *60*, 5995-5996.

(24) Sato, S.; Sakamoto, T.; Miyazawa, E.; Kikugawa, Y. *Tetrahedron* **2004**, *60*, 7899-7906.

(25) Julka, S.; Regnier, F. *J. Proteome Res.* **2004**, *3*, 350-363.

(26) Yu, Y.-Q.; Gilar, M.; Lee, P. J.; Bouvier, E. S. P.; Gebler, J. C. *Anal. Chem.* **2003**, *75*, 6023-6028.

(27) Yu, Y.-Q.; Gilar, M.; Gebler, J. C. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 711-715.

Chapter 5 – Reciprocal Labelling for Comparison of Samples from Aerobic and Anaerobic *E. coli*

5. 1 Introduction

Quantitative mass spectrometric methods used in the analysis of small molecules and peptides are well-established in commercial and research labs around the world. Targeted approaches utilizing selected reaction monitoring (SRM) for known or defined analytes exploit the inherent sensitivity gained from quadrupoles and ion traps in the first stage of MS analysis by selectively filtering for desired ions of interest.[1] More recently, quantitative mass spectrometry-based approaches have sought to utilize the exquisite capability of MS-based platforms to identify thousands of peptides from complex protein digests. Although scientists have succeeded in developing a range of techniques, quantitative mass spectrometric methods can be broadly classified as label-based (isotope incorporating) and label-free (typically spectral counting or ion current intensity). Among label-based methods, the actions of cells, enzymes, and reagents have been successfully exploited to introduce isotopically labelled groups *in vitro,*[2, 3] during enzymatic digestion,[4-6] or by chemical derivatization.[7, 8] Label free methods typically use spectral counting or extracted ion currents and have benefitted from newly development algorithms for feature alignment based on retention time.

Regardless of the strategy employed, appropriate design of a discovery-based quantitative proteomics experiment allows for a broad survey of peptide concentration changes to be determined in a non-targeted fashion, revealing information about proteins that is often new or unexpected, thereby generating avenues for further research. Unlike targeted approaches, discovery-based experiments have to contend with the concomitant identification and quantification of several peptides over a short span of time, resulting in comparatively poor S/N ratios and incomplete analysis of components in a sample. While the undersampling problem is well-documented for the identification of peptides, the additional information required for quantification increases demands on the already congested instrument duty cycle. Precursor, or MS-

based, methods can suffer from infrequent sampling across the peptide elution window, leading to poor peak reconstruction and diminished precision. Methods using MS/MS ion intensity can be adversely affected by low ion counting statistics for reporter ions in the MS/MS leading to quantization errors.[9] With these challenging conditions in mind, larger variations in accuracy and precision are tolerated and data analysis is performed with these considerations in mind.

Most methods used to validate MS-based quantification results, such as Western blot or immunoassay, help correct potential biases arising from using a mass spectrometry based method. It has been reported that MS-based quantification methods are characterized by a truncated dynamic range with ratios that tend toward unity.[10] It has been previously suggested that analyzing a replicate sample with inversion of the isotopic labelling assignments can be used to evaluate the consistency of quantification ratios; this strategy is known as "reciprocal labeling" or "reverse labelling".[11, 12] By inverting the isotopic labels used, the ratio of peptide peaks observed should be inverted; ratios that are invariant after the labelling may be due to chemical interferences or false positive peptide assignments. While the strategy is conceptually simple, reports implementing reverse labelling have been comparatively uncommon.

The labelling chemistry applied in this study is the previously reported 2MEGA (dimethylation after guanidinylation) strategy,[13] which is an automatable two step reaction involving conversion of lysine to homoarginine followed by reductive methylation. Light and heavy labels are incorporated using one of two variants of formaldehyde ($^{12}CH_2O$ or $^{13}CD_2O$) for the binary comparison of two peptide samples. The main advantage of the 2MEGA method is the comparatively low cost of the isotope source used for sample coding and the ability to automate the chemistry, which can allow for high-throughput processing of samples. Particular to the application of reverse labelling for sample quantification is a potential correction for the slight kinetic isotopic effect due to the presence of deuteriums in the labelling groups: ($(CH_3)_2$- vs. ($^{13}CD_2H)_2$). Although the kinetic isotopic effect for dimethyl groups has been reported to be relatively short when compared to the timescale of peptide peak widths,[13, 14] it is possible that these slight differences can affect the quantification

accuracy, since peptides are not precisely co-eluting from the HPLC and the observed ratio will change as the peak pairs elute.[15] Most software programs will average the ratio (usually weighted by intensity) over the entire elution profile of the paired peaks. The program used in this study, MASCOT Distiller, calculates the best line of fit for the scatterplot of light and heavy peptide intensities over the peptide elution profile. As the slight kinetic isotope effect will increase the scatter in the data points used for fitting, reverse labelling can correct for the systematic bias of isotopic labels containing deuteriums.

Here, we use the 2MEGA labelling protocol (dimethylation after guanidinylation) to label *E. coli* protein digests grown under aerobic or anaerobic conditions. By systematic evaluation, the consistency in the reported quantification results is examined. Variations at the instrumental level during acquisition and during the combined sample labelling/initial chromatography steps are considered. For the first set of experiments, peptide digests were separately labelled with either a light chain label or a heavy chain label and mixed together in a one to one ratio, with an expected ratio of unity for all peptides. After analyzing the same sample twice, differences in the reported quantification ratios are examined to determine which values fall outside of the expected range and have significant differences in measurement. In the second set of experiments, digests from *E. coli* grown aerobically or anaerobically are differentially labelled, as in a standard comparison experiment, and analyzed twice to reveal inconsistencies resulting primarily from instrumentation variation for a system in which ratios are not necessarily unity. Reverse labelling was then performed by switching the isotopic labelling assignments. Here, it was expected that contributions to inconsistency came from both instrumental analysis and differences in sample handling. Ratios which showed a large difference between the forward and reverse cases were then analyzed to determine whether replicate analysis or reverse labelling is preferable for checking the consistency in reported ratios.

5.2 Experimental

5.2.1 Chemical and Reagents

LC-MS grade solvents (water, methanol, and acetonitrile) and BCA Assay Kit were obtained from Fisher Scientific (Edmonton, AB). *O*-methylisourea hemisulfate, ^{12}C formaldehyde, 2-picoline borane complex, LC grade acetone, and LC-MS grade formic acid and trifluoroacetic acid were obtained from Sigma Aldrich (Oakville, ON). $^{13}CD_2O$ formaldehyde was obtained from Cambridge Isotope Laboratories (Andover, MA).

5.2.2 Cell Culture

E. coli K12 digests were prepared by culturing cells under aerobic or anaerobic conditions until OD_{600} = 0.5. Cells were lysed using an Emulsiflex homogenizer and the soluble fraction was frozen and kept at -80 °C prior to further processing. Proteins were precipitated by the addition of five volumes of acetone to the thawed samples and stored overnight at -80 °C, followed by centrifugation at 3 800 *g* for 90 minutes to pellet proteins. Proteins were re-solubilized in a minimal amount of 1% SDS and protein concentration was determined using the BCA assay.

5.2.3 Protein Digestion

Water and ammonium bicarbonate (1 M) were used to adjust protein samples to a final concentration of 100 mM NH_4HCO_3 (pH 8) and 0.1% SDS, followed by reduction with dithiothreitol, alkylation with iodoacetamide, and digestion by trypsin (50:1, protein:enzyme). After digestion overnight at 37 °C, samples were acidified with 10% formic acid to pH < 2 and SDS was removed from the peptide digest by strong cation exchange using a 2.0 mm i.d. x 150 mm PolySULFOETHYL A column (PolyLC; Columbia, MD). Mobile phase A was 10 mM KH_2PO_4 in 30% acetonitrile (pH 2.8) and mobile phase B was mobile phase A with 500 mM KCl (pH 2.8). A flow rate of 0.2 mL/min was used with the following gradient program (time in min, % B): 0, 0%; 20, 0%; 20.01, 100%; 30, 100%; 35, 0%; 45, 0%. A single fraction was collected between 27 and 37 minutes. Samples were processed in batches and pooled back into the aerobic and anaerobic samples. The samples were dried in a vacuum centrifuge and reconstituted using 200 mM NH_4HCO_3 to a peptide concentration of ~0.5 µg/µL. A small portion of the sample was used for peptide quantification using a LC-UV method as previously described.[16]

5.2.4 Isotopic Labelling of Peptide Digests

Based on the quantification results, 500 µg (10 x 50 µg samples) of the aerobic and anaerobic *E. coli* digests were labelled using the automated 2MEGA labelling protocol using a Gilson 215 liquid handler. Peptide solutions were adjusted to pH 11 using 8 µL 2M NaOH. *O*-methylisourea hemisulfate solution (~3M) was prepared in a 1:1 (v/v) of 2M NaOH and 1M Na_2CO_3 (pH 12) and 100 µL was added to each sample. The guanidinylation reaction was allowed to proceed for 90 minutes at 37 °C before adjustment to pH 7 using 30 µL 6M HCl and further adjustment to pH 5 using 24 µL 1M acetate buffer. Formaldehyde (4% in H_2O, (w/w)) was added to solution followed by subsequent addition of 2-picoline borane. For the aerobic *E. coli* digest the light label $^{12}CH_2O$ formaldehyde was used and for the anaerobic *E. coli* digest the heavy label $^{13}CD_2O$ was used. The dimethylation reaction was complete at 37 °C for 20 minutes before addition of 1M NH_4HCO_3 and reacted for another 20 minutes. Finally, samples were adjusted to pH <2 using 10% TFA. Samples were desalted and quantified by LC-UV as previously described. This process was repeated, but with the labelling assignments reversed: $^{12}CH_2O$ was used for the anaerobic digests and $^{13}CD_2O$ was used for the aerobic digests.

5.2.5 Sample Mixing and Strong Cation Exchange Analysis

One control sample was prepared by mixing together light and heavy labelled aerobic *E. coli* digest (denoted as $Aerobic_L$:$Aerobic_H$) in a 1:1 ratio by weight based on the peptide quantification results; similarly, the second control sample was prepared with anaerobic *E. coli* digest ($Anaerobic_L$:$Anaerobic_H$). For the first comparison sample, the light labelled aerobic *E. coli* digest was mixed with the heavy labelled anaerobic *E. coli* digest in 1:1 ratio ($Aerobic_L$/$Anaerobic_H$) and digests with the reversed labelling assignments were similarly prepared ($Anaerobic_L$/$Aerobic_H$). The mixtures were independently separated by strong cation exchange using a 2.0 mm i.d. x 150 mm PolySULFOETHYL A column. Mobile phase A was 10mM KH_2PO_4 in 30% ACN (pH 2.8) and solvent B was solvent A with 500 mM KCl (pH 2.8) and the following gradient program with a flow rate of 0.2 mL/min was used (time in min, % B): 0 , 0%; 7, 0%; 8, 3%; 36, 14%; 44, 20 %; 49, 30%; 53, 50%; 58, 50 %; 60, 0%; 70, 0%. Fractions were collected between 17 to 65 minutes in 1 minute fractions. Fractions were dried in a

vacuum centrifuge to remove acetonitrile before being acidified with 10% TFA and desalted and quantified by LC-UV as previously described. The desalted samples were dried down, reconstituted in 0.1% formic acid, and pooled together to generate 23 fractions with sufficient peptide sample in each fraction for at least two runs. Samples were aliquoted into two separate microcentrifuge tubes and stored at -80 °C pending MS analysis.

5.2.6 Mass Spectrometric Analysis

Samples were separated on a 300 µm i.d. x 150 mm Discovery C_{18} column using a Waters nanoAcquity LC followed by analysis on a Waters ESI-QTOF Premier mass spectrometer. Fractions were analyzed using the precursor ion exclusion (PIE) strategy. Peaks lists were initially processed to generate exclusion lists for the analysis of adjacent fractions using ProteinLynx. Peak lists were searched using MASCOT v2.2 (enzyme: trypsin; missed cleavages: 2, fixed modifications: Carbamidomethyl (C) Guanidinyl (K); variable modifications: Dimethylation_Light (N-term; $+C_2H_4$), Dimethylation_Heavy (N-term; $+^{13}C_2D_4$); MS tolerance: 30 ppm; MS/MS tolerance: 0.2 Da). A modified instrument-type setting using standard ESI-QTOF fragmentations also allowing a-ions was used. A database containing only *E. coli* K12 sequences (4337 entries) was used for searching.

5.2.7 Data Analysis

After fractions were analyzed by MS, peak lists were processed for identification and quantification using MASCOT Distiller v2.2. Raw data was processed into peak lists using standard ESI-QTOF conditions and searched using MASCOT v2.2 using the same conditions as for the exclusion lists. Quantification was performed for all peptides and a minimum correlation threshold of 0.9 was required for isotopic peak fitting. Peptides in the final list have at least one assigned spectrum with an ion score above the threshold. Reported quantification ratios were extracted and processed in Microsoft Excel to generate the final results. Occurrences of the same peptide across multiple fractions were consolidated by using the geometric average. Different charge states of the same peptide were averaged together using the geometric average.

Identified peptides and their quantification values were then arranged into a parsimonious list of proteins that accounted for all peptide identifications.

5.3 Results and Discussion

5.3.1 Control Dataset

The overall experimental workflow is shown in Figure 5.1. In order to examine the contribution to overall error strictly from instrumental analysis, the aerobic and anaerobic protein digests were labelled with both the light and heavy chain modification. The aerobic light and heavy labelled digests were mixed together in a 1:1 ratio and similarly for the anaerobic digests. The peptide mixtures were independently fractionated by strong cation exchange chromatography into 49 one minute fractions and desalted and quantified by LC-UV analysis. The desalted fractions were pooled to generate a set of 23 fractions for MS analysis; the fractions were pooled such that there was sufficient peptide quantity in each fraction for two analyses. Each fraction was divided into two aliquots and each set of the 23 fractions constituted one dataset. For each set of samples, the precursor ion exclusion (PIE) strategy was used to maximize the number of identifications.[17] In brief, an initial sample was analyzed by LC-MS and peptides identified by database searching. The *m/z* ratios of the identified peptides with their retention times were then excluded during analysis of neighbouring fractions. This process was continued until all of the samples were analyzed for a particular dataset and independently repeated for the second set of fractions for both the aerobic and anaerobic control samples.

Once the samples were analyzed, the data was processed for quantification using MASCOT Distiller. Each raw LC-MS data file was processed to generate a peak list and searched with MASCOT against an *E. coli* sequence database. To be considered for quantification, either the light or heavy labelled form of a peptide had to be identified with an ions score above the MASCOT identity threshold. Quantification was based off the peak fitting by MASCOT Distiller. Since MASCOT Distiller uses a fit to the entire isotopic envelope for quantification, a correlation score for the fit is reported and used as a quality filter to eliminate peaks for which peak fitting is poor or if peaks are

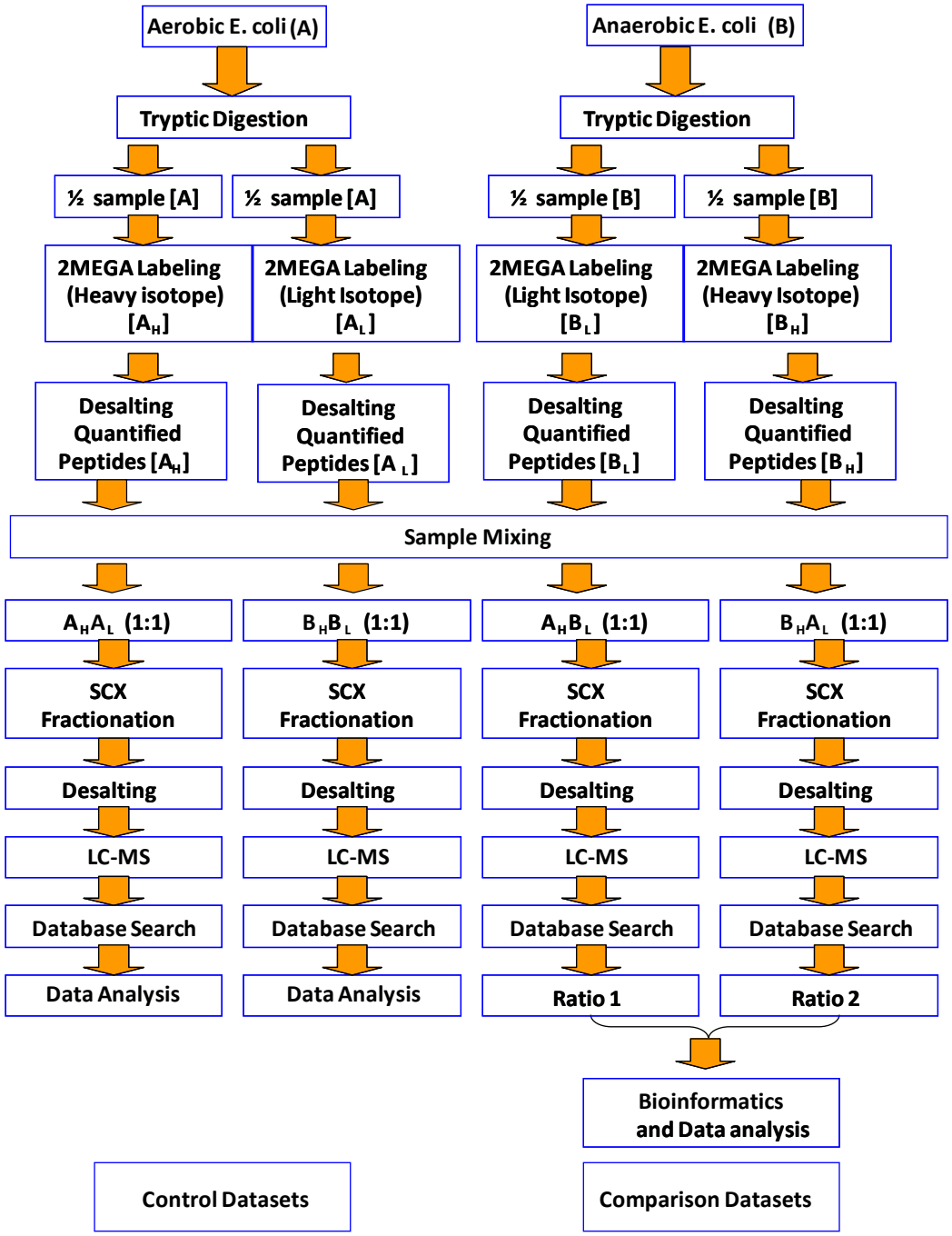


Figure 5.1 Experimental workflow for the control and comparison datasets

overlapped. It was empirically determined for the data that a correlation score of 0.90 gave suitable peak fitting. All peptides with a lower correlation score were discarded.

5.3.1.1 Identification Consistency

The overlap for the identified peptide sequences between the aerobic datasets was 71.7% and the anaerobic datasets was 71.1%. This result is similar to reports that re-analysis of a single fraction of a complex digest mixture is typically around 70%, indicating that although fractionation improves the number of identifications, the undersampling issue remains. For protein identifications, the overlap was slightly higher, with 81.6% for the aerobic dataset and 80.4% for the anaerobic dataset. Every peptide is present in the sample with a light or heavy N-terminal tag, although identification of only one form was needed for quantification of the pair. Approximately 45% and 49% of the peptides from the aerobic and anaerobic datasets, respectively, were identified with both heavy and light tags.

5.3.1.2 Quantification Method

For the data processing of quantified peptides, occurrences of the same peptide across multiple fractions were treated as equivalent and the geometric average of the ratios was taken. When the reported intensity was used for the weighting of peptides observed over multiple fractions, no significant difference in the overall consistency of the data was noted (data not shown). Since quantification ratios are based on the appearance of pairs in the MS spectrum, increased peak intensity may not necessarily improve reproducibility, beyond the minimum required for accurate peak fitting. Furthermore, although MASCOT Distiller reports an intensity value, the value represents the intensity of the precursor ion at the point the MS/MS spectrum was taken and is not characteristic of the entire elution profile. The geometric average was used versus the intensity weighted values due to the overall ease in calculating the ratios. Instances of the same peptide in different charge states were treated as being equivalent and averaged geometrically.

