Assessment of Universal Approaches to Proteome Prefractionation

by

Fang Liu

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
May 2011

DALHOUSIE UNIVERSITY

DEPARTMENT OF CHEMISTRY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Assessment of Universal Approaches to Proteome Prefractionation" by Fang Liu in partial fulfilment of the requirements for the degree of Master of Science.

Dated: May 13th, 2011

Supervisor: ______________________________

Readers: ______________________________

____

____

DALHOUSIE UNIVERSITY

DATE: May 13th, 2011

AUTHOR: Fang Liu

TITLE: Assessment of Universal Approaches to Proteome Prefractionation

DEPARTMENT OR SCHOOL: Department of Chemistry

DEGREE: MSc CONVOCATION: October YEAR: 2011

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions. I understand that my thesis will be electronically available to the public.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

The author attests that permission has been obtained for the use of any copyrighted material appearing in the thesis (other than the brief excerpts requiring only proper acknowledgement in scholarly writing), and that all such use is clearly acknowledged.

____
Signature of Author

Table of Contents

List of Tables

List of Figures

Abstract

Protein prefractionation is a popular and effective strategy for improved MS analysis of complex proteome mixtures. A challenge of prefractionation is the even partitioning with high recovery of all components of the mixture, particularly hydrophobic proteins. This thesis assesses various proteome prefractionation platforms, with a goal of comprehensive proteome analysis. A more reliable dataset of 1136 *S. cerevisiae* transmembrane proteins was computationally generated, and used to assess two gel-based platforms (GeLC/MS and GELFrEE/MS). These platforms were determined to be comparable for proteome analysis. The requirement for high-throughput, automated fractionation demands a gel-free separation workflow. Here, a LC-based workflow was optimized, relying on SDS-assisted yeast extraction, organic solvent protein precipitation, and reversed phase separation in a formic acid/isopropanol solvent system. Though this workflow afforded improvements over conventional LC strategies to proteome fractionation, the gel-based platforms were demonstrated to be superior, in terms of their unbiased separation of hydrophobic *vs* hydrophilic proteins.

List of Abbreviations and Symbols Used

1D	one dimensional
2D	two dimensional
2DGE	two dimensional gel electrophoresis
AC	alternating current
ACN	acetonitrile
APFO	ammonium perfluorooctanoate
BSA	bovine serum albumin
CE	capillary electrophoresis
CF	chromatofocusing
CID	collision induced dissociation
CMW	chloroform/methanol/water
DC	direct current
DDA	data dependent acquisition
DTT	dithiothreitol
ESI	electrospray ionization
ExPASy	Expert Proteomics Analysis System
FA	formic acid
FFE	free-flow electrophoresis
GELFrEE	gel-eluted liquid fraction entrapment electrophoresis
GRAVY	grand average of hydropathy
HIC	hydrophobic interaction chromatography
HILIC	hydrophilic interaction chromatography
HMM	hidden Markov model
HPLC	high-performance liquid chromatography
ID	inner diameter
IEX	ion exchange chromatography
IEF	isoelectric focusing
IMPs	integral membrane proteins
IPA	isopropanol
LC	liquid chromatography
LIT	linear ion trap
MALDI	matrix-assisted laser desorption/ionization
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MudPIT	multidimensional protein identification technology
MW	molecular weight
m/z	mass to charge ratio
NCBI	National Center for Biotechnology Information
NMR	nuclear magnetic resonance
NSI	nano-electrospray ionization
pI	isoelectric point
PTM	post-translational modification
PSDAVB	poly styrene-divinyl benzene