Across the four datasets, an average of ~85% of the identified peptides were reported with quantification values. For peptides without ratios, the most common

reason for non-reporting of ratios was due to correlation scores that did not meet the peak fitting criterion (R = 0.900). Peak fitting was compromised for peptides with low abundance or which overlapped with other peaks. Reported quantification ratios were discarded for a small percentage of peptides (<0.5% of total peptides) due to certain, specified quality issues. Some reported ratios were negative or zero, which could be due to a calculation error, poor peak fitting, missed detection of a peptide pair, or the rare misidentification of a peptide. Since identification of either the light or heavy labelled form of a peptide is sufficient for quantification, rare misassignments in which an incorrect N-terminal tag is assigned (e.g., a heavy labelled peptide was identified as a light labelled peptide) would result in the absence of the expected matching component 6 Da away. The absence of the expected paired peak gives a ratio that approaches zero or is arbitrarily high. While spuriously large values are often the result of these types of data processing errors, it is suggested that manual inspection of these values be undertaken to ensure that no peptides are lost when analyzing data with genuine peptide abundance changes. Proline N-terminal peptides were discarded for quantification, due to generally low ion scores and high likelihood of being incorrect assignments. Approximately three peptides in the dataset were quantified proline N-terminated peptides.

For each of the datasets, after data processing, a global average close to unity was observed. For the first and second aerobic datasets, averages of 1.02 and 1.04 were observed, respectively, and in the anaerobic datasets, averages of 1.00 and 1.00 were observed. Each of the four datasets suggest a lognormal distribution, but are not considered lognormal at 95% confidence due to the high kurtosis (thick tails) in all four distributions. Internal consistency between peptide quantification results were compared by calculating the relative difference between the duplicate experiments by taking the square root of the ratio between the duplicate experiments. Figure 5.2 shows the cumulative percentage of peptides less than a given relative fold difference in the quantification ratio. For the aerobic dataset, in which most of the ratios are close to the average of 1.03, approximately 95% of all datapoints are between 0.84 and 1.24 (1.21-fold difference) and 97.5% of all datapoints are within 0.78 and 1.34 (1.31-fold difference). For the anaerobic dataset, 95% of all datapoints lie within 0.72 and

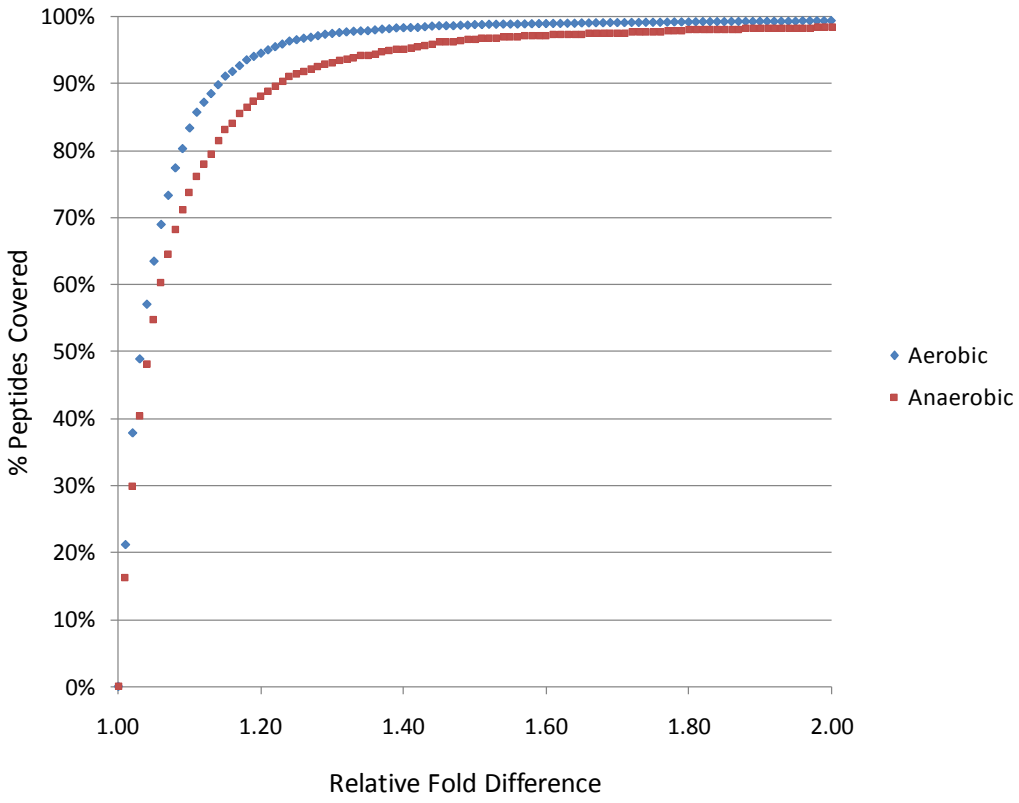


Figure 5.2 Relative difference plot for aerobic (blue diamonds) and anaerobic (red squares) control datasets

1.39 (1.39-fold difference) and 97.5% of all datapoints are within 0.60 and 1.67 (1.67-fold difference). A closer examination of peptides for both datasets that exist within the extreme 5% did not find any significant correlation with their charge state in strong cation exchange separation, length, or hydrophobicity (data not shown).

After examining the qualities of each independent dataset, the quantification results were then compared across replicates to determine the consistency between the two runs. Quantified, identical peptides from the first and second runs of the aerobic datasets are plotted on $\log_2$-$\log_2$ plot shown in Figure 5.3; a similar result was obtained for the anaerobic samples. Despite a global average close to 1.00, there are a number of data points that show greater than a two-fold difference. Even though reported values lie away from the expected value of zero, the ratios obtained appear to be relatively consistent, either being near the origin or within the first and third quadrant. With most of the data points clustered near the origin, the few outlier data points control the overall line of best fit. As such, it appears that most of the deviation observed is due to random variations in instrument performance or data processing. When considering the average protein ratios for peptides common to both the first and second analysis of the aerobic dataset, a nearly lognormal distribution is observed (Figure 5.4A). Although the average protein ratio is close to unity, the range for 95% of all protein values is between 0.61 and 1.63. While this range is not a particularly large range for the control experiments, the distinct outliers in the peptide data implies that some data filtering may greatly improve the fit. Here, the 1% of data with the largest relative difference between the two replicates was discarded, without regard to their average ratio. Since data comparing two different samples will not necessarily have a known value, a metric was selected that would not use the measured value, but rely strictly on the measurement consistency. Figure 5.4B shows the distribution of protein ratios after 40 peptides, corresponding to 9 proteins, were removed. It is noted that while nine of the forty removed peptides were single hit identifications, five peptides were two peptide identifications, and the remaining twenty-six peptides were from protein identifications with at least three peptides and have at least two consistent, quantifiable peptides remaining. This data indicates that poor peptide

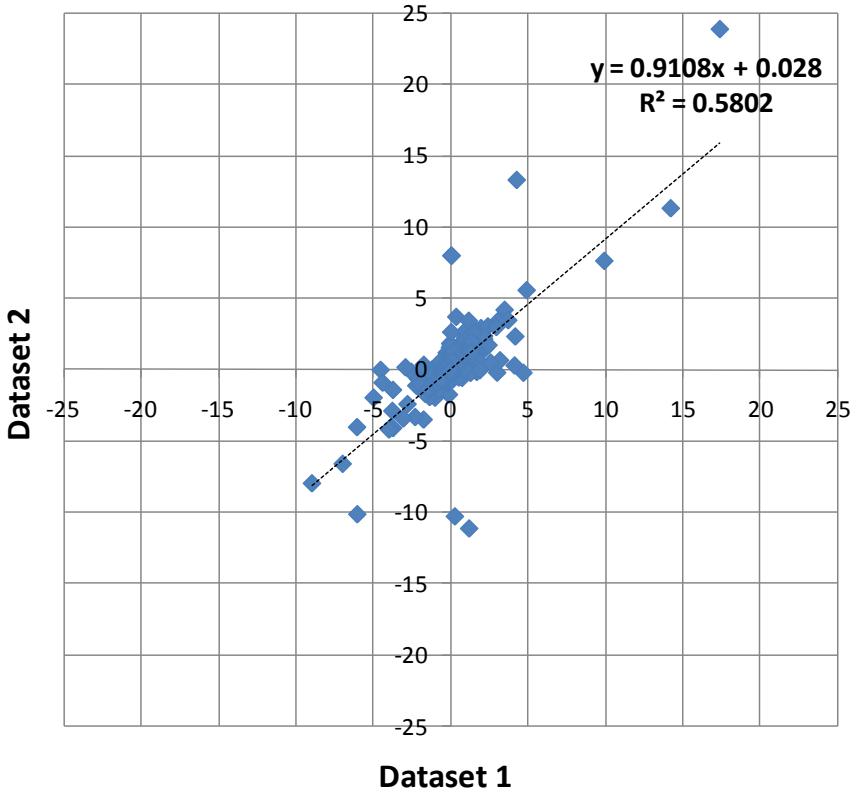


Figure 5.3 $\log_2$-$\log_2$ plot of peptide ratios from the two aerobic datasets.

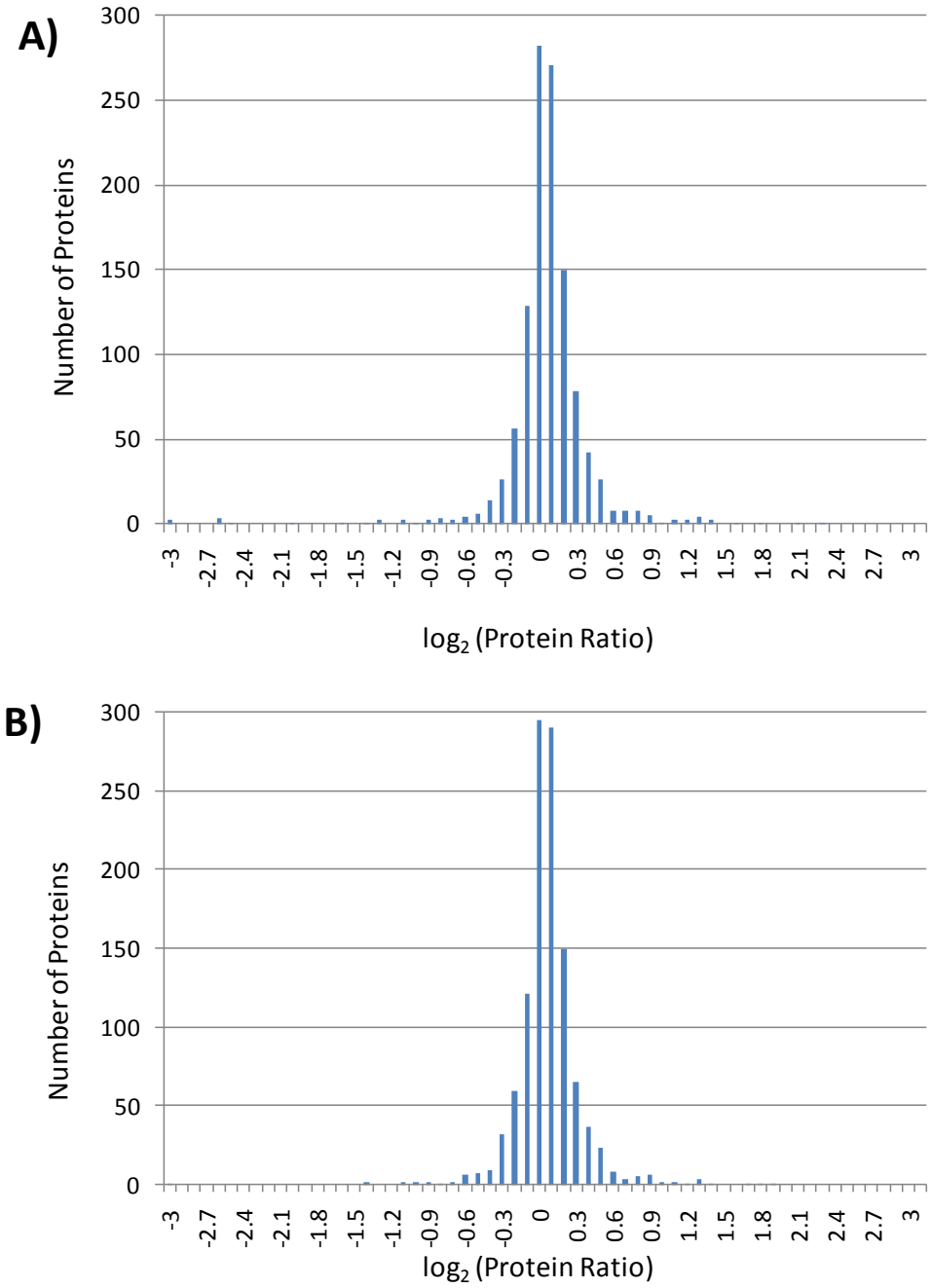


Figure 5.4 Average protein ratios using different percentages of data. Average protein ratios (log_2 scale) using all all peptides and b) average protein ratios with 99% of peptides.

quantification consistency is not limited to presumably lower concentration single hit matches and can occur for peptides from multi-hit proteins. After discarding only 1% of the data, the 95% range is between 0.70 and 1.43, which is a drastic reduction in the range of the confidence interval. For the ease of data analysis, proteins with ratios outside of the range of 0.67-1.50 will be considered as being significantly different than one based on this control data.

5.3.2 Comparison Dataset

In order to perform the comparison between the aerobically and anaerobically grown *E. coli* samples, four experiments were performed. The aerobic and anaerobic peptide digests were separately labelled with the light or heavy reagents. After desalting and peptide quantification, the aerobic light labelled digest was mixed together with the anaerobic heavy labelled digest (denoted as $Aerobic_L/Anaerobic_H$) in equal amounts by peptide weight and the anaerobic light labelled digest was mixed together with the aerobic heavy labelled digest ($Anaerobic_L/Aerobic_H$). After mixing, the samples were separated by strong cation exchange chromatography into 51 fractions and pooled together into 23 final fractions, ensuring that each fraction had enough peptide content for two analyses. The samples were aliquoted into two sets and stored prior to analysis. Fractions were run using the previously described precursor ion exclusion method. Peptide quantification ratios were processed in the same method as with the control datasets.

5.3.2.1 Consistency Between Replicate Analyses of the Same Sample

Since each sample was analyzed twice, the same metrics of relative difference and global average can be determined. For the $Aerobic_L/Anaerobic_H$ datasets, the global ratio averages within each dataset were 1.11 and 0.945, which is within the expected variation of the method; the global averages for the reverse labelling were similar at 1.15 and 1.17. The relative difference between each replicate dataset was also plotted and is shown on Figure 5.5. It is noted that normalization was not applied to the quantification results obtained. While equal weights of the peptides were mixed in a 1:1 ratio, an average ratio close to, but not necessarily equal to, unity was expected.

From Figure 5.5, the relative difference is expectedly larger than for the control datasets. The relative differences for 95% of the peptides for the $Aerobic_L/Anaerobic_H$ and $Anaerobic_L/Aerobic_H$ datasets were 1.42 and 1.36, respectively. The larger differences in measurement precision are expected, due to the variations in the peptide ratios away from unity. Taken together, it appears that the data is internally consistent between replicate measurements, even if the relative peptide concentrations are not necessarily unity. More interestingly, with the two analyses of the two different labelling assignments, there are three different comparison sets for any one dataset: one replicate analysis with the same labelling assignment and two datasets with the reversed labelling assignment. Although there are six possible comparisons possible, three of comparisons will be discussed. While the data is not shown, the analyses for the other possible dataset combinations show qualitatively similar results. To compare the data, the first run of the $Aerobic_L/Anaerobic_H$ was used as the basis for the three other sets. As with all other analyses, the lists of peptides from each dataset were compared against each other and matching, quantified peptides were determined and the protein quantification values for each dataset determined.

5.3.2.2 Identification and Quantification Consistency

The overlap in the peptide identifications from the first run of the $Aerobic_L/Anaerobic_H$ seems to be relatively consistent against the three other experiments, when considering the absolute number of overlapping peptides. The replicate analysis of the $Aerobic_L/Anaerobic_H$ sample overlaps with 4118 peptides (1141 proteins) from the $Aerobic_L/Anaerobic_H$ first run, whereas the first and second $Anaerobic_L/Aerobic_H$ analyses had 3969 peptides (1136 proteins) and 3799 peptides (1094 proteins) overlapping, respectively. Similarly, the commonality in the number of identified and quantified peptides is also relatively consistent, with a slight decrease in percentage overlap in the two reversed labelling conditions, around 85% of the identified peptides. These data suggest that the labelling and fractionation by strong cation exchange results in only a slight difference between the number of peptides identified and quantified. The consistency in the reported quantification ratios was

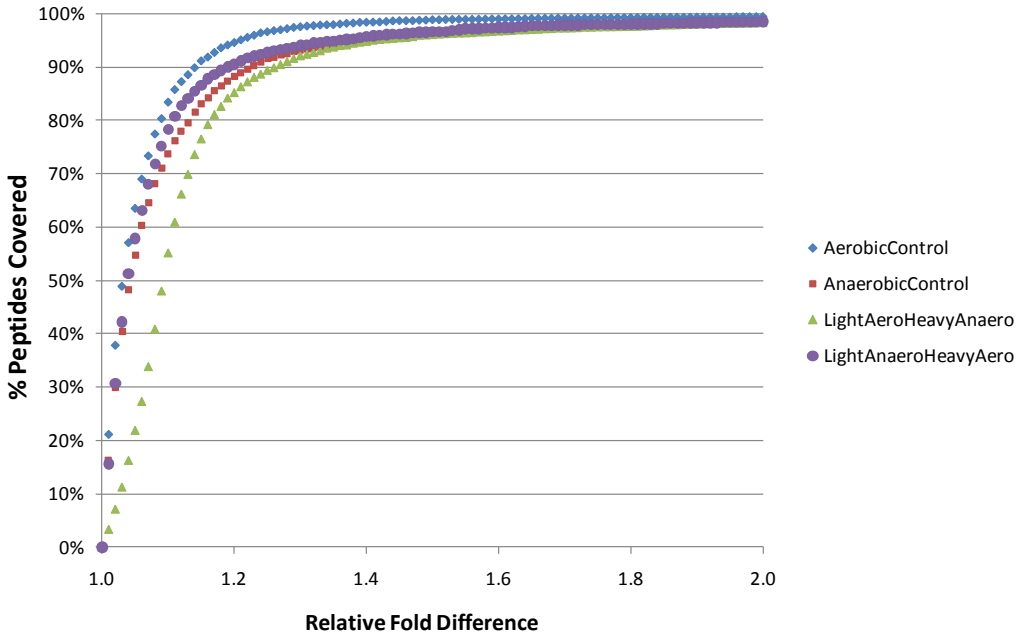


Figure 5.5 Relative difference plot for aerobic control ($Aerobic_{Heavy}Aerobic_{Light}$; blue diamonds), anaerobic control ($Anaerobic_{Heavy}Anaerobic_{Light}$; red squares), comparison set 1 ($Anaerobic_{Heavy}Aerobic_{Light}$; green triangles), and comparison set 2 ($Aerobic_{Heavy}Anaerobic_{Light}$; purple circles)

examined. The matching quantification ratios between the $Aerobic_L/Anaerobic_H$ and its replicate, along with the two reverse cases were calculated and plotted in Figure 5.6. Not surprisingly, the relative difference for the $Aerobic_L/Anaerobic_H$ dataset against its replicate analysis is smaller than against either of the reverse labelling datasets. Over the entire range of values, the replicate $Aerobic_L/Anaerobic_H$ run has a relative fold difference less than relative difference of 1.24, whereas the two $Anaerobic_L/Aerobic_H$ datasets each have an average relative difference of 1.29, meaning that each of the entries in the reverse labelling tend, on average, have an additional error of ~5%. For ratios deviating from unity, the relative standard deviation increases, particularly for greater than fivefold relative changes.[10] These data are consistent with the assertion that the labelling method described and the chromatography steps used are relatively reproducible and are not a significant source of observed variation when compared to other sources of variation.