QIT	quadrupole ion trap
RF	radio frequency
RP	reversed phase
RPLC	reversed phase liquid chromatography
RPµLC	reversed phase microcapillary liquid chromatography
rpm	rotations per minute
Rsp	ranked preliminary score
SCX	strong cation exchange
SDS	sodium dodecyl sulfate
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
sIEF	solution isoelectric focusing
S/N	signal to noise
SEC	size exclusion chromatography
SPE	solid phase extraction
STD	standard deviation
TFA	trifluoroacetic acid
TMH	transmembrane alpha-helix
TMHMM	transmembrane hidden Markov model
TOF	time-of-flight
TPCK	tosyl phenylalanyl chloromethyl ketone
Tris	tris(hydroxymethyl)aminomethane
UV	ultraviolet
UniProt	Universal Protein Resource
UniProtKB	UniProt Knowledgebase
Xcorr	cross-correlation value
YPD	yeast extract-peptone-dextrose
Å	Angstrom
amu	atomic mass unit
χ^2	chi-square
Da	Dalton
ΔCN	Delta correlation value
ºC	degrees Celcius
fmol	femtomole
g	gravity
L	litre
kDa	Kilodalton
µg	microgram
µL	microlitre
nm	nanometer
min	minutes
mL	millilitre
mm	millimetre
M	molarity
%C	percent cross-linker (*w*/*w*)
%T	percent total acrylamide (*w*/*v*)
psi	pounds per square inch

Acknowledgements

I would first like to thank my supervisor, Dr. Alan Doucette, for his encouragement and helpful criticisms over the course of my time at Dalhousie. I really appreciated the considerable amount of effort he spent guiding me through research. Learning has become a wonderful experience.

I would like to thank past and present members of Dr. Doucette's research group, namely Dr. John Tran, Dr. Mark Wall, Dr. Diane Botelho, Dennis Orton, Douglas Vieira, Shayla Fitzsimmons, Dr. Aaron Steevensz, and Jason Smith. I especially would like to thank Dr. John Tran, Dr. Mark Wall, Dr. Diane Botelho and Dennis Orton for their great help in many aspects of my research projects. I cannot express how grateful I am to work with and learn from them.

I would like to thank my committee members, Dr. Peter D. Wentzell and Dr. Robert D. Guy, for their advice over the years. Also, I thank Dr. Peng Zhange for taking the time to read this thesis as well as the adives I learned from them towards my research projects.

I also like to thank people who provide a great support to my research work. Thanks to Ken Chisholm of the NRC-IMB and Elden Rowland of the Proteomics and Mass Spectrometry Center for their valuable help and time. Thanks to Simon Smith (Dr. Lois Murray's former student), Marissa LeBlanc (Dr. Christopher McMaster's former student), and Dr. Melanie Dobson, of Dalhousie University, for providing yeast cells. Thanks to Kelly L. Resmer and Dr. Robert L. White, of Chemistry Department, for helping to culture yeast cells. For some work in the lab, Machinists Mike Boutilier and Rick Conrad, of the Chemistry Department, provided great help. Also, I would like to thank the office staff of the Department of Chemistry, particularly Giselle Andrews, for their incredible patience and help over the years. Thanks to all of the lab instructors, especially Cathy Byers, during my time as a teaching assistant.

Finally, I would like to thank my entire family for all of their support throughout my degree. To my parents, thank you for encouraging me to pursue higher education. To my husband and sons, thank you for your support and understanding, while I went back university to pursue this higher education. I really thank you all and love you very much.

Chapter 1

Introduction

With many completed genome projects, including the *Saccharomyces cerevisiae* genome [1], proteomics research was accelerated as a more direct, information-rich approach to rapid and robust characterization of gene products (i.e., proteins). Concurrent with genome sequencing, the rapid development of mass spectrometry (MS)-based technologies for protein characterization is a driving force to help expand the potential of proteomics studies. MS techniques permit identification of proteins and peptides in a high-throughput, comprehensive fashion. Methodological and technological advances have allowed MS analysis of proteins to be brought to the forefront of biological research; however, complete proteome analysis remains a formidable task. The analysis of integral membrane proteins (IMPs), in particular transmembrane helices (TMHs) in IMPs or transmembrane proteins, is considered one of the more challenging aspects of proteomics. The physico-chemical properties of transmembrane proteins, such as hydrophobicity, make conventional protein analysis techniques less effective [2,3], hence making these proteins under-represented in conventional proteomic analyses [4,5]. Given the complexity of biological samples, an unbiased proteome analysis platform needs to consider proper sample preparation, separation and identification to permit representative profiling of the proteome.