5.3.3 Final Quantification Results Processing

With the four particular datasets comparing the same set of samples, there are numerous ways to perform the data analysis. For ease of comparison, the two $Aerobic_L/Anaerobic_H$ datasets will combined together and compared against the dataset with the reversed labelling assignments ($Anaerobic_L/Aerobic_H$). Unlike with the control system, the ratios of the light and heavy labelled components are not expected to be unity and can span the entire range of values. As expected, the relative percentage error for the comparison datasets is greater than for the control set. Notably, there exists a long tail even at higher relative differences (>200%), indicating that there is a small, but significant percent of peptides that do not show consistency between the forward and reverse labelling.

By using all of the data from 5048 peptides and generating a parsimonious protein list explaining all peptide identifications (fewest proteins, 1296 proteins total), the correlation of the $\log_2$-$\log_2$ plot between the $Aerobic_L/Anaerobic_H$ datasets and the $Anaerobic_L/Aerobic_H$ gives a line of best fit (Figure 5.7a, $y = 0.135x - 0.142$, $R^2 = 0.022$) where both the slope and intercept deviate from ideal behaviour (1 and 0, respectively). A visual inspection of the data shows that the majority of data points

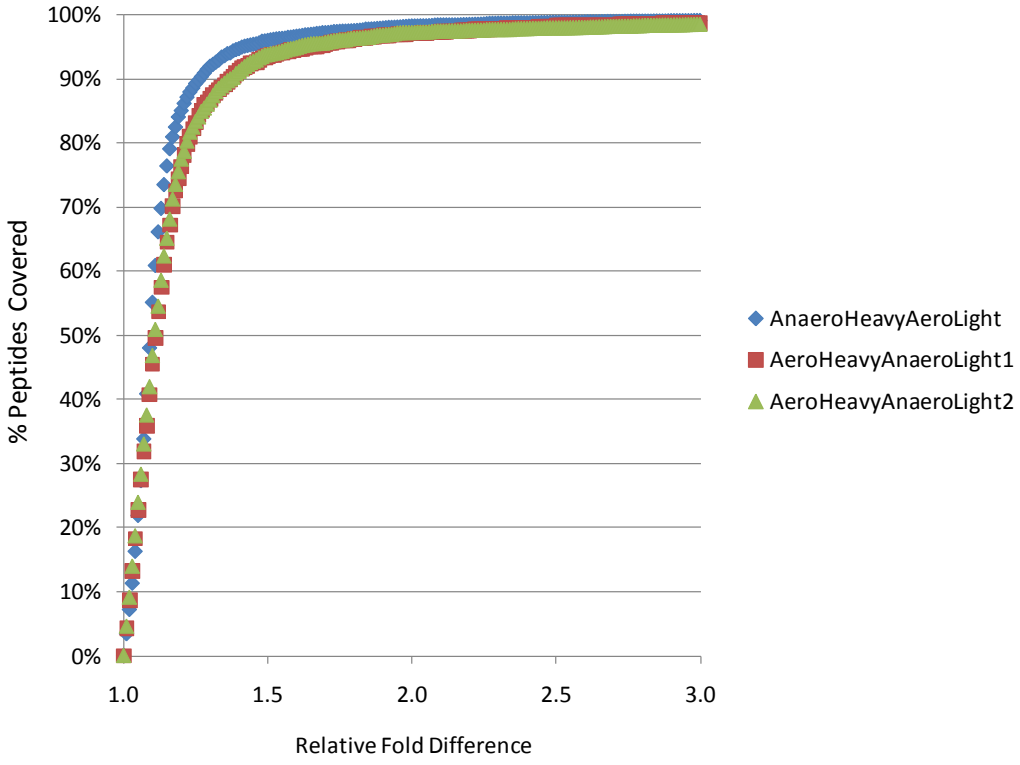


Figure 5.6 Relative difference plot for the four different comparison sets. The relative difference of the $Anaerobic_{Heavy}Aerobic_{Light}$ data to the second $Anaerobic_{Heavy}Aerobic_{Light}$ (blue diamonds), the first $Aerobic_{Heavy}Anaerobic_{Light}$ (red squares), and the second $Aerobic_{Heavy}Anaerobic_{Light}$ (green triangles) datasets.

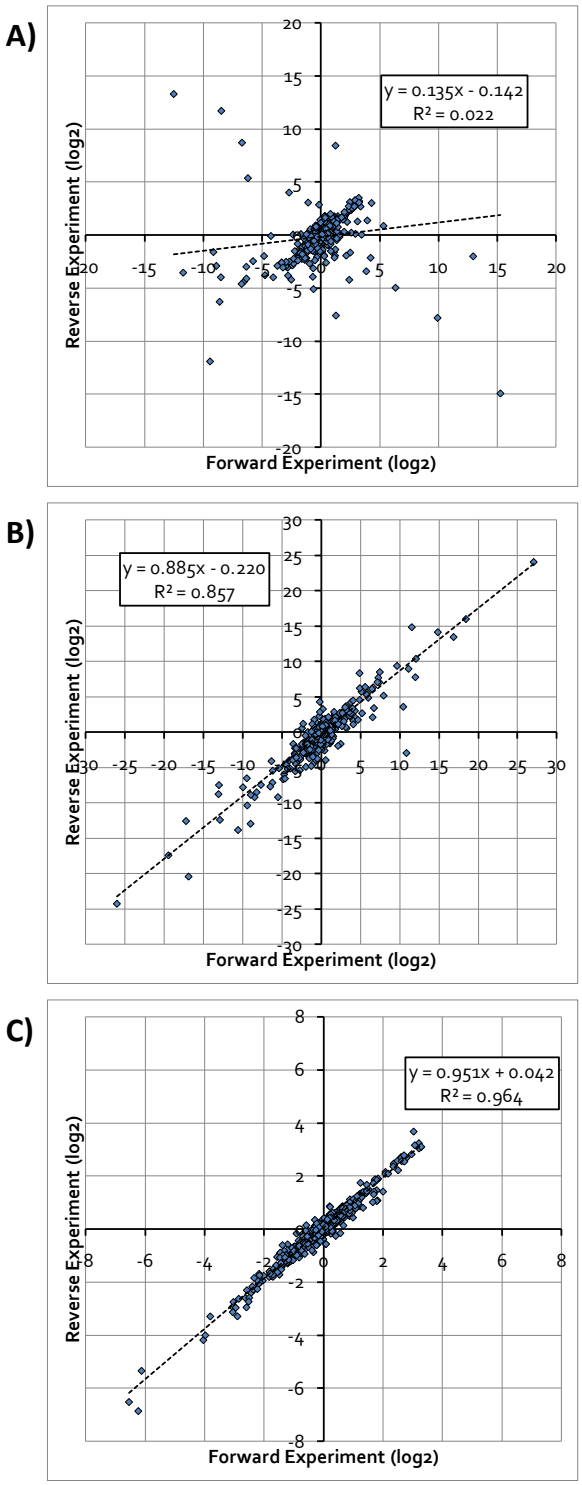


Figure 5.7 Protein ratios from $Aerobic_{Heavy}Anaerobic_{Light}$ versus $Aerobic_{Light}Anaerobic_{Heavy}$ using a) 100%, b) 99%, and c) 95% of peptide data

show the expected behaviour with some outliers that drive the poor fit. With any type of data processing, considerable care has to be taken not to over-process data. However, it seems that removal of a handful of outliers will vastly improve the consistency in the data. Since peptide identification and quantification ratios are used as surrogates for proteins identification and quantification, an additional peptide-level quality filter was applied to improve the results at protein level. Initially, the data was selectively reduced by eliminating 1% of all peptides (relative difference greater than 479%), which resulted in the removal of 50 peptides and 14 proteins. The line of best fit is $y = 0.885x – 0.220$ with an $R^2 = 0.857$ (Figure 5.7b). By discarding 5% of peptides, all of the peptides have less than 171% relative difference, and the number of proteins is reduced to 1250. The resultant line of fit is $y = 0.951x + 0.042$ with an $R^2 = 0.964$.

While this fit is very close to the ideal behaviour, the 252 discarded peptides were examined to determine if there were features common to these outliers to develop an independent metric that could be used to discriminate peptides with large relative differences that may reflect genuine biological changes. Of the 46 proteins removed from the final list, 44 were single peptide identifications and the quantification ratios could not be checked against the remaining peptide dataset for internal consistency. The two other proteins, HYCB_ECOLI and PTHA_ECOLI, were identified solely by three and two discarded peptides, respectively. The discarded peptides were all internally consistent in the forward and reverse cases, but the quantification values had a large scatter. For HYCB_ECOLI, the three peptides ranged from 2.76 to 178-fold relative up-regulation in anaerobic sample, with an average of up-regulation factor of 25.2. For PTHA_ECOLI, the range of values was 3.36 to 16.3 times up-regulation in the aerobic system and an average of 10.3.

The remaining 203 peptides were part of multiple hit protein identifications and the values from these discarded peptides were qualitatively compared against the more consistent data. Peptides and proteins were broadly classified as having up-regulated, unchanged, or down-regulated expression depending whether the relative ratio of the aerobic to anaerobic expression was found to be >1.5, between 1.5 and 0.67, or less than 0.67, respectively. Twenty-nine peptides were matched with proteins

that had up-regulated ratios; ten peptides were found to have both forward and reverse ratios classified as up-regulated, seven were found to have one up-regulated and one unchanged, two were found with one unchanged and one down-regulated, and ten peptides were found with one up-regulated and one down-regulated ratio. Similar qualitatively consistent results were observed with the 40 peptides matched to down-regulated proteins: 20 had both forward and reverse peptides with down-regulated values, 14 had one down-regulated ratio and one unchanged, two had one unchanged and one up-regulated, and 4 had one down-regulated and one up-regulated. These data suggest that while some qualitatively correct data (i.e., both up-regulated values) is being discarded (30 of 69 peptides; 43%), there is other data does not provide strong evidence for differential behaviour (i.e., at least one unchanged ratio, 25 of 69; 36%). More importantly, there is some data (14 of 69; 20%) that suggest a contradiction where one experiment suggests a significant up-regulation, whereas the other suggests significant down-regulation.

For the 134 discarded peptides matched to unchanged proteins, the data can be similarly divided into three cases. Forty-five peptides (34%) have at least one peptide where a forward or reverse ratio is reported as being unchanged. Approximately 18 or 13% of peptides show differential expression with up-regulation or down-regulation values with both labelling schemes. An analysis of these 18 peptides showed that 14 are peptides with missed cleavages (all at the N or C-terminus) and that 11 of these peptides are already represented in the dataset without the missed cleavage. The remaining 71 peptides (53%) constitute the cases where data is mutually contradictory, up-regulated in one labelling scheme and down-regulated in the other. These peptides highlight one main reason to utilize the forward and reverse labelling for quantitative proteomics experiments.

5.3.4 Justification of forward-reverse replicates versus forward-forward replicates

Misidentification of a peak, whether it be a genuine peptide or chemical interference, with a peptide assignment with the incorrect N-terminal label can be easily flagged when using the reverse labelling. As the reversal of the labelling assignments leads to a swap in the heights of the theoretical pair, “pairs” for which the

relative intensities do not change can be easily detected. For chemical interferences, in which the *m/z* ratio does not change between replicate runs, this will lead to an obvious contradiction in the observed ratio. On occasion, one peak from a genuine peptide pair will have an incorrect identification with the incorrect N-terminal label. Upon analysis of the reverse sample, re-analysis of that peak will result in a similar spectrum that will again identify the same peptide with the erroneous label. However, since the components have been switched, an obvious contradiction in the relative difference is observed, as shown for proteins is quadrants II and IV in Figure 5.7a.

Previous work utilizing reductive methylation at the N-termini peptides has shown the formation of strong, diagnostic a_1 ions for 99% of all non-glycine N-terminated peptides that provide information about the N-terminal residue.[18] While the MASCOT quantification methods have "minimal a_1 ion intensity" listed as a quality filter for peptide assignments, this feature has not been implemented for the MASCOT v2.2 search engine that was used in this study. Once this quality filter is supported, it is envisioned that this should allow for a simple check of the N-terminal amino acid assignment and confirm the light or heavy labelling tag attached to the N-terminus. These two additional pieces of information should improve the overall data quality from quantification experiments, by screening out potentially incorrect peptide assignments and preventing peak pair misidentification.

5.4 Bioinformatics Analysis

In terms of biological significance, the data of interest are the proteins that are differentially expressed in *E. coli* under aerobic and anaerobic growth conditions. Here, the final reported ratios used for comparison are the average results of replicates of the $Aerobic_L/Anaerobic_H$ datasets against $Anaerobic_L/Aerobic_H$ datasets. The final results are shown in Table 5.1. Bioinformatics processing was performed by examining proteins that were found to be consistently changed in both datasets. Tentative protein functions were assigned based on previously reported functions for identified genes.[19] It is noted that various databases have attempted to consolidate the wide range of *E. coli* genomic and transcriptomic information from several data repositories

and studies. Online tools, such as EchoBASE[20] and EcoCyc,[21] aggregate data from a range of sources and are continually updated.[22, 23]

5.4.1 Response to Oxygen

Aerobic respiration control protein ARCA_ECOLI (gene name: *arcA*), which was found to be significantly more abundant in the anaerobic samples (1.59 times), works with *arcB* to form a two component regulatory system. As a kinase, *arcB* can phosphorylate *arcA* under anaerobic or microaerobic conditions in order to regulate the expression of over 100 genes.[24] The change in the relative abundance of *arcA* between aerobic and anaerobic conditions is consistent with its function in modulating response to oxidative stress[25] and its downstream effects on certain metabolic pathways. Although *arcB* was not found to be significantly changed in this study, *arcB* changes phosphorylation state in response to changing quinone pool[26] and may not have significantly altered expression, as measured in this study. OXYR_ECOLI (*oxyR*) is responsible for response to oxidative stress as was only found to be at a slightly higher concentration in the aerobic samples (1.34-fold). Similarly, catalyase-peroxidase (KATG_ECOLI, *katG*), whose function is the controlled catalysis of hydrogen peroxide into oxygen and water, was only slightly more abundant in aerobic samples (1.40-fold). The absence of oxygen in the environment leads to reduced expression of proteins involved in dealing with oxidative stress, such as superoxide dismutase (SODM_ECOLI, *sodA*) which was found eightfold higher in the aerobic samples. Similarly, TRXB_ECOLI (*trxB*), a thioredoxin reductase, was found to be 1.52 times more abundant in the aerobic samples. However, SODF_ECOLI (*sodB*), another superoxide dismutase, was found to be relatively unchanged between aerobic and aerobic samples (0.78-fold in aerobic samples).

Two components of the cytochrome *o* oxidase complex (*cyoA* and *cyoB*) were found to be expressed approximately six-fold greater under aerobic conditions, which is qualitatively consistent with expression data from *cyoA-lacZ* and *cyoB-lacZ* fusions (140-fold relative expression in aerobic to anaerobic conditions).[27] However, two proteins of the cytochrome d terminal oxidase system (*cydA* and *cydB*), were found to have expression levels were found to be almost unchanged (0.90 and 1.03-fold change

Table 5.1 Proteins ratios from aerobically and anaerobically grown *E. coli*

Protein ID (_ECOLI)	# of Peptides	Forward (Aero/ Anaero)	Reverse (Aero/ Anaero)	Avg	Protein ID (_ECOLI)	# of Peptides	Forward (Aero/ Anaero)	Reverse (Aero/ Anaero)	Avg
ACON1	12	0.52	0.54	0.53	DING	1	1.03	0.43	0.67
ACON2	23	1.68	1.56	1.62	DINJ	1	0.45	0.61	0.52
ACRD	1	1.67	1.75	1.71	DKGA	1	0.58	0.64	0.61
ADHE	33	0.34	0.29	0.32	DLDH	13	2.19	2.08	2.13
AER	1	0.53	0.46	0.49	DMSA	9	0.29	0.29	0.29
AGAL	1	1.54	1.62	1.58	DMSB	5	0.20	0.18	0.19
AHPF	11	0.75	0.55	0.64	DMSD	1	0.25	0.36	0.30
ALDA	2	3.78	5.67	4.63	DNAC	1	0.73	0.55	0.63
AMIA	1	1.74	1.84	1.79	DNAJ	8	1.68	1.52	1.60
APAH	1	0.56	0.76	0.65	DNAK	32	1.78	1.70	1.74
ARCA	5	0.66	0.60	0.63	DPIA	1	0.37	0.38	0.37
ARLY	1	1.49	1.64	1.57	DPPA	1	2.40	2.69	2.54
ARTI	1	1.72	1.51	1.62	DPS	10	0.54	0.41	0.47
ASPA	10	0.65	0.69	0.67	DSBC	2	1.54	1.48	1.51
ATDA	2	2.64	2.71	2.68	DSDC	1	2.89	2.75	2.82
ATMA	5	0.67	0.59	0.63	DYR	1	0.76	0.58	0.67
ATPE	3	0.63	0.64	0.64	ECOT	1	1.77	1.30	1.52
AVTA	2	1.42	1.90	1.64	EFG	28	1.68	1.49	1.58
BARA	1	0.58	0.57	0.58	EFPL	4	0.67	0.66	0.66
BIOD2	1	0.47	0.41	0.44	EXBB	1	5.93	5.15	5.53
CADC	2	0.22	0.20	0.21	EXBD	1	2.84	1.15	1.81
CAN	2	3.90	3.55	3.72	FADE	1	0.61	0.66	0.64
CBPA	1	1.75	1.80	1.78	FADL	2	0.65	0.63	0.64
CBRA	1	2.34	2.08	2.20	FDHF	2	0.01	0.01	0.01
CH10	2	1.90	2.26	2.07	FDNG	5	0.28	0.24	0.26
CH60	14	2.20	1.81	1.99	FDNH	1	0.64	0.69	0.67
CHEZ	2	0.43	0.47	0.45	FDOG	6	2.32	2.43	2.38
CILA	1	0.04	0.01	0.02	FDOH	2	2.70	2.42	2.55
CIRA	1	9.56	9.21	9.38	FEOB	5	1.98	1.73	1.85
CISY	8	1.90	1.84	1.87	FEPB	1	7.45	9.66	8.48
CLPA	11	2.55	2.44	2.49	FHUA	1	19.84	8.14	12.71
CN16	2	1.59	1.82	1.70	FKBB	1	0.58	0.54	0.56
COABC	1	2.07	2.24	2.15	FLGE	3	0.40	0.31	0.35
CORC	3	0.64	0.50	0.57	FLGH	1	0.75	0.58	0.66
CSPA	1	0.38	0.28	0.33	FLGK	1	0.36	0.37	0.37
CSPD	1	2.19	3.45	2.75	FLIA	1	0.74	0.51	0.61
CSPE	1	0.51	0.86	0.66	FLIC	9	0.48	0.45	0.47
CUSA	1	0.64	0.69	0.67	FLIG	1	0.64	0.61	0.63
CUSR	1	0.56	0.48	0.52	FLIM	2	0.66	0.63	0.65
CYCA	1	2.52	2.58	2.55	FNR	1	1.52	1.49	1.50
CYOA	1	7.03	6.09	6.54	FOCA	2	0.31	0.35	0.33
CYOB	4	5.70	6.64	6.15	FOLD	2	0.46	0.48	0.47
CYSJ	2	0.58	0.46	0.52	FRDA	5	0.29	0.28	0.29
CYSN	2	2.18	3.22	2.65	FRDB	7	0.26	0.23	0.25
DADA	2	1.67	2.08	1.86	FTNA	3	0.35	0.25	0.29
DAPB	1	1.51	1.75	1.62	FUCO	1	1.58	1.47	1.52
DAPD	6	1.81	1.93	1.87	FUCR	1	1.68	1.53	1.60
DCEA	5	0.56	0.45	0.50	FUMB	10	0.19	0.18	0.19
DCOR	3	0.62	0.64	0.63	GADC	3	0.51	0.46	0.49
DEAD	20	0.69	0.62	0.66	GAL7	1	1.98	1.74	1.86
DGAL	1	4.64	4.22	4.43	GARR	1	2.10	1.56	1.81
DHE4	1	0.52	0.74	0.62	GATD	2	1.20	1.94	1.52
DHNA	14	3.82	3.29	3.55	GATY	11	0.71	0.62	0.67
DHSA	5	2.28	2.12	2.20	GATZ	12	0.65	0.63	0.64
DING	1	1.03	0.43	0.67	GCH1	3	2.53	1.85	2.16

Table 5.1 continued

Protein ID (_ECOLI)	# of Peptides	Forward (Aero/ Anaero)	Reverse (Aero/ Anaero)	Avg	Protein ID (_ECOLI)	# of Peptides	Forward (Aero/ Anaero)	Reverse (Aero/ Anaero)	Avg
GCSP	4	1.90	1.81	1.85	MCP1	4	0.55	0.43	0.49
GLF	9	0.66	0.57	0.61	MDAB	1	1.56	1.60	1.58
GLGX	1	0.77	0.36	0.53	MENB	4	0.59	0.57	0.58
GLPA	18	0.61	0.56	0.58	MEND	1	0.77	0.51	0.63
GLPB	3	0.50	0.49	0.49	MGLA	2	6.67	6.11	6.39
GLPC	9	0.53	0.52	0.52	MNTR	2	1.66	1.98	1.82
GLPD	16	1.82	1.76	1.79	MOTA	2	0.54	0.47	0.51
GLPF	1	2.71	1.16	1.78	MPPA	2	0.79	0.51	0.64
GLPK	20	1.71	1.64	1.67	MPRA	3	1.48	1.63	1.55
GLPX	3	0.38	0.43	0.41	MREC	2	0.75	0.58	0.66
GLRX1	1	0.40	0.38	0.39	MRP	1	0.70	0.60	0.65
GLSA1	2	0.39	0.44	0.41	MRR	1	1.74	1.56	1.65
GLTD	1	2.07	1.53	1.78	MTLD	7	0.64	0.52	0.58
GLTS	1	0.57	0.60	0.58	NAGB	1	1.38	2.35	1.80
GPMA	6	4.04	3.32	3.66	NANA	7	7.25	9.23	8.18
GPMI	6	0.63	0.55	0.59	NANE	1	6.25	5.56	5.90
GTRB	1	0.50	0.41	0.45	NANK	1	5.97	6.91	6.42
HCHA	2	0.58	0.62	0.60	NARG	14	2.03	1.86	1.94
HCP	3	0.66	0.55	0.60	NARH	4	2.67	2.31	2.48
HDEA	1	0.63	0.42	0.51	NARJ	3	5.82	5.11	5.45
HDEB	2	0.57	0.38	0.46	NARL	4	1.89	1.91	1.90
HTPG	21	2.21	2.18	2.20	NARP	1	0.62	0.61	0.61
HTPX	1	1.83	1.54	1.68	NARX	1	1.20	2.51	1.74
HYBA	6	0.14	0.12	0.13	NARY	1	1.26	3.51	2.10
HYBD	1	0.19	0.14	0.16	NARZ	2	1.97	1.98	1.98
HYBE	1	0.14	0.07	0.10	NDK	4	2.48	2.28	2.37
HYCE	6	0.05	0.06	0.05	NFSA	3	3.68	2.74	3.17
HYCG	1	0.01	0.01	0.01	NIFU	2	1.50	1.62	1.56
HYPB	2	0.36	0.30	0.33	NIKA	2	0.37	0.35	0.36
HYPD	1	0.22	0.19	0.20	NIKC	1	0.26	0.22	0.24
HYPE	1	0.27	0.31	0.29	NIKR	2	0.59	0.52	0.55
IBPA	3	1.47	1.69	1.58	NIRD	1	1.84	2.12	1.97
IBPB	2	1.84	2.05	1.94	NLPC	1	1.67	1.59	1.63
IDH	13	1.92	2.02	1.97	NLPI	1	0.67	0.54	0.60
IMDH	12	1.51	1.51	1.51	NRDD	2	0.33	0.21	0.26
INSL	1	0.51	0.56	0.53	NRDR	2	2.05	1.70	1.87
IPYR	4	1.94	1.59	1.75	NUOE	2	1.84	2.19	2.01
ISPF	1	1.81	1.41	1.60	NUPC	1	0.55	0.53	0.54
KHSE	1	1.61	1.45	1.53	ODO1	12	1.61	1.40	1.50
KPYK2	10	0.61	0.58	0.59	ODO2	6	2.03	2.10	2.06
KUP	1	0.79	0.50	0.63	ODP1	30	2.76	2.86	2.81
LEP	4	0.54	0.75	0.64	ODP2	17	5.15	5.07	5.11
LEXA	2	0.72	0.52	0.62	OMPT	4	1.76	1.96	1.86
LLDD	1	4.71	2.36	3.33	OMPW	1	0.59	0.55	0.57
LOLE	1	0.64	0.68	0.66	OPDA	7	1.67	1.58	1.62
LONH	2	0.62	0.68	0.65	OSMC	1	0.65	0.48	0.56
LPXB	1	1.46	1.58	1.52	OSME	4	0.61	0.50	0.55
LRHA	1	0.43	0.22	0.31	OTSA	1	0.65	0.65	0.65
LRP	2	1.73	1.53	1.63	PARE	2	0.40	0.46	0.43
MALK	3	1.73	1.59	1.66	PCKA	9	2.36	1.99	2.17
MASY	8	0.57	0.62	0.59	PEPE	2	0.40	0.57	0.47
MBHL	3	0.51	0.41	0.46	PEPT	4	0.58	0.47	0.52
MBHM	7	0.17	0.16	0.16	PFLA	4	0.73	0.60	0.66
MBHS	1	0.54	0.48	0.51	PFLB	21	0.55	0.46	0.50
MBHT	3	0.13	0.13	0.13	PFLF	1	0.18	0.12	0.15

Table 5.1 continued

Protein ID (_ECOLI)	# of Peptides	Forward (Aero/ Anaero)	Reverse (Aero/ Anaero)	Avg	Protein ID (_ECOLI)	# of Peptides	Forward (Aero/ Anaero)	Reverse (Aero/ Anaero)	Avg
PHEP	1	0.64	0.64	0.64	SSEB	1	1.69	2.25	1.95
PHNA	1	0.42	0.33	0.37	STPA	3	2.62	2.74	2.68
PITA	1	0.74	0.58	0.65	SUCC	8	1.78	1.56	1.67
PLSX	1	0.51	0.77	0.63	SUCD	5	2.26	2.35	2.30
POXB	3	0.54	0.41	0.47	TALA	5	0.50	0.58	0.54
PPA	1	0.49	0.59	0.54	TDCE	10	0.55	0.49	0.52
PPIC	2	0.77	0.43	0.58	TEHB	3	0.44	0.40	0.42
PPSA	1	0.39	0.20	0.28	TESB	2	2.15	2.39	2.26
PROA	3	1.77	1.81	1.79	THD2	3	1.88	1.57	1.72
PROV	2	1.97	1.55	1.75	TONB	1	7.19	6.50	6.84
PTHB	3	6.37	7.78	7.04	TPIS	5	0.32	0.37	0.35
PTKA	4	0.71	0.59	0.65	TPPB	1	0.62	0.41	0.51
PTTBC	4	2.74	2.50	2.62	TPX	3	2.09	2.35	2.21
PTW3C	4	2.41	2.41	2.41	TRPB	6	0.65	0.61	0.63
PURA	11	0.44	0.46	0.45	TRPH	1	0.80	0.37	0.54
PURU	4	1.64	1.39	1.51	TRUB	1	1.82	1.42	1.61
PUTA	9	5.58	6.17	5.87	TRUD	1	1.41	2.16	1.75
PUTP	1	4.21	4.22	4.22	TRXB	4	1.68	1.38	1.52
PYRD	3	0.57	0.68	0.62	TYPH	8	1.69	1.69	1.69
PYRE	1	2.05	1.76	1.90	TYRA	2	0.53	0.65	0.58
RAIA	1	3.73	3.45	3.59	UBID	4	1.88	1.56	1.71
RCMNS	1	1.77	2.03	1.90	UDP	1	1.76	1.34	1.53
RFAL	1	0.68	0.34	0.48	UIDR	1	0.42	0.65	0.52
RFFA	3	0.51	0.47	0.49	URK	5	0.69	0.65	0.67
RHLE	7	0.64	0.59	0.62	USPA	1	1.42	1.84	1.62
RIBB	2	2.00	1.62	1.80	USPG	4	0.64	0.56	0.60
RIHC	1	1.51	1.49	1.50	UVRB	3	1.79	1.83	1.81
RIR1	11	2.50	2.25	2.37	WBBJ	1	0.59	0.58	0.59
RIR2	2	2.05	2.16	2.10	WBBL	2	0.32	0.33	0.33
RL11	5	0.61	0.59	0.60	XERD	1	1.50	1.61	1.55
RL14	3	1.60	2.05	1.81	YACC	1	0.65	0.61	0.63
RL27	3	0.41	0.43	0.42	YACF	3	1.91	1.23	1.53
RL29	3	0.69	0.65	0.67	YACL	2	1.57	1.63	1.60
RL30	2	0.16	0.18	0.17	YAGE	2	1.38	2.11	1.71
RL31	2	0.72	0.50	0.60	YAHK	1	0.53	0.55	0.54
RL32	1	0.08	0.13	0.10	YBAQ	1	1.63	2.07	1.84
RL33	1	0.06	0.06	0.06	YBBN	5	2.18	2.43	2.30
RODZ	5	0.66	0.66	0.66	YBGI	2	1.64	1.45	1.54
ROF	1	1.89	3.22	2.47	YCBB	1	1.11	0.38	0.65
RPE	1	0.64	0.62	0.63	YCCU	2	0.55	0.71	0.62
RRMA	1	0.52	0.47	0.49	YCEF	1	0.65	0.39	0.50
RS18	3	0.71	0.61	0.66	YCIH	1	0.44	0.37	0.41
RS19	3	0.63	0.62	0.62	YCIO	2	1.71	1.77	1.74
RS20	2	0.24	0.17	0.20	YDCP	1	0.61	0.36	0.47
RS21	2	0.13	0.17	0.15	YDFH	1	0.37	0.79	0.54
SDAC	4	0.63	0.57	0.60	YDFZ	1	0.46	0.43	0.45
SDHD	6	1.86	1.40	1.62	YDGH	3	0.59	0.55	0.57
SLP	4	0.38	0.28	0.33	YDJI	1	0.20	0.21	0.21
SLYB	3	0.65	0.51	0.58	YEBE	2	0.46	0.41	0.43
SLYD	1	0.40	0.31	0.35	YEED	1	0.40	0.49	0.44
SMPA	1	0.80	0.40	0.56	YEEE	1	0.46	0.44	0.45
SODM	2	7.54	9.71	8.56	YEEF	2	0.52	0.55	0.54
SPEE	2	1.96	1.91	1.94	YEEX	2	0.66	0.56	0.61
SPY	1	0.54	0.43	0.48	YEFI	3	0.57	0.52	0.54
SRLD	1	5.14	6.56	5.81	YEGP	1	0.45	0.33	0.39

Table 5.1 continued

Protein ID (_ECOLI)	# of Peptides	Forward (Aero/ Anaero)	Reverse (Aero/ Anaero)	Avg	Protein ID (_ECOLI)	# of Peptides	Forward (Aero/ Anaero)	Reverse (Aero/ Anaero)	Avg
YFBT	2	0.50	0.85	0.65	YGAC	1	0.55	0.60	0.57
YFCC	1	0.26	0.19	0.22	YGAM	1	0.77	0.48	0.61
YFFB	1	1.98	1.73	1.85	YGAU	3	0.60	0.52	0.56
YGAC	1	0.55	0.60	0.57	YGIQ	1	0.30	0.34	0.32
YGAM	1	0.77	0.48	0.61	YGIW	2	0.46	0.37	0.41
YGAU	3	0.60	0.52	0.56	YHAM	1	1.54	1.51	1.53
YGIQ	1	0.30	0.34	0.32	YHBT	2	0.47	0.54	0.50
YGIW	2	0.46	0.37	0.41	YHDH	1	0.43	0.76	0.57
YHAM	1	1.54	1.51	1.53	YHIR	3	0.74	0.55	0.64
YHBT	2	0.47	0.54	0.50	YHJH	1	0.64	0.64	0.64
YHDH	1	0.43	0.76	0.57	YIAD	3	1.53	1.64	1.58
YHIR	3	0.74	0.55	0.64	YICG	1	1.70	1.68	1.69
YHJH	1	0.64	0.64	0.64	YIFB	1	0.39	0.23	0.30
YIAD	3	1.53	1.64	1.58	YIFE	4	1.70	1.84	1.77
YICG	1	1.70	1.68	1.69	YIHW	1	1.76	4.00	2.65
YIFB	1	0.39	0.23	0.30	YIID	1	3.39	3.20	3.29
YIFE	4	1.70	1.84	1.77	YJBJ	2	0.63	0.50	0.56
YIHW	1	1.76	4.00	2.65	YJEE	1	1.24	3.44	2.07
YIID	1	3.39	3.20	3.29	YJIM	1	0.30	0.31	0.30
YJBJ	2	0.63	0.50	0.56	YJJI	4	0.55	0.41	0.48
YJEE	1	1.24	3.44	2.07	YLAC	1	3.99	4.51	4.24
YJIM	1	0.30	0.31	0.30	YNCE	2	9.47	8.37	8.90
YJJI	4	0.55	0.41	0.48	YNFE	3	0.10	0.17	0.13
YLAC	1	3.99	4.51	4.24	YNFF	5	0.11	0.12	0.11
YNCE	2	9.47	8.37	8.90	YNJE	2	0.33	0.29	0.31
YNFE	3	0.10	0.17	0.13	YQJD	3	0.72	0.57	0.64
YNFF	5	0.11	0.12	0.11	YRBA	1	1.29	3.14	2.01
YNJE	2	0.33	0.29	0.31	YRBA	1	1.29	3.14	2.01
YQJD	3	0.72	0.57	0.64	YCBB	1	1.11	0.38	0.65
YRBA	1	1.29	3.14	2.01	YCCU	2	0.55	0.71	0.62
YCBB	1	1.11	0.38	0.65	YCEF	1	0.65	0.39	0.50
YCCU	2	0.55	0.71	0.62	YCIH	1	0.44	0.37	0.41
YCEF	1	0.65	0.39	0.50	YCIO	2	1.71	1.77	1.74
YCIH	1	0.44	0.37	0.41	YDCP	1	0.61	0.36	0.47
YCIO	2	1.71	1.77	1.74	YDFH	1	0.37	0.79	0.54
YDCP	1	0.61	0.36	0.47	YDFZ	1	0.46	0.43	0.45
YDFH	1	0.37	0.79	0.54	YDGH	3	0.59	0.55	0.57
YDFZ	1	0.46	0.43	0.45	YDJI	1	0.20	0.21	0.21
YDGH	3	0.59	0.55	0.57	YEBE	2	0.46	0.41	0.43
YDJI	1	0.20	0.21	0.21	YEED	1	0.40	0.49	0.44
YEBE	2	0.46	0.41	0.43	YEEE	1	0.46	0.44	0.45
YEED	1	0.40	0.49	0.44	YEEF	2	0.52	0.55	0.54
YEEE	1	0.46	0.44	0.45	YEEX	2	0.66	0.56	0.61
YEEF	2	0.52	0.55	0.54	YEFI	3	0.57	0.52	0.54
YEEX	2	0.66	0.56	0.61	YEGP	1	0.45	0.33	0.39
YEFI	3	0.57	0.52	0.54	YEIA	1	0.50	0.51	0.51
YEGP	1	0.45	0.33	0.39	YEIE	1	0.66	0.61	0.63
YEIA	1	0.50	0.51	0.51	YEIQ	1	2.17	1.84	2.00
YEIE	1	0.66	0.61	0.63	YEIR	3	2.53	2.18	2.35
YEIQ	1	2.17	1.84	2.00	YEIT	2	0.77	0.51	0.63
YEIR	3	2.53	2.18	2.35	YFBT	2	0.50	0.85	0.65
YEIT	2	0.77	0.51	0.63	YFCC	1	0.26	0.19	0.22
YFBT	2	0.50	0.85	0.65	YFFB	1	1.98	1.73	1.85
YFCC	1	0.26	0.19	0.22	YGAC	1	0.55	0.60	0.57
YFFB	1	1.98	1.73	1.85	YGAM	1	0.77	0.48	0.61

Table 5.1 continued

Protein ID (_ECOLI)	# of Peptides	Forward (Aero/ Anaero)	Reverse (Aero/ Anaero)	Avg	Protein ID (_ECOLI)	# of Peptides	Forward (Aero/ Anaero)	Reverse (Aero/ Anaero)	Avg
YRBA	1	1.29	3.14	2.01	YGAM	1	0.77	0.48	0.61
YCBB	1	1.11	0.38	0.65	YGAU	3	0.60	0.52	0.56
YCCU	2	0.55	0.71	0.62	YGIQ	1	0.30	0.34	0.32
YCEF	1	0.65	0.39	0.50	YGIW	2	0.46	0.37	0.41
YCIH	1	0.44	0.37	0.41	YHAM	1	1.54	1.51	1.53
YCIO	2	1.71	1.77	1.74	YHBT	2	0.47	0.54	0.50
YDCP	1	0.61	0.36	0.47	YHDH	1	0.43	0.76	0.57
YDFH	1	0.37	0.79	0.54	YHIR	3	0.74	0.55	0.64
YDFZ	1	0.46	0.43	0.45	YHJH	1	0.64	0.64	0.64
YDGH	3	0.59	0.55	0.57	YIAD	3	1.53	1.64	1.58
YDJI	1	0.20	0.21	0.21	YICG	1	1.70	1.68	1.69
YEBE	2	0.46	0.41	0.43	YIFB	1	0.39	0.23	0.30
YEED	1	0.40	0.49	0.44	YIFE	4	1.70	1.84	1.77
YEEE	1	0.46	0.44	0.45	YIHW	1	1.76	4.00	2.65
YEEF	2	0.52	0.55	0.54	YIID	1	3.39	3.20	3.29
YEEX	2	0.66	0.56	0.61	YJBJ	2	0.63	0.50	0.56
YEFI	3	0.57	0.52	0.54	YJEE	1	1.24	3.44	2.07
YEGP	1	0.45	0.33	0.39	YJIM	1	0.30	0.31	0.30
YEIA	1	0.50	0.51	0.51	YJJI	4	0.55	0.41	0.48
YEIE	1	0.66	0.61	0.63	YLAC	1	3.99	4.51	4.24
YEIQ	1	2.17	1.84	2.00	YNCE	2	9.47	8.37	8.90
YEIR	3	2.53	2.18	2.35	YNFE	3	0.10	0.17	0.13
YEIT	2	0.77	0.51	0.63	YNFF	5	0.11	0.12	0.11
YFBT	2	0.50	0.85	0.65	YNJE	2	0.33	0.29	0.31
YFCC	1	0.26	0.19	0.22	YQJD	3	0.72	0.57	0.64
YFFB	1	1.98	1.73	1.85	YRBA	1	1.29	3.14	2.01
YGAC	1	0.55	0.60	0.57					

in the aerobic samples, respectively) by mass spectrometry, while the same study found threefold up-regulation under anaerobic conditions.[27] Similarly, another study examining the effect of oxygen on a *cyd-lacZ* fusion expression found highest expression at 7% oxygen saturation (microaerophilic conditions), which was reduced to 60% of maximum under anaerobic conditions (0% oxygen) and eight times lower under aerobic conditions (>20% oxygen).[28] The expected fivefold ratio of anaerobic to aerobic expression, or even a general increase under anaerobic conditions,[29] was not observed in the current study. A similar experiment was performed using a *cyo-lacZ* fusion and the ratio of aerobic to anaerobic expression was found to be about tenfold, which is consistent with the current data. The exact rationale for the consistency in the cytochrome *o* oxidase results, but not the cytochrome *d* oxidase results, is unclear.

5.4.2 Iron Regulation

Under oxidizing conditions, it is expected that the available iron in the extracellular space should be the ferric form (Fe^{3+}), rather than the more water-soluble ferrous form (Fe^{2+}). This has practical considerations for bacteria, which require iron for function. Protein *tonB* (TONB_ECOLI, *tonB*) is involved in transport of bound metal complexes[30] in an energy dependent fashion by interacting with membrane proteins. Outer membrane transport proteins for ferrienterobactin (ferric iron bound to the siderophore enterobactin), such as FEPB_ECOLI (*fepB*) and FHUA_ECOLI (*fhuA*), were found to have an 8.48-fold and 12.71-fold increase, respectively, under aerobic conditions. Furthermore, functional stabilizers of *tonB*, EXBB_ECOLI (*exbB*, 5.23-fold increase) and EXBD_ECOLI (*exbD*, 1.81-fold increase), were also found to be present in higher quantities under aerobic conditions. A putative iron transport protein (CIRA_ECOLI, *cirA*) was also identified as 9.38-fold higher under the same conditions.