In this work, a solution-based proteome characterization strategy is presented. This established approach identifies IMPs concurrently with the more soluble, hydrophilic components of the sample. The strategy relies on an optimized extraction, solubilization and proteome prefractionation ahead of mass spectrometry. Considering the need to identify and characterize transmembrane proteins, the first step of this work is to find an appropriate computational TMH prediction method for obtaining a reliable dataset of yeast transmembrane proteins. This provides a basic reference point for all further experimental comparisons. The computational approach is used to assess the performance of gel as well as solution-based proteome prefractionation strategies, together with compatible methods for proteome extraction, precipitation, solubilization, and identification with tandem mass spectrometry (MS/MS).

1.1 Proteomics

The term 'proteomics' refers to the large-scale study of the 'proteome' which represents both the total number of proteins expressed by a cell or organism, and all protein isoforms and post-translational modifications (PTMs) [6]. Proteomics has thus allowed key insights into the composition, regulation and function of biological systems, allowing a greater fundamental understanding of the role specific proteins play in subcellular processes [6]. In contrast to a genome, a proteome may differ from cell to cell and is constantly changing throughout the lifespan of the organism in response to environmental conditions. In addition, proteins are subject to various chemical modifications or PTMs such as splicing and translation isoforms and differential modification and processing species, which alter the physical and chemical properties of the molecule. Neurexins (a family of highly polymorphic neuronal cell surface proteins), for example, can experience greater than 1,000 PTMs [7]. In fact, some 50-90% of the proteins in the human body are estimated to be post-translationally modified [8].

Modifications tailor the function of the protein, as seen through phosphorylation, in which a negative charge is added to the protein altering its conformation to act as a molecular switch to control diverse cellular processes [8]. In addition to PTMs, the proteome has a vast dynamic range. Using the human plasma proteome as an example, the dynamic range is up to 10 orders of magnitude (serum albumin, 35-50×10^9 pg/ml vs Interleukin 6 at 0-5 pg/ml) [9]. Less complex organisms, such as yeast have a lower dynamic range; however, it is still estimated to be 5 to 6 orders of magnitude [10]. Therefore, because of its enormous diversity, a more versatile analytical platform is required to conduct an effective characterization of the proteome.

The systematic analysis of all proteins in a complex biological sample has been stimulated by MS-based proteomics [11]. Mass spectrometry has been the analytical chemist's workhorse for protein research due to its high sensitivity, resolution, throughput, and ease of protein identification. Surely the most significant advances related to MS of biological molecules have been the introduction of two soft ionization techniques: electrospray ionization (ESI) [12] and matrix-assisted laser desorption/ionization (MALDI) [13], which were developed in the late 1980s. Coupled with these ionization techniques, the development of mass analyzers and multistage instruments (for example, the linear ion trap (LIT), Orbitrap [14] and LIT-Orbitrap [15]) provides high mass accuracy, rapid scanning, and enormous instrumental versatility. These developments have led to improvements in protein identification, making MS an indispensable tool for proteomic research.

Because of the high sample complexity and variability, whole proteome analyses require protein separation prior to MS for high resolution proteomic identification.

Proteins from complex biological sources can be partially resolved by either gel-based (i.e., electrophoresis) or solution-based separation such as, liquid chromatography (LC) [11]. In contrast to gel-based separation, solution-phase separation is conventionally applied at the peptide level, requiring proteolytic digestion of proteins prior to separation. Solution-based separation of intact proteins (i.e., proteome prefractionation) is also possible, but less commonly practiced. These approaches are discussed in Section 1.2.3. The vast majority of current proteomic strategies employ peptide separation and identification by MS/MS for analysis of complex protein samples.

1.2 Methods Used in Proteomics

The rapid development of proteomics research has been due in large part to the coupling of two fundamental analytical platforms [11,16]: MS and LC. As a sensitive and selective detector, MS not only affords an ability to 'read' amino acid sequences of proteins/peptides based on MS fragmentation patterns (see Section 1.2.2.2), but it also enables protein characterization with a high degree of sample throughput. Analysis speed is also critical, and LC/MS offers a convenient approach to profile the incredible number of components present in the complex mixture, over