5.4.3 Energy Metabolism

Changes in the metabolism of *E. coli* are expected as the system shifts from aerobic to anaerobic respiration. Succinate dehydrogenase (SDHD_ECOLI, *sdhD*), succinate-CoA synthetase (SUCC_ECOLI, *sucC*; SUCD_ECOLI, *sucD*) were found to be present under aerobic conditions in greater amounts by approximately 60%, which is consistent with previous reports.[31] While fumarate hydratase A (FUMA_ECOLI; *fumA*)

was present at essentially unchanged concentrations between the two conditions (1.30-fold increase in aerobic samples), fumarate hydratase B (FUMB_ECOLI; *fumB*) was significantly lower in aerobic samples (0.30-fold). Both enzymes catalyze the inter-conversion between malate and fumarate in the citric acid cycle, but *fumB* is subject to anaerobic control for expression.[32, 33] Furthermore, anaerobic conversion of pyruvate into lactate and formate is qualitatively consistent with the observation of relative up-regulation of formate acyl transferase proteins (*pflA*, *pflB*, and *pflF*) under anaerobic conditions.[34] Carbohydrate transport proteins were also found to be differentially expressed. Among the most significantly changed proteins were PTHA_ECOLI (*gutB*) and PTHB_ECOLI (*gutE*), which are both down-regulated approximately tenfold in anaerobic bacteria. Fumarate reductase (FRDA_ECOLI, *frdA*; FRDB_ECOLI, *frdB*) was approximately four times up-regulated under anaerobic conditions, but to a reduced degree than predicted by gene fusion studies.[35, 36] Altogether, the data suggest the many of the proteins expected to change concentration in response to growth under anaerobic conditions were found in the current dataset.

5.5 Conclusion

Reverse labelling was evaluated as a simple method to check the consistency and quality of quantification values obtained from labelled peptides in a 2D-LC-MS experiment. Overall, the consistency in quantification values obtained appears to be internally consistent with over 87% varying by less than 50% at peptide level for all data, before any data analysis measures. It was observed that approximately 5% of peptide data shows at least a 2-fold difference in the observed ratio of replicates. While some large relative differences were genuine, at least 40% of peptides with large differences were found to have qualitatively contradictory ratios and may be artifacts from poor peak fitting or incorrect peptide assignments. By eliminating the 5% of peptides with the highest relative errors, 252 peptides or 46 proteins were removed. However, the remaining proteins form a high quality dataset with consistent behaviour over multiple experiments. At protein level, the similarity in quantification improves to less than a 30% difference for over 95% of all proteins.

These data suggest that peak fitting and data processing should be carefully scrutinized and considered potential sources of error, as a relatively low error rate can become problematic across the large datasets generated from complex LC-MS experiments. Just as reversed/randomized database searching estimates the false positive rate for peptide identifications, errors in reported quantification values are also possible, however infrequent they may be. While it is impractical to manually verify all results from an LC-MS experiment, it is suggested that the *de facto* practice of selective result checking continue, especially for key results that will be the subject of further scientific work. While this chapter is limited to the particular mass spectrometer and the software used, it is believed that the conclusions regarding data processing programs can be readily generalized. Although data processing has become increasingly automated, it appears that careful analyst intervention cannot be readily substituted.

As a discussion of other software programs capable of generating quantitative information from MS-based experiments is beyond the scope of this paper, there have been recent reviews on this topic.[37] It is expected that each program has its own features, advantages, and reliability in results reporting. As the demand for quantification experiments becomes increasingly commonplace, it is expected software capabilities will becoming increasingly accurate as the research field matures.

5.6 Literature Cited

(1) Picotti, P.; Lam, H.; Campbell, D.; Deutsch, E. W.; Mirzaei, H.; Ranish, J.; Domon, B.; Aebersold, R. *Nat. Methods* **2008**, *5*, 913-914.

(2) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen Dan, B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell. Proteomics* **2002**, *1*, 376-386.

(3) Jiang, H.; English Ann, M. *J. Proteome Res.* **2002**, *1*, 345-350.

(4) Fenselau, C.; Yao, X. *J. Proteome Res.* **2009**, *8*, 2140-2143.

(5) Reynolds, K. J.; Yao, X.; Fenselau, C. *J. Proteome Res.* **2002**, *1*, 27-33.

(6) Rao, K. C. S.; Carruth, R. T.; Miyagi, M. *J. Proteome Res.* **2005**, *4*, 507-514.

(7) Thompson, A.; Schaefer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. *Anal. Chem.* **2003**, *75*, 1895-1904.

(8) Morano, C.; Zhang, X.; Fricker Lloyd, D. *Anal. Chem.* **2008**, *80*, 9298-9309.

(9) Kuzyk, M. A.; Ohlund, L. B.; Elliott, M. H.; Smith, D.; Qian, H.; Delaney, A.; Hunter, C. L.; Borchers, C. H. *Proteomics* **2009**, *9*, 3328-3340.

(10) White, C. A.; Oey, N.; Emili, A. *J. Proteome Res.* **2009**, *8*, 3653-3665.

(11) Kierszniowska, S.; Walther, D.; Schulze, W. X. *Proteomics* **2009**, *9*, 1916-1924.

(12) Hebeler, R.; Oeljeklaus, S.; Reidegeld, K. A.; Eisenacher, M.; Stephan, C.; Sitek, B.; Stuehler, K.; Meyer, H. E.; Sturre, M. J. G.; Dijkwel, P. P.; Warscheid, B. *Mol. Cell. Proteomics* **2008**, *7*, 108-120.

(13) Ji, C.; Guo, N.; Li, L. *J. Proteome Res.* **2005**, *4*, 2099-2108.

(14) Boersema, P. J.; Raijmakers, R.; Lemeer, S.; Mohammed, S.; Heck, A. J. R. *Nat. Protoc.* **2009**, *4*, 484-494.

(15) Zhang, R.; Sioma Cathy, S.; Thompson Robert, A.; Xiong, L.; Regnier Fred, E. *Anal. Chem.* **2002**, *74*, 3662-3669.

(16) Wang, N.; Xie, C.; Young, J. B.; Li, L. *Anal. Chem.* **2009**, *81*, 1049-1060.

(17) Wang, N.; Li, L. *Anal. Chem.* **2008**, *80*, 4696-4710.

(18) Ji, C.; Lo, A.; Marcus, S.; Li, L. *J. Proteome Res.* **2006**, *5*, 2567-2576.

(19) Serres, M. H.; Gopal, S.; Nahum, L. A.; Liang, P.; Gaasterland, T.; Riley, M. *Genome Biol* **2001**, *2*, Research0035.0031-Research0035.0037.

(20) Misra, R. V.; Horler, R. S. P.; Reindl, W.; Goryanin, I. I.; Thomas, G. H. *Nucleic Acids Res.* **2005**, *33*, D329-D333.

(21) Salgado, H.; Santos-Zavaleta, A.; Gama-Castro, S.; Peralta-Gil, M.; Penaloza-Spinola, M. I.; Martinez-Antonio, A.; Karp, P. D.; Collado-Vides, J. *BMC Bioinf.* **2006**, *7*, No pp given.

(22) Karp, P. D.; Keseler, I. M.; Shearer, A.; Latendresse, M.; Krummenacker, M.; Paley, S. M.; Paulsen, I.; Collado-Vides, J.; Gama-Castro, S.; Peralta-Gil, M.; Santos-Zavaleta, A.; Penaloza-Spinola, M. I.; Bonavides-Martinez, C.; Ingraham, J. *Nucleic Acids Res.* **2007**, *35*, 7577-7590.

(23) Keseler, I. M.; Bonavides-Martinez, C.; Collado-Vides, J.; Gama-Castro, S.; Gunsalus, R. P.; Johnson, D. A.; Krummenacker, M.; Nolan, L. M.; Paley, S.;

Paulsen, I. T.; Peralta-Gil, M.; Santos-Zavaleta, A.; Shearer, A. G.; Karp, P. D. *Nucleic Acids Res.* **2009**, *37*, D464-D470.

(24) Salmon, K. A.; Hung, S.-p.; Steffen, N. R.; Krupp, R.; Baldi, P.; Hatfield, G. W.; Gunsalus, R. P. *J. Biol. Chem.* **2005**, *280*, 15084-15096.

(25) Loui, C.; Chang, A. C.; Lu, S. *BMC Microbiol.* **2009**, *9*, doi:10.1186/1471-2180-9-183.

(26) Georgellis, D.; Kwon, O.; Lin, E. C. C. *Science* **2001**, *292*, 2314-2316.

(27) Cotter, P. A.; Cherpuri, V.; Gennis, R. B.; Gunsalus, R. P. *J. Bacteriol.* **1990**, *172*, 6333-6338.

(28) Tseng, C.-P.; Albrecht, J.; Gunsalus, R. P. *J. Bacteriol.* **1996**, *178*, 1094-1098.

(29) Govantes, F.; Albrecht, J. A.; Gunsalus, R. P. *Mol. Microbiol.* **2000**, *37*, 1456-1469.

(30) Wooldridge, K. G.; Morrissey, J. A.; Williams, P. H. *J. Gen. Microbiol.* **1992**, *138*, 597-603.

(31) Park, S.-J.; Chao, G.; Gunsalus, R. P. *J. Bacteriol.* **1997**, *179*, 4138-4142.

(32) Tseng, C.-P. *FEMS Microbiol. Lett.* **1997**, *157*, 67-72.

(33) Tseng, C.-P.; Yu, C.-C.; Lin, H.-H.; Chang, C.-Y.; Kuo, J.-T. *J. Bacteriol.* **2001**, *183*, 461-467.

(34) Sawers, G.; Bock, A. *J. Bacteriol.* **1988**, *170*, 5330-5336.

(35) Jones, H. M.; Gunsalus, R. P. *J. Bacteriol.* **1987**, *169*, 3340-3349.

(36) Condon, C.; Weiner, J. H. *Mol. Microbiol.* **1988**, *2*, 43-52.

(37) Mueller, L. N.; Brusniak, M.-Y.; Mani, D. R.; Aebersold, R. *J. Proteome Res.* **2008**, *7*, 51-61.

Chapter 6 – Preliminary Evaluation of Elk Plasma Biomarkers

6.1 Introduction

Bovine spongiform encephalopathy, informally known as BSE or mad cow disease, gained worldwide prominence in the mid-1990s as a fatal food-borne infection that became widespread in the United Kingdom. Bovine spongiform encephalopathy was found to be transmitted through meat-and-bone meal used to supplement cattle feed. The unsellable portions of cattle, including the carcass, spinal column, and brains, were homogenized into a ground meal that was added to conventional cattle feed. The infectious agent was transmitted to other animals through the feed system, amplified *in vivo*, and returned to the feed system after the cows were brought to slaughter. In 2003, Canada had its first case of BSE and since then infrequent cases have occurred at the rate of about two cases per year. Food safety concerns in foreign countries and a lack of consumer confidence in Canadian beef products resulted in economic losses in excess of $6 billion.[1]

The causative agent for mad cow disease was found to be the misfolded form of the prion protein, in which the prion protein adopts a non-native conformation that slowly develops plaques within the central nervous system. This misfolded, infectious form of the prion protein is characterized as having the same primary sequence as the native prion protein with no additional post-translation modifications, but is noted for being highly protease resistant.[2] Most enzymes used for general protein degradation, such as proteinase K, cannot digest the misfolded prion protein entirely; as such, the misfolded form is known as PrP^{res} (protease resistant prion protein) or PrP^{Sc} (prion protein from scrapie, the analogue of BSE in sheep). The cellular, native form of the prion protein (PrP^{C}) does not have a clear biological role, but has been shown to have functions such as copper ion binding.[3, 4] Furthermore, unlike PrP^{Sc}, cellular PrP is protease sensitive and can be digested entirely by proteinase K. PrP^{Sc} is thought to propagate in one of two mechanisms: template mediated synthesis and seed-mediated synthesis. In both cases, a misfolded prion protein acts as a base for further conversion.

Prion-based diseases, such as chronic wasting disease (CWD) in elk, remain an issue for wildlife conservation.[5] The spread of the chronic wasting disease in deer and elk populations has been steadily increasing, both in terms of the number of animals infected and the geographical range. Furthermore, the rapidity of transmission in wild animal populations has been largely unexpected. Unlike cattle in the food system, there is no obvious route of infection or contact between infectious material in elk and the general populations. While various vectors, such as feces,[6] urine,[7] and environmental contamination,[8-10] have been suggested, there has been no clear indication of the exact mechanism of infection. One of the primary challenges in addressing the spread of chronic wasting disease is the lack of an effective ante mortem test. Without adequate tools to track and monitor progression of the disease, it may be difficult to implement population control strategies. To look for potential markers of CWD infection, the plasma of orally infected elk was analyzed at various time points using MS-based proteomics methods. Plasma is a potential source for biomarker discovery, as various circulating proteins have immunological functions. However, there are three key challenges to address when working with elk plasma.

Plasma is a challenging biological matrix to work with due to the presence of several high abundance proteins, such as albumin, immunoglobulins, and various components of the complement system. In humans, the twenty proteins highest in concentration constitute over 98% of the total protein content in plasma by weight.[11] If lower concentration proteins of interest are to be studied, protein separation must be performed to remove or separate these high abundance components from other proteins. As an additional step performed prior to isotopic labelling for quantitative proteomics, this can potentially introduce errors into the quantification ratios observed. The second challenge is the lack of an elk genome sequence database. Since MS/MS spectra rarely contain sufficient information for complete *de novo* sequencing of peptides, spectra are searched against genomic or protein sequence databases that drastically reduce the search space of potential peptides. Without a sequence database available, fewer peptide identifications are expected as a combination of *de novo* sequencing and searches against other similar, but non-exact, sequence databases can only partially compensate for the missing information.

Lastly, due to the nature of the plasma samples, there are restrictions on where and how the samples may be processed for analysis. Since prions have been identified by the Canadian Food Inspection Agency as an infectious agent, they must be treated in a Biosafety Level II plus laboratory and have the infectious prion protein eliminated from the sample prior to analysis in a lower level laboratory. Due to limitations on laboratory space and equipment available, there were restrictions on the available analytical methods. Our initial efforts to describe proteomic changes in elk plasma after oral infection with chronic wasting disease material are presented. A combination of protein separation methodologies and isotopic labelling were evaluated for quantitative analysis of protein changes.

6.2 Experimental

6.2.1 Chemical and Reagents

LC-MS grade solvents (water, methanol, and acetonitrile), Pierce BlueDye™ Albumin depletion columns, and BCA Assay Kit were obtained from Fisher Scientific (Edmonton, AB). LC-MS grade formic acid, LC-MS grade trifluoroacetic acid, ^{12}C formaldehyde, and 2-picoline borane complex were obtained from Sigma Aldrich (Oakville, ON). ^{13}C formaldehyde was obtained from Cambridge Isotope Laboratories (Andover, MA). Precast 12% gels were obtained from Invitrogen (Burlington, ON) and 4-15% SDS-PAGE gels were obtained from Bio-Rad (Mississauga, ON).

6.2.2 Elk Plasma Samples

Three plasma samples were obtained from three uninfected elk and used in preliminary studies and albumin depletion tests. Samples of plasma from infected elk were obtained from the Canadian Food Inspection Agency with the generous assistance of Dr. Catherine Graham at the Lethbridge CFIA laboratory. A set of three different elk were orally infected with brain homogenate from a CWD-infected elk and blood samples were taken at zero, seven, and twelve months post-infection (mpi) and at the terminal end stage. The terminal end stage was determined by clear observation of clinical symptoms of chronic wasting disease (e.g., ataxia) and ranged between

twenty to twenty-four months. Clarified plasma samples were aliquoted for separate analyses and stored at -80 °C pending further analysis.

6.2.3 Albumin Depletion

Albumin depletion was evaluated using uninfected elk plasma samples. The depletion was originally performed as per the manufacturer's instructions. During method optimization, modifications were made by doubling the amount of the binding resin used and increasing the strength of the stripping buffer by doubling the salt concentration to 1 M KCl or using brine. Depletion efficiency was evaluated by separating the original sample, flow-through, and eluted bound fractions by SDS-PAGE on a 4-15% gradient gel run at constant current using 100V for 100 minutes.

6.2.4 In-Gel Digestion

Sample preparation of plasma from orally infected elk was carried out at the Centre for Prions and Protein Folding Diseases at the University of Alberta in a Biosafety Level II plus laboratory with the appropriate safety precautions. Personal protective equipment used included a lab gown, face shield, shoe covers, and double gloves. Modified laboratory procedures were used to minimize potential contamination of the laboratory with samples. Protein concentrations for the twelve plasma samples from infected elk (3 animals at four time points) were obtained using the BCA assay. A reference sample was prepared by mixing together equal protein weights of each of the twelve samples. Each of the timepoint samples was separated on four lanes of a 12% linear gel along with four lanes of the reference sample; a total of 200 µg of sample was used per lane based on the BCA quantification results. The gel was run at constant current using 75V for 30 minutes followed by 150V for 80 minutes. The gels were visualized by Coomassie dye and the top portion of the gel corresponding to the high molecular weight portion (>148 kDa, based on molecular weight standards), was discarded due to the potential of infectious prion multimers in the higher molecular weight bands. The remaining portions of the gel were cut into six molecular weight ranges using a scalpel with cut-resistant gloves. Two lanes of each timepoint or reference were pooled for a single extraction and were marked for either light or heavy isotopic labelling. Gel pieces were minced into 1 mm^3 cubes, dehydrated

with acetonitrile twice, rehydrated with trypsin (8 ng/µL) in 100 mM NH_4HCO_3, covered with 100 mM NH_4HCO_3 and incubated overnight at 37 °C. Peptides were extracted using 50% acetonitrile and 0.25% TFA in water, followed by a second extraction with 75% acetonitrile and 0.25% TFA; extracts were pooled together for downstream processing. Tubes with peptide digests were soaked in 1M NaOH for 1 hour to decontaminate vials prior to removal from the laboratory.

6.2.5 Isotopic Labelling

Peptide digests were dried down in a vacuum centrifuge to near dryness and reconstituted using 0.1% TFA and peptide amounts were desalted and quantified.[12] The desalted peptides were dried down and reconstituted using 100 mM $NH(CH_3)_3HCO_3$. Samples were labelled using reductive methylation using 8 µL 1 M 2-picoline borane and 4 µL 4% formaldehyde. Each set of extracts was labelled with light labelled formaldehyde ($^{12}CH_2O$) or heavy labelled formaldehyde ($^{13}CD_2O$), based on the original labelling assignment from the gel extraction phase. Samples were acidified with TFA to pH <2 and desalted/quantified as previously described. Desalted peptides were dried in a vacuum centrifuge and reconstituted to 0.2 µg/µL using 0.1% formic acid.

6.2.6 Sample Mixing, MS Analysis, and Data Processing

Paired timepoint samples and reference samples were mixed together in a 1:1 ratio, by peptide weight, using both the forward labelling scheme (light labelled timepoint digest with heavy labelled reference digest) and reverse labelling scheme (light labelled reference digest with heavy labelled timepoint digest). Samples were separated on a 300 µm i.d. x 150 mm Discovery C_{18} column using a Waters nanoAcquity LC followed by analysis on a Waters ESI-QTOF Premier mass spectrometer. Data was processed using MASCOT Distiller for quantification and searched against the MASCOT v. 2.2 search engine using the following search parameters (enzyme: trypsin; missed cleavages: 2, variable modifications: Dimethylation_Light (N-term; $+C_2H_4$), Dimethylation_LightK (K; $+C_2H_4$), Dimethylation_Heavy (N-term; $+^{13}C_2D_4$), Dimethylation_HeavyK (K; $+^{13}C_2D_4$); MS tolerance: 30 ppm; MS/MS tolerance: 0.2 Da). The dimethylation modifications were

searched as a paired set (i.e., light and heavy dimethylation were not allowed on the same peptide). A modified instrument-type setting using standard ESI-QTOF fragmentations further allowing a-ions was used. A confidence threshold of 95% was used for initial peptide sequencing. Various search databases (Swiss-Prot and NCBInr) were used with different taxa (*bos*, other mammals, and *homo sapiens*) and the results were pooled together for further analysis. Quantification results were taken from MASCOT Distiller and processed using Microsoft Excel. Identical peptides from different fractions were considered equivalent and averaged geometrically. Identical sequences in different charge states were also averaged geometrically.

6.3 Results and Discussion

Plasma samples were obtained from three CWD orally infected elk at 0, 7, and 12 months post-infection and at the onset of clinical symptoms. A time course experiment design allows for comparison of results from multiple points that tracked disease progression (Figure 6.1). A mixed reference was prepared by mixing together equal aliquots, by weight, of each of the twelve samples. The rationale was to allow for equal comparison of all twelve timepoints against a common sample, which would improve the ability to track changes across the samples, since the average of the ratios from the twelve samples should be unity. From a quantification perspective, the use of a mixed reference would restrict the theoretical range of quantification from zero (absent in the sample) to twelve (a single timepoint contributes all of the signal intensity found in the reference) and allow erroneous or spurious values to be easily detected. This is important in cases where only a single peak, instead of a pair, was observed. Since peptide sequence misassignments were possible, particularly given the limited sequence information available in the study, this provided an additional criterion for discriminating peaks suggesting significant changes.

6.3.1 Albumin Depletion

As with the plasma of most mammals, albumin is a primary component of elk plasma (Figure 6.2). Since the presence of albumin would mask the signal from lower abundance proteins, removal of albumin using a protein-level separation method was

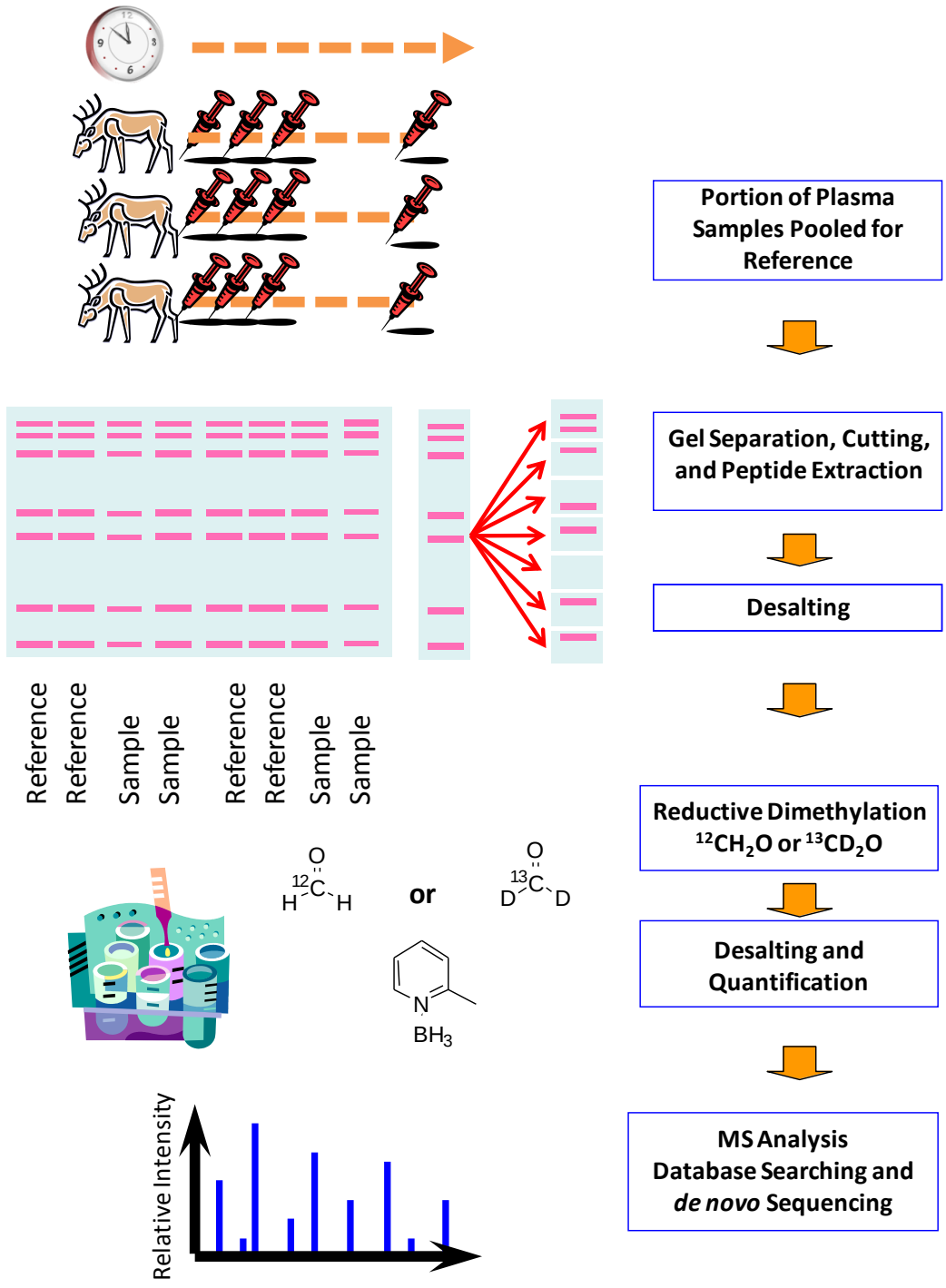
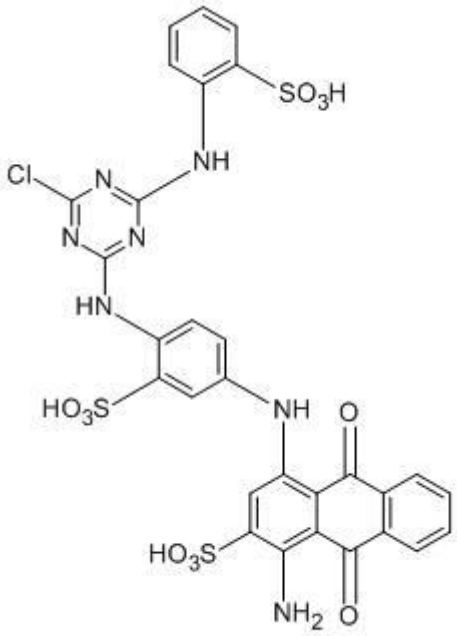


Figure 6.1 Experimental workflow for time-course experiment to compare plasma from CWD-infected elk

A)



B)

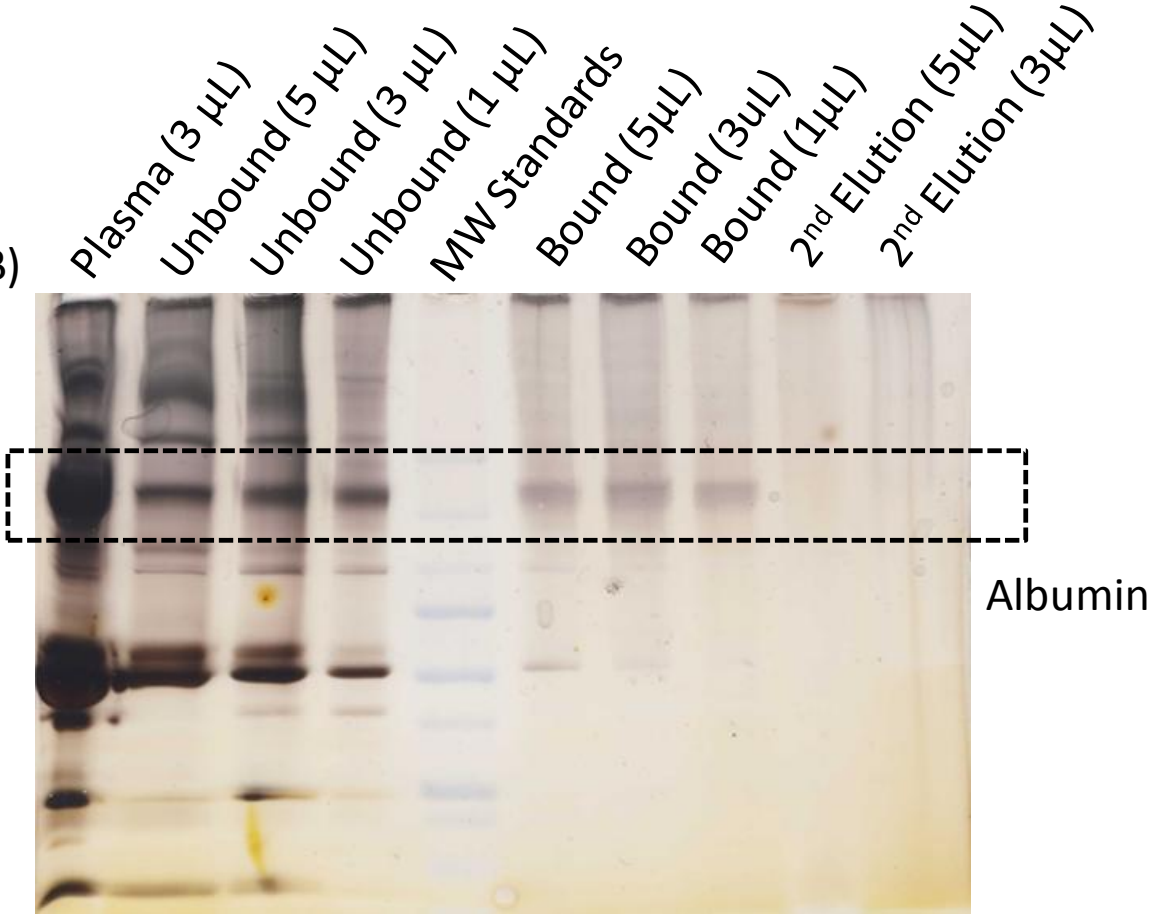


Figure 6.2 a) Structure of Cibacron Blue dye used for albumin binding and b) initial attempts to remove albumin using dye columns

explored. With the additional restriction of the limited equipment within the enhanced Biosafety Level II plus facility, spin column formats were the main option for albumin depletion. The most frequently used method for albumin and high-abundance protein depletion is immunoaffinity columns that use antibodies raised against specific high-abundance plasma proteins.[13] Commercially available products are available for analyzing plasma from various research and pharmaceutical applications: human, rat, and mouse. Since antibodies grown against elk albumin were not commercially available, traditional depletion methods that exploit the anion binding capacity of albumin were considered. Albumin is known to have a binding affinity for certain anionic species, such as RNA and particular anionic dyes. We considered the use of a dye resin packed with Cibacron Blue dye (Figure 6.2A), which has been previously reported for albumin depletion.[14]

Initial tests using conditions outlined in the manufacturer's instructions led to only a modest albumin depletion effect as shown in Figure 6.2B. The flow-through fraction was found to contain a significant portion of albumin and the eluted fraction of bound proteins showed non-specific binding. To improve resin binding, samples were diluted with a neutral pH buffer that reduces inter-protein interactions. Different salt concentrations were used in the dilution buffer along with plain deionized water. In all cases, no significant improvement in protein binding was observed. The amount of resin was also increased in order to aid in albumin retention. However, only a slight improvement in albumin binding was observed (data not shown). This effect was not entirely unexpected, since it is known that the albumin from different species have variable binding affinities to Cibacron Blue, with some species showing little to no binding. Since proteins other than albumin can bind with anionic dyes and non-specific binding can occur to the resin support,[15] quantification accuracy can be affected by the inconsistent loss of protein between samples, as losses would occur before they can be tracked by the peptide-level labelling. Due to the limited efficiency of the albumin removal and potentially deleterious impact on quantification accuracy, this dye-based albumin depletion method was not employed.

6.3.2 SDS-PAGE Analysis of Samples

In order to separate albumin from other proteins, SDS-PAGE was used to separate samples at protein level. As with any protein separation methodology, care must be taken to minimize variability since protein loss or unevenness in fractionation can cause artifacts in peptide quantification. Although the loading capacity of a single SDS-PAGE gel lane is typically 40-50 µg, an increased loading of 200 µg was used as most of the proteins in plasma, by weight, are albumin and other high abundance proteins. While increasing the loading capacity diminished chromatographic resolution, sufficient material was required for downstream processing and MS analysis. It is also noted the highest molecular weight proteins (>148 kDa), were discarded, since these bands could be contaminated with high molecular weight prion multimers.[16] Although the presence of prion multimers could affect the protein concentration, especially when comparing 0 mpi samples to later samples, variations in multimer abundance are believed to be small when compared to potential variations in high abundance proteins (i.e., albumin or immunoglobulins). Furthermore, by mixing equal peptide amounts post-extraction, as discussed below, the protein amounts are effectively normalized in each fraction with the other proteins in a similar molecular weight range.

Since inconsistency in separation behaviour, gel cutting, and in-gel digestion can lead to quantification inaccuracy, gels were loaded with a total of four lanes of a single time point and four lanes of the mixed reference sample, as shown in Figure 6.3. The gels were halved and each half used for a forward or reciprocal labelling experiment. The final isotopic labels were reversed between the two replicates in order to minimize sample processing variations. It was anticipated that performing a technical replicate would minimize analyst variability from a variety of inconsistencies in sample preparation, particularly with fractionation and extraction of peptides from gels. Peptide quantification results have revealed that differences in fractionation between samples can be a major source of error, if extraction efficiency was assumed to be equal.[17] The LC gradient used for desalting and quantification was modified to elute peptides at 60% acetonitrile, rather than at 90%, in order to separate peptides from residual Coomassie dye used for gel visualization and SDS that was strongly retained to the column.

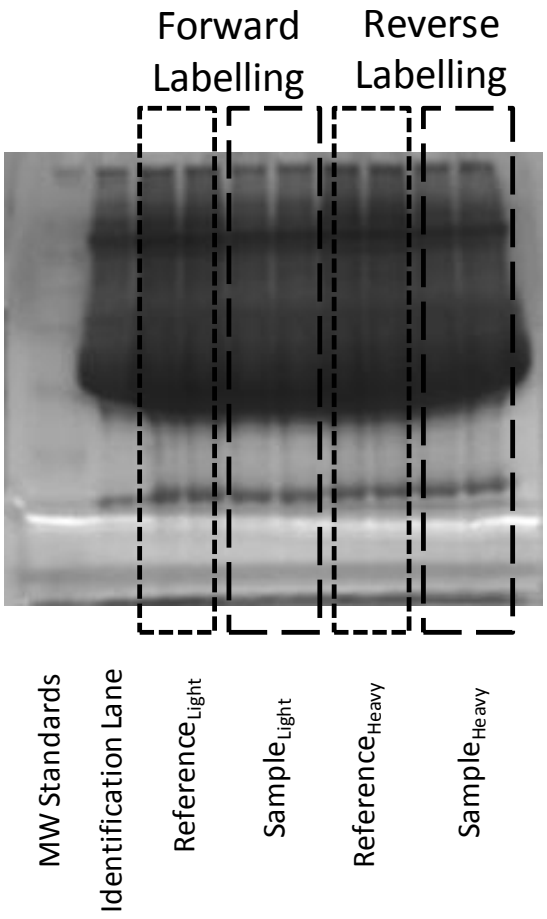


Figure 6.3 SDS-PAGE separation of the reference and a plasma sample (elk #8, 0mpi)

Although the original study design intended to use the 2MEGA labelling method for quantitative isotopic labelling, there was concern that the low peptide amount would lead to inconsistencies in the labelling. Dividing the signal intensity across correctly and incorrectly labelled variants of the same peptide could also have an effect on identification efficiency; high quality spectra would be required for sequencing peptides *de novo* or against non-specific databases. As additional sample was not readily available, reductive methylation was chosen for isotopic labelling. If the reaction was found to be incomplete by MS, the samples could be dried down, reconstituted in the reaction buffer, and reacted again without loss of overall conversion efficiency.

6.3.3 Peptide Sequencing

Since there was no elk sequence database available, peak lists were searched against the two largest databases, SwissProt and NCBInr, using three different taxa: *bos*, other mammals, and *homo sapiens*. A significant portion of the bovine genome has been sequenced and sequence overlaps are expected due to the genetic similarity of cows and elk.[18] Similarly, the "other mammals" database was also exploited in order to evaluate similar sequences that may have amino acid substitutions. While not genetically similar, human genetic sequences were also utilized, since the human database is the largest and most complete sequence database available. An initial search requiring exact sequence matches was used in order to identify proteins with high confidence. A second "error tolerant" search was then performed utilizing only the identified proteins, but allowing single amino acid substitutions and other modifications in order to increase sequence coverage. Even with exact sequence matches, manual verification was used in order to confirm assignments. With careful manual analysis, some sequences were identified with high confidence. An initial spectral assignment to a peptide from the bovine database was manually interpreted in order to solve the middle stretch of the peptide sequence that was originally assigned to an isobaric stretch of amino acids as shown in Figure 6.4. As noted in Chapter 2, reductive dimethylation at peptide N-termini generally leads to MS/MS spectra with a pronounced a_1 ion and primarily y-ions in the remainder of spectra, which simplifies *de novo* sequencing.[19]

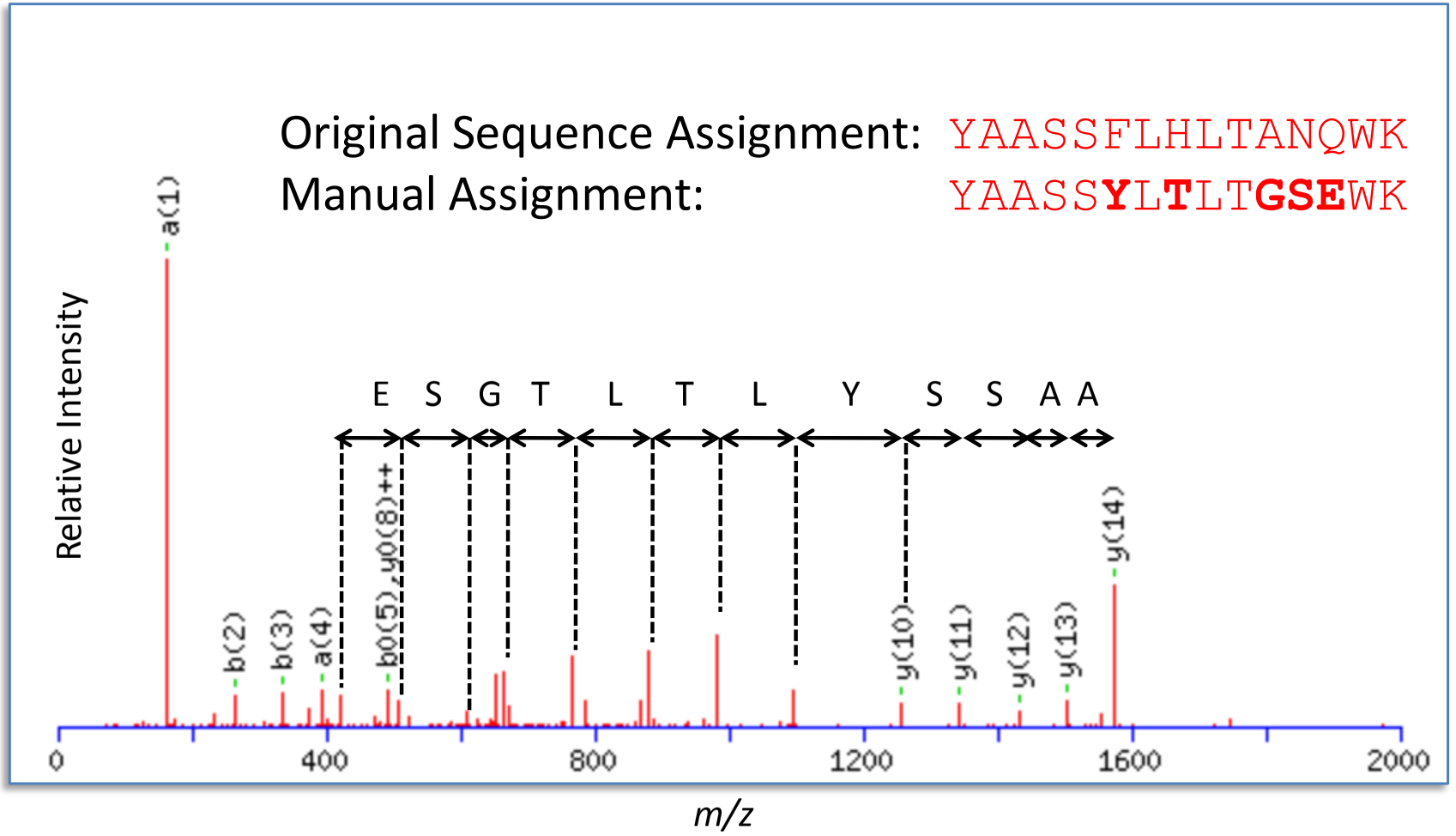


Figure 6.4. Peptide sequencing using a combination of MASCOT and manual *de novo* sequencing.

De novo sequencing was also attempted in order to find sequences containing more than one amino acid substitution. An initial pass of the peak list was processed and *de novo* sequence assignments were evaluated based on their ion score. An ion score is based on the number of peaks in a given spectrum consistent with an assigned peptide sequence. While matching peaks increase the score, unassigned peaks, either real or noise, decrease the ions score, with minor adjustments for peak intensities. A minimal ion score of 65 was used as the quality filter and the *de novo* sequences were submitted for a BLAST search against a sequence database to determine potential protein matches.[20-22] Although very few new proteins were identified with the strategy, an example peptide is shown in Figure 6.5. An initial *de novo* search of the spectrum yielded confident assignments for the N-terminus of the peptide, as evidenced by the short stretch of assigned b- and y-ions. Here, the unclear amino acid assignment is marked with a lower case "i" to denote the isobaric isoleucine and leucine. A BLAST search of the potential *de novo* sequence tags against the *bos* database yields a match to a predicted protein CD63 antigen which has a sequence stretch of VSITKGCGINFSIK, which is similar to one potential interpretation of the spectrum: VSITKGCVIESK. The *de novo* sequence assignment, if correct, would have three different mutations: conversion of a glycine to a valine, conversion of an asparagine to a glutamine, and deletion of the isoleucine. Confident amino acid stretches not already found in the database searches were comparatively rare and there is no simple way to gauge the accuracy of the protein assignments made with the various potential *de novo* interpretations. As such, these tentative protein matches were not used for quantification.

6.3.4 Quantification

Peptide quantification was performed using MASCOT Distiller. Peptide exact matches found through database searching were processed normally, while peptides identified with single amino acid substitutions were manually entered into a local search database and processed separately. The ratios observed between the forward and reciprocal labelled experiments was found to be less consistent than with the *E. coli* study in Chapter 5, likely due to inconsistencies in gel separation and extraction. Although the sample mixing was based on equivalent weights of a time point and its

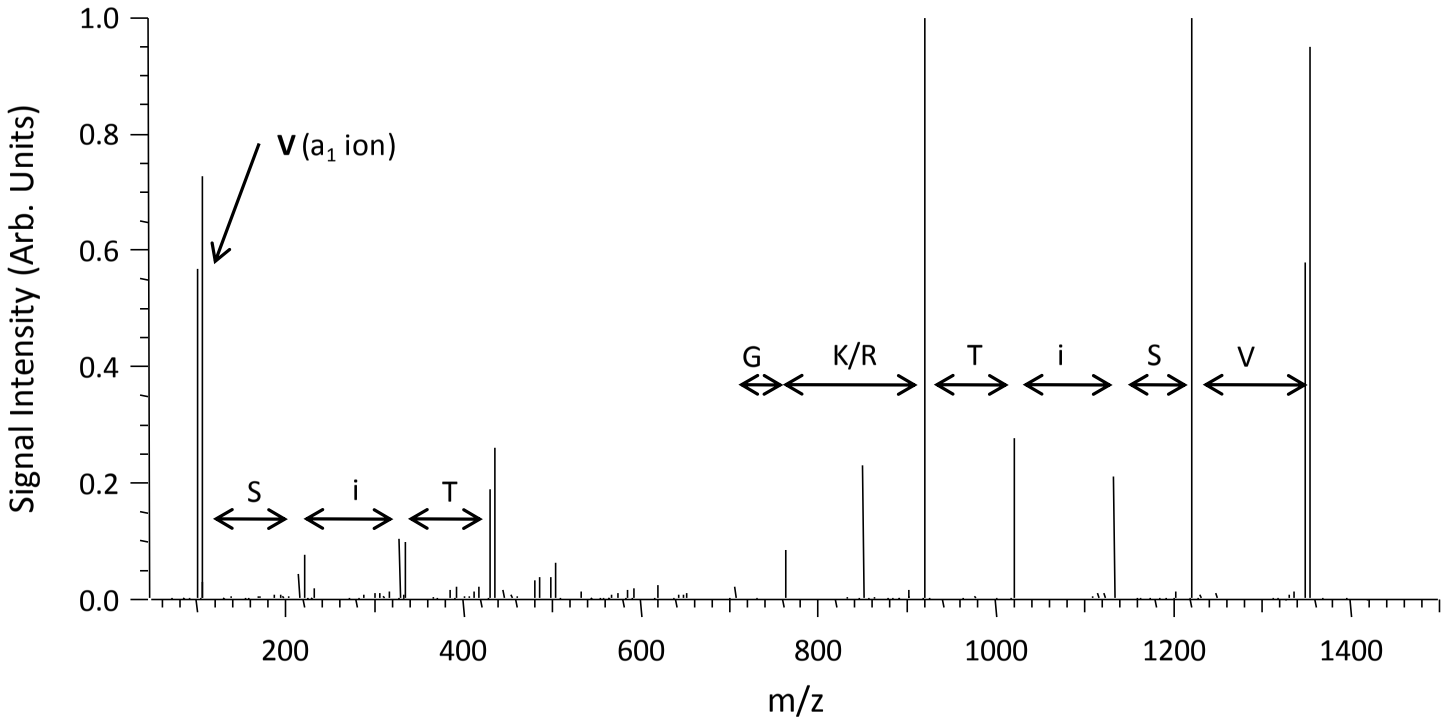


Figure 6.5. ***De novo*** **sequencing yields a short stretch of amino acids, confirmed by both b-ion and y-ion matches; the short stretch was used as the basis for a BLAST search. It is noted that the a_1 ion and b-ions show up as 6 Da pairs due to both the light and heavy forms of the peptides being (unintentionally) allowed through during MS/MS.**

corresponding lane in the reference sample, slight differences in extraction efficiency would be inappropriately reflected as genuine concentration changes in the biological sample. Not surprisingly, the relative differences tend to be larger than for other systems, due to the additional sample preparation steps prior to sample labelling and mixing.

One of the disadvantages of using reductive methylation, when compared to 2MEGA, is that arginine and lysine dimethylated with $^{12}CH_2O$ formaldehyde are nearly isobaric[23] (Δm = 0.0251 Da) and difficult to distinguish with an ESI-QTOF mass spectrometer with a database search tolerance of 30 ppm for the precursor/MS scan. Hence, identification of the light form of a peptide does not necessarily provide unambiguous assignment of the C-terminal residue. However, the heavy labelled forms have different masses, as a lysine containing peptide dimethylated with $^{13}CD_2O$ will be +12 Da heavier than the light form (two heavy N-terminal methyl groups and two heavy groups on the lysine side chain), whereas an arginine containing peptide will only be +6 Da heavier (two heavy N-terminal methyl groups). The assignment of this residue is important when looking for the second peak for quantification (Figure 6.6). If peptides were reported with ratios toward the outside of the acceptable range (i.e., 0 to 12), their sequences were added to a local database by allowing both the lysine and arginine sequence variants. If one of the two values was clearly outside of the allowable range, the other value was used; in cases where both values were within the range, the peptide quantification was manually verified.

6.3.5 Protein Level Results

Through the forward and reverse labelling of samples, consistency between the two replicates was determined by use of $\log_2$-$\log_2$ plot of the protein level ratios obtained from both experiments (Figure 6.7). Proteins were grouped based on the gene identifiers, even if the species used to identify the sequences were different. Most of the values are internally consistent and show the expected behaviour and lie near the line with slope equal to one. However, one protein was found to have inconsistent behaviour and was identified as human keratin, a common contaminant in gel-based workflows.

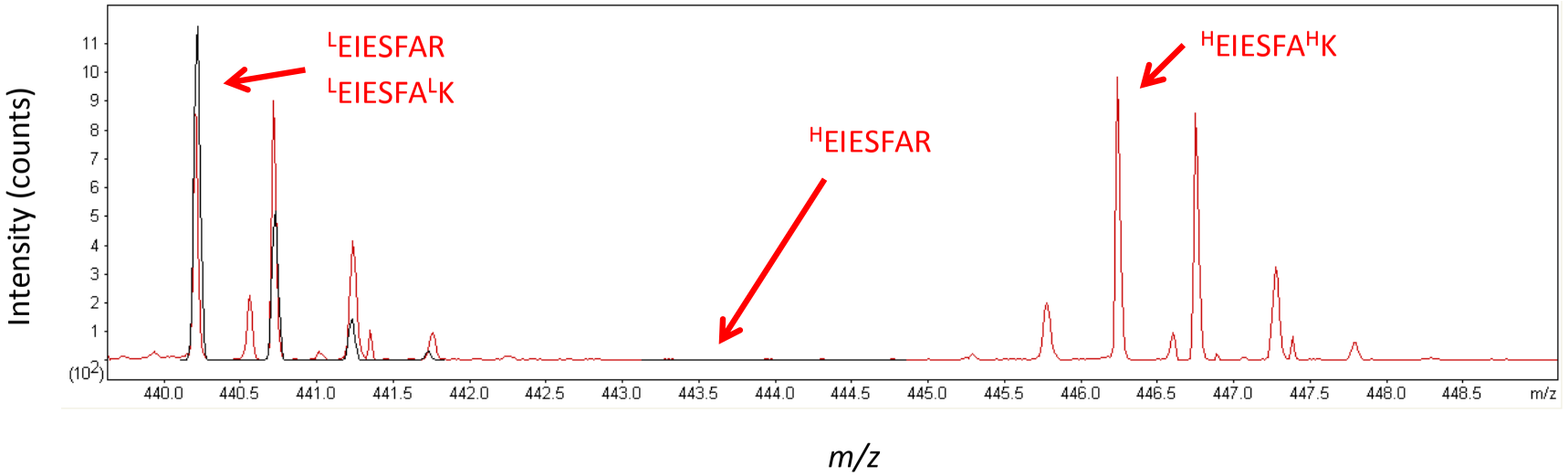


Figure 6.6 MS spectrum of EIESFAK/EIESFAR. The initial spectral assignment to the light peak was EIESFAR, which does have a corresponding pair. If the assignment is switched to the isobaric EIESFAK, the paired peak can be found. Superscripts denote modification at the subsequent residue with either the light methyl groups ($+C_2H_4$, L) or the heavy methyl groups ($+^{13}C_2D_4$, H)

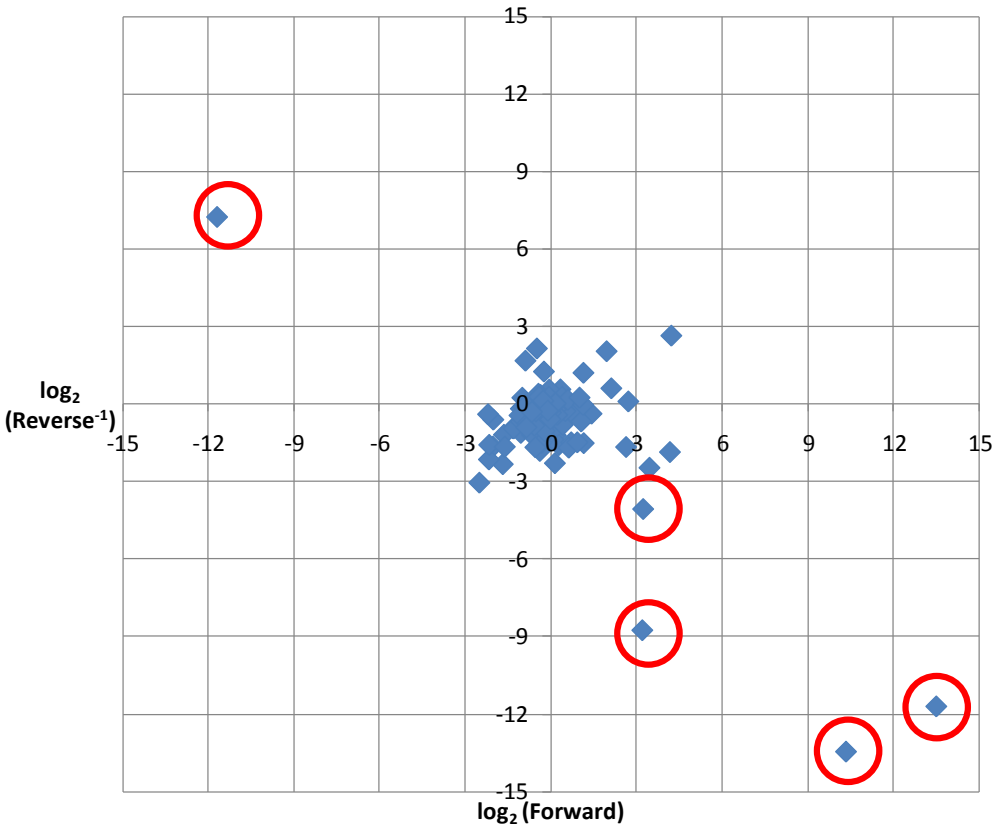


Figure 6.7 log_2-log_2 plot of proteins from elk #8 at the terminal time point. Final protein ratios from the forward and reverse labelling should yield consistent results. Circled proteins represent proteins with mismatched peptides that had lysine/arginine substitutions or were contaminants (keratin).

After half of the samples were processed, the preliminary results were evaluated to determine the state of the completed progress and whether it would be worthwhile to complete analysis of the remaining samples. Although the dataset was incomplete, the data could be processed pairwise to evaluate whether or not consistent ratios were being obtained. The data for each animal was normalized against its 0 mpi sample and the trend for the proteins were taken across the three different time points (7 mpi, 12 mpi, and at the terminal end stage). The inconsistency between the normalized forward and reverse values obtained was with a 1.50-fold error on average. For most of the proteins identified, no clear trend with respect to time was observed (Table 6.1). Another issue encountered for the consistent analysis of proteins in plasma was the lack of proteome coverage. After removal of keratins, which are common contaminants during gel processing, only 48 proteins were consistently identified across all of the samples and could be used as a basis for comparison. One protein previously reported to changed with disease state was haptogloblin.[24] However, a clear pattern over time was not observed.

6.4 Conclusion

The applicability of protein-level separation for use with peptide-level labelling in quantification experiments is ultimately limited by the consistency in the protein separation method employed. Although the cutting of gels and peptide extraction was found to introduce some errors into the overall quantification accuracy of the method, the results were found to be qualitatively consistent in the limited dataset studied. By splitting a protein across two different fractions, the observed ratio for that protein in each of the two fractions may be different, especially when compared against a reference run in an adjacent lane. It is noted that only six fractions were taken, which is typically far less than for a standard identification experiment. While the pattern of the sample loading used in the gels can minimize these variations, care must be taken to ensure that the samples in each lane run straight and that band excision is perpendicular to the sample lane. Most reports today utilizing SDS-PAGE front end separations tend to use metabolic or protein-level labelling approaches that will not be affected by variations in chromatographic separation.

Table 6.1 Consistently identified proteins across all six elk plasma sample analyzed

		Elk #18 (7 mpi)		Elk #37 (12 mpi)		Elk #8 (Terminal)	
Gene ID	**# of Peptides**	**Forward**	**Reverse**	**Forward**	**Reverse**	**Forward**	**Reverse**
A1AG	**1**	0.47	0.13	0.41	0.95	0.48	0.41
A1AT	**9**	0.82	1.03	0.90	1.26	0.75	0.94
A1BG	**4**	0.73	0.93	0.79	0.92	0.88	0.53
ADIPO	**2**	0.73	0.76	1.63	1.95	0.71	1.68
ALBU	**21**	0.83	0.72	1.14	1.26	1.42	1.65
ANGT	**1**	0.64	0.78	0.47	0.95	0.46	0.68
AOCX	**1**	1.08	0.33	0.99	1.15	0.87	0.62
APOA1	**14**	1.14	0.99	2.52	0.70	1.84	0.74
APOD	**1**	0.90	0.93	1.20	1.41	0.76	2.22
AT8A1	**1**	1.24	0.84	0.98	0.64	0.88	0.96
CERU	**2**	0.56	4.27	2.01	1.24	0.55	0.67
CFAB	**5**	1.26	1.12	0.72	1.24	0.80	0.98
CFAH	**1**	0.98	0.64	1.02	28.25	0.73	0.87
CLUS	**2**	0.78	0.87	0.71	1.49	1.33	0.95
CO3	**5**	1.19	1.92	4.57	1.17	0.34	0.35
CO4	**1**	0.35	2.29	0.83	1.11	0.96	0.55
CO7	**1**	1.01	3.44	0.66	0.69	4.60	0.80
CP11A	**1**	2.63	0.61	1.61	1.40	1.12	2.70
CRP	**2**	5.76	1.33	1.63	2.97	18.58	0.34
FBXL2	**1**	0.32	0.05	0.64	1.61	2.20	1.16
FETUA	**2**	0.36	1.82	0.42	1.06	0.61	0.83
FETUB	**1**	0.71	1.06	0.65	1.18	0.48	1.48
FGL2	**1**	0.69	0.70	0.58	0.90	0.75	1.19
GELS	**5**	3.96	0.43	3.17	1.06	9.53	0.33
GRLF1	**1**	1.17	0.65	0.76	1.45	1.16	1.19
HA1A	**1**	0.97	0.76	1.01	1.33	1.12	1.35
HBA	**1**	17.22	0.32	1.84	0.80	0.99	0.89
HBB	**1**	3.97	0.75	1.04	1.05	0.79	3.18
HEMO	**5**	0.57	1.37	0.74	1.30	0.51	1.49
HRG	**1**	1.08	1.24	0.94	1.34	0.96	0.94
IFNG	**1**	0.69	0.70	0.58	0.90	0.75	1.19
ITIH1	**4**	3.24	0.21	1.42	1.26	0.74	1.06
ITIH2	**3**	0.95	1.68	1.23	1.32	1.52	0.56
ITIH4	**6**	0.97	0.90	0.84	1.28	0.57	1.10
KCRM	**1**	1.11	0.77	0.82	0.87	0.76	0.94
KNG1	**1**	0.57	0.54	1.68	3.13	1.50	0.90
RET4	**1**	1.13	1.17	1.78	1.30	1.00	2.50
RPAP1	**1**	0.85	1.25	0.97	1.54	0.95	1.31
SMC3	**1**	0.85	1.16	0.94	1.77	0.75	1.30
SORL	**1**	0.96	0.96	1.16	1.09	1.10	1.06
SPA31	**4**	0.81	0.65	0.65	0.76	0.65	0.77
THBG	**1**	0.95	0.70	0.86	1.41	0.69	1.34
THRB	**2**	1.02	0.85	1.03	1.08	1.02	1.04
TRFE	**8**	0.92	1.52	1.92	1.16	1.33	0.63
TRYP	**3**	1.16	0.92	1.36	0.94	0.83	1.37
TTHY	**3**	4.67	0.87	1.59	1.56	0.93	2.21
VTDB	**2**	0.54	1.33	0.71	1.10	0.63	1.28
VTNC	**2**	0.87	1.68	0.93	1.09	0.84	1.61

*All values have been normalized against the matching 0 mpi timepoint

Due to the low number of proteins identified, it is unknown whether or not this conclusion can be made for samples of higher complexity, since the limited number of proteins identified may not include those occurring at the cut boundaries. It is possible that proteins near the cut boundaries were present in the samples, but could not be identified and their relative expression ratios observed. Also, relatively few fractions were taken, limited by the distribution of proteins within the sample.

Due to the incomplete dataset, it was not possible to identify biomarkers of chronic wasting disease in elk. It is anticipated that finding such markers in plasma would be difficult given the current state of the elk genomic information and the challenges involved in depleting high abundance proteins from elk plasma. Without considerably more effort in addressing these primary issues, other better defined study systems, such as cows with bovine spongiform encephalopathy[25] or cervidized mice expressing elk PrP^{C},[26, 27] may be feasible in the near term. A recent report used urine from BSE infected cattle to determine a set of five protein-based biomarkers of BSE disease progression.[25] Given that the bovine genome has been sequenced, identification of spectra should be relatively efficient. With potential increases in sensitivity from using a solution based 2D-LC platform rather than a 2D gel-based platform, additional and more complete proteome coverage may be possible.

6.5 Literature Cited

(1) Mitra, D.; Amaratunga, C.; Sutherns, R.; Pletsch, V.; Corneil, W.; Crowe, S.; Krewski, D. *J. Toxicol. Env. Heal. A* **2009**, *72*, 1106-1112.

(2) Prusiner, S. B. *Science* **1982**, *216*, 136-144.

(3) Brown, D. R.; Qin, K.; Herms, J. W.; Madlung, A.; Manson, J.; Strome, R.; Fraser, P. E.; Kruck, T.; von Bohlen, A.; Schulz-Schaeffer, W.; Giese, A.; Westaway, D.; Kretzschmar, H. *Nature* **1997**, *390*, 684-687.

(4) Viles, J. H.; Cohen, F. E.; Prusiner, S. B.; Goodin, D. B.; Wright, P. E.; Dyson, H. J. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 2042-2047.

(5) Watts, J. C.; Balachandran, A.; Westaway, D. *PLoS Pathogens* **2006**, *2*, 152-163.

(6) Tamgueney, G.; Miller, M. W.; Wolfe, L. L.; Sirochman, T. M.; Glidden, D. V.; Palmer, C.; Lemus, A.; DeArmond, S. J.; Prusiner, S. B. *Nature* **2009**, *466*, 652.

(7) Gonzalez-Romero, D.; Barria, M. A.; Leon, P.; Morales, R.; Soto, C. *FEBS Lett.* **2008**, *582*, 3161-3166.

(8) Rigou, P.; Rezaei, H.; Grosclaude, J.; Staunton, S.; Quiquampoix, H. *Environ. Sci. Technol.* **2006**, *40*, 1497-1503.

(9) Johnson, C. J.; Pedersen, J. A.; Chappell, R. J.; McKenzie, D.; Aiken, J. M. *PLoS Pathogens* **2007**, *3*, 874-881.

(10) Johnson, C. J.; Phillips, K. E.; Schramm, P. T.; McKenzie, D.; Aiken, J. M.; Pedersen, J. A. *PLoS Pathogens* **2006**, *2*, 296-302.

(11) Anderson, N. L.; Anderson, N. G. *Mol. Cell. Proteomics* **2002**, *1*, 845-867.

(12) Wang, N.; Xie, C.; Young, J. B.; Li, L. *Anal. Chem.* **2009**, *81*, 1049-1060.

(13) Bjorhall, K.; Miliotis, T.; Davidsson, P. *Proteomics* **2005**, *5*, 307-317.

(14) Cordwell, S. J.; Thingholm, T. E. *Proteomics* **2010**, *10*, 611-627.

(15) Granger, J.; Siddiqui, J.; Copeland, S.; Remick, D. *Proteomics* **2005**, *5*, 4713-4718.

(16) Silveira, J. R.; Raymond, G. J.; Hughson, A. G.; Race, R. E.; Sim, V. L.; Hayes, S. F.; Caughey, B. *Nature* **2005**, *437*, 257-261.

(17) Zhang, G.; Fenyo, D.; Neubert, T. A. *J. Proteome Res.* **2009**, *8*, 1285-1292.

(18) Brenn, A.; Karger, A.; Skiba, M.; Ziegler, U.; Groschup, M. H. *Proteomics* **2009**, *9*, 5199-5205.

(19) Hsu, J.-L.; Huang, S.-Y.; Chow, N.-H.; Chen, S.-H. *Anal. Chem.* **2003**, *75*, 6843-6852.

(20) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917-1926.

(21) Grossmann, J.; Fischer, B.; Baerenfaller, K.; Owiti, J.; Buhmann, J. M.; Gruissem, W.; Baginsky, S. *Proteomics* **2007**, *7*, 4245-4254.

(22) Waridel, P.; Frank, A.; Thomas, H.; Surendranath, V.; Sunyaev, S.; Pevzner, P.; Shevchenko, A. *Proteomics* **2007**, *7*, 2318-2329.

(23) Boersema, P. J.; Raijmakers, R.; Lemeer, S.; Mohammed, S.; Heck, A. J. R. *Nat. Protoc.* **2009**, *4*, 484-494.

(24) Yerbury, J. J.; Kumita, J. R.; Meehan, S.; Dobson, C. M.; Wilson, M. R. *J. Biol. Chem.* **2009**, *284*, 4246-4254.

(25) Simon, S. L. R.; Lamoureux, L.; Plews, M.; Stobart, M.; LeMaistre, J.; Ziegler, U.; Graham, C.; Czub, S.; Groschup, M.; Knox, J. D. Proteome Sci. 2008, 6, doi:10.1186/1477-5956-6-23.

(26) Seelig, D. M.; Mason, G. L.; Telling, G. C.; Hoover, E. A. *Am. J. Pathol.* **2010**, *176*, 2785-2797.

(27) LaFauci, G.; Carp, R. I.; Meeker, H. C.; Ye, X.; Kim, J. I.; Natelli, M.; Cedeno, M.; Petersen, R. B.; Kascsak, R.; Rubenstein, R. *J. Gen. Virol.* **2006**, *87*, 3773-3780.

Chapter 7 – Conclusion and Future Work

This thesis describes the characterization and applications of a differential labelling chemistry technology for MS-based quantitative proteomics experiments. The dimethylation after guanidinylation (2MEGA) labelling method was successfully used for quantitative comparison of samples from a human colon cancer cell line and *E. coli* grown under differential conditions and is generally applicable to all types of protein samples. Certain features of the labelling method may find currency for high throughput applications, particularly the consistent observation of the a_1 fragment ion for N-terminal residue identification, the low reagent cost of the isotopically labelled formaldehyde used, and facile method automation for reduced analyst intervention. The potential for sample preparation in a high-throughput manner has been suggested and is a readily achievable goal.

From a technical perspective, isotopic label switching (reciprocal labelling) was used to confirm the consistency obtained from MS-based quantitative experiments. It was found that the quantification ratios reported for most peptide matches were qualitatively correct and that over 95% of common peptides were within 67% of the average value between two runs. The consistency observed for ratios when using forward and reverse labelling at protein level was significantly improved when discarding a small percentage (~5%) of inconsistent peptides. Furthermore, reverse labelling was found to be useful for detection of peptides with large, but incorrect, ratios due to peptide misidentifications or incorrect peak fitting used for area calculations.

A definitive conclusion regarding the applicability of protein level separation methods with peptide-level labelling could not be established. The low numbers of peptides and proteins identified from a plasma sample from elk was limited by the unavailability of a genomic or sequence database. SDS-PAGE separation followed by peptide level labelling gave qualitatively accurate results for most peptides; however, quantitative measurements on closeness of fit could not be determined. The use of a mixed reference standard was also explored as an alternative to using a single time

point as the control/reference sample. The construction of a mixed standard creates an acceptable range for abundance ratios and values outside of this range can be easily identified as likely to be incorrect.

More generally, the applicability of a particular labelling method will ultimately depend on specific requirements of any analytical challenge. The 2MEGA labelling method seems particularly well suited for applications for which a) low method cost is a premium, b) reasonable quantification accuracy is required, c) instrument analysis time is not a major consideration, and d) a standard or reference is available. For applications in which one of these conditions may not hold true, there are various methods that can be considered and implemented with equal or greater overall utility.

7.1 Future Work

7.1.1. Revisiting the LC-MALDI Platform

In Chapter 3, the LC-MALDI platform was used to selectively identify peptides from peak pairs with implied differential expression. One of the driving forces for peak selection, rather than total identification of all peaks, was the comparatively slow MS/MS acquisition speed of MALDI instrumentation used at the time of the work. Current MALDI TOF/TOF instrumentation can acquire considerably more spectra per unit time and vendor supplied software packages can perform automated data analysis to enable exclusive selection of differentially expressed pairs for MS/MS sequencing. Both of these factors together should greatly reduce analysis time.

Interestingly, the refinement of MALDI- and ESI-based platforms may ultimately lead to both ionization methods serving different research areas. The development of robust nanoESI interfaces and capillaries have improved detection sensitivity for low sample loading amounts due to the increased ionization efficiency of analytes in the flow regime of ~100 nL/minute.[1] While beneficial for sub-microgram quantities of samples, a lower flow rate necessitates the use of smaller columns for optimal chromatography. Conversely, the decoupling of chromatographic separation and MS analysis means MALDI-based platforms will be competitive for comparatively

high loadings, when higher LC flow rates (2 - 40 µL/min) can be used. Provided that the chromatography is comparable, higher loadings may lead to more peptide identifications, as samples can be continually analyzed until the sample spots are exhausted.

For study systems with abundant sample amounts, it would be worthwhile to analyze samples by both ESI and MALDI to exploit the complementary ionization behaviour to increase proteome coverage.[2, 3] Reports from our laboratory and other sources suggest there is only 25% peptide overlap and 60% protein overlap by MALDI and ESI.[4] The 2MEGA method is relatively low cost when significant amounts of sample need to be prepared, so the limiting factor would be instrument analysis time on both platforms.

7.1.2 High Throughput Studies and Clinical Applications

While mass spectrometry has been used to elucidate protein changes from a variety of conditions in a diverse range of biologically relevant study systems, this has not necessarily translated into clinically relevant assays. Although many reported studies discuss biomarker identification, protein-based biomarker candidates are rarely validated in any significant fashion.[5] Comprehensive longitudinal studies using mass spectrometry across large sample pools at different time points are uncommon in the literature. Current technologies are suited for the biomarker discovery phase (generally <10 samples) and MRM-based analyses provide the added sensitivity and specificity for precise biomarker quantification for large populations (>1000 samples). However, the verification stage, requiring hundreds of samples, sits as the interface between these two phases and presents some unique challenges.

Although it is unlikely that MS-based protein assays will gain widespread adoption in the short term, development of such technology may allow for critical evaluation of the merits and drawbacks of an MS-based analytical platform for peptide biomarkers. The majority of clinical assays currently performed use selective electrodes for inorganic species, spectrophotometric assays on 96-well plates, or are antibody-based immunoassays. These analytical methods offer speed considerations

or are easily multiplexed; these are critical considerations for clinical applications. Clinically relevant biomarker development is a rigorous and time consuming process and takes on the order of several years to bring to bear. While most MS-based assays require LC separation prior to analysis, one potential advantage is that simultaneous analysis of several candidate protein markers can be measured over the course of a single run. By expanding the number of potential candidates, biomarker panels can be developed, which may have more predictive accuracy than single protein markers. This is particularly advantageous during the verification phase, where a variety of biomarker candidates may have been indicated during the discovery phase. While protein-based MS strategies are currently relegated to a discovery role in clinical applications, increasing speed and sensitivity of MS instrumentation may make it a competitive analytical platform for addressing future clinical challenges.

7.2 Literature Cited

(1) Juraschek, R.; Dulcks, T.; Karas, M. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 300-308.

(2) Bodnar, W. M.; Blackburn, R. K.; Krise, J. M.; Moseley, M. A. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 971-979.

(3) Shirran, S. L.; Botting, C. H. *J. Proteomics*, *73*, 1391-1403.

(4) Young, J. B.; Li, L. *Anal. Chem.* **2007**, *79*, 5927-5934.

(5) Siu, K. W. M.; De Souza, L. V.; Scorilas, A.; Romaschin, A. D.; Honey, R. J.; Stewart, R.; Pace, K.; Youssef, Y.; Chow, T.-f. F.; Yousef, G. M. *J. Proteome Res.* **2009**, *8*, 3797-3807.

Appendix 1 - 2MEGA Labelling Protocol

This protocol details the automated dimethylation after guanidinylation (2MEGA) protocol. The reagent amounts required for a particular liquid handler configuration will depend on the minimum volume required for the liquid handler, the working range of accurate solution dispensation for the liquid handler, and the number of samples to be prepared. The solutions can be scaled, as appropriate. Note that the prepared reagent concentrations stated are only approximate values.

For small samples (5 to 25 µg of peptides), the low concentration solutions (3M HCl, 0.5% CH_2O, and 0.5 M 2-picoline borane) are suggested to maintain a minimum dispensing volume of 5 µL. For larger samples (100 – 200 µg of peptides), high concentration solutions (6M HCl, 8% CH_2O, and 2M 2-picoline borane) are suggested to keep reagent volumes reasonable (<200 µL). If desired, larger sample solutions can be aliquoted into several tubes. Please consult Section 3) Liquid Handler Protocol for additional information.

A1.1 Reagents Required

Sodium Hydroxide
Sodium Carbonate (monohydrate)
O-methylisourea hemisulfate (99%)
Hydrochloric Acid (37%)
Sodium acetate (trihydrate)
Acetic acid (glacial)
Formaldehyde (37%, w/v)
2-Picoline borane
Ammonium bicarbonate
LC-MS grade water
LC-MS grade methanol

A1.2 Reagent Preparation (Not all solutions will be required; consult Section 3 for required solutions)

a) 2M NaOH

Dissolve 1.60g of NaOH in 20 mL LC-MS grade water.

b) 1M $NaCO_3$ (solution should be prepared weekly)

Dissolve 2.48g of $NaCO_3 \bullet H_2O$ in 20 mL LC-MS grade water.

c) *O*-methylisourea hemisulfate solution (~3M, must be freshly prepared)

Dissolve 0.3683g *O*-methylisourea hemisulfate in 261.2 µL 2M NaOH and 261.2 µL 1M Na_2CO_3. The solution pH should be 11.5, by pH paper. The pH can be further adjusted by addition of 2M NaOH, if required (< 20 µL).

d) 3M HCl

Deliver 15 mL room temperature LC-MS grade water into a clean container and cool the water in an ice bath. Slowly add 5 mL HCl (37% w/v) and allow the solution to come to room temperature before usage.

e) 6M HCl

Deliver 10 mL room temperature LC-MS grade water into a clean container and cool the water in an ice bath. Slowly add 10 mL HCl (37% w/v) and allow the solution to come to room temperature before usage.

f) 1M Acetate Buffer (~pH 5)

Dissolve 20.3 mg of sodium acetate trihydrate in 952 µL LC-MS grade water. After the salt is fully dissolved, add 48 µL of glacial acetic acid. The pH should be tested with pH paper.

g) 4% (w/v) Formaldehyde (must be freshly prepared)

Add 48 µL of 37% (w/v) formaldehyde to 396 µL LC-MS grade water.

h) 0.5% (w/v) Formaldehyde (must be freshly prepared)

Add 6 µL of 37% (w/v) formaldehyde to 438 µL LC-MS grade water.

i) 8% (w/v) Formaldehyde (must be freshly prepared)

Add 80 µL of 37% (w/v) formaldehyde to 290 µL LC-MS grade water.

j) 2M 2-Picoline borane (must be freshly prepared)

Dissolve 184.4 mg of 2-picoline borane in 861.4 µL LC-MS grade methanol.

k) 0.5M 2-Picoline borane (must be freshly prepared)

Dissolve 92.2 mg of 2-picoline borane in 861.4 µL LC-MS grade methanol.

l) 1M Ammonium bicarbonate (must be freshly prepared)

Dissolve 79.06 mg of ammonium bicarbonate in 1 mL LC-MS grade water.

A1.3 Liquid Handler Protocol

For initial experiments, it is suggested a standard protein digest mixture be used to check labelling performance and adjustments made, as required. The test sample should be 100 µL of 0.5 µg/µL of protein digest in 100mM NH_4HCO_3.

Ideally, reagents amounts should be scaled for the peptide amount in the sample for optimal performance. Concentrated samples should be diluted with 100 mM NH_4HCO_3 to a final concentration of 0.5 µg/µL. Three different ranges are suggested for optimal labelling (see Table A1.1 below). Reagent amounts may need to be adjusted slightly for particular applications.

Table A1.1 Reagents for 2MEGA Labelling of Various Peptide Amounts

Peptide Amount	Sample Volume	2M NaOH	OMIS Soln	3M HCl	Acetate Buffer	0.5% CH2O	0.5M Picoline Borane	1M NH4HCO3
(in µg)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)
5	100	12	5	9.6	24	8	8	16
10	100	12	10	11.2	24	16	16	16
15	100	12	15	12.8	24	24	24	16
20	100	12	20	14.4	24	32	32	16
25	100	12	25	16.0	24	40	40	16

Peptide Amount	Sample Volume	2M NaOH	OMIS Soln	3M HCl	Acetate Buffer	4% CH2O	2M Picoline Borane	1M NH4HCO3
(in µg)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)
25	100	12	25	16	24	5	10	16
50	100	12	50	24	24	10	15	16
75	100	12	75	32	24	15	20	16

Peptide Amount	Sample Volume	2M NaOH	OMIS Soln	6M HCl	Acetate Buffer	8% CH2O	2M Picoline Borane	1M NH4HCO3
(in µg)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)	(in µL)
50	100	12	50	12	24	5	20	16
75	150	18	75	16	24	7.5	30	16
100	200	24	100	20	24	10	40	16
150	300	36	150	28	24	15	60	16
200	400	48	200	36	24	20	80	16

The reaction protocol can be summarized into the following steps:

Samples are heated to 37 °C and temperature is kept constant throughout

Reagent solutions are placed into the reagent racks and sequence is started.

Addition of NaOH to all samples

Addition of *O*-methylisourea solution to all samples

Incubation for 75 minutes

Addition of HCl to all samples to stop guanidinylation reaction

Addition of acetate buffer to all samples to prepare samples for dimethylation

Addition of formaldehyde to all samples

Addition of 2-picoline borane to all samples

Incubation for 30 minutes

Addition of ammonium bicarbonate

Incubation for 15 minutes

Downstream preparation

It is suggested that samples are desalted before LC-MS analysis (see Wang, N.; Xie, C.; Young, J. B.; Li, L. *Anal. Chem.* 2009, *81*, 1049-1060). This method can also be used to quantify samples to determine how samples should be mixed for quantitative experiments.

A1.4 Troubleshooting

a) Guanidinylation (either incomplete at lysine or excess N-terminal guanidinylation)

The guanidinylation step is primarily controlled through the solution pH after adjustment with 2M NaOH and the addition of the *O*-methylisourea solution. If incomplete guanidinylation at lysine is observed (defined as either unmodified lysines or dimethylated lysines for >2% of all identifications from the MS/MS search result), the solution pH during guanidinylation should be checked. The pH should be 11.5 during guanidinylation and should be adjusted by adding additional 2M NaOH to the *O*-methylisourea solution. If required, the *O*-methylisourea solution can be adjusted to pH 12 by dissolving *O*-methylisourea hemisulfate salt 3:1 solution (v/v) of 2M NaOH/1M Na_2CO_3 instead of a 1:1 solution (v/v).

Observation of N-terminal guanidinylation at glycine or alanine is anticipated and should account for >60% of all N-terminal guanidinylation instances. If guanidinylation at the N-terminus of peptide is observed in excess for non-glycine or non-alanine peptides (defined as non-G/non-A N-terminal guanidinylation for >2% of all identifications from the MS/MS search result), the initial incubation time can be reduced to 60 minutes from 75 minutes.

b) Missing Dimethylation

Missing dimethylation is a comparatively rare occurrence, given the excess of reagent used in this protocol. If missing dimethylation is observed (defined as >2% unmodified N-termini of all identifications from the MS/MS search result), it is suggested that the pH of the solution is checked after addition of acetate buffer (pH ~5-7) and the formaldehyde and 2-picoline borane solutions be prepared fresh and re-tested. If the issue persists, fresh reagents should be purchased.

c) Excessive Precipitation

While some precipitation is expected after completion of the reaction, it can be reduced by scaling down the amount of 2-picoline borane used by half while keeping

the formaldehyde concentration constant. However, the reaction conditions should be carefully tested to ensure that the dimethylation remains complete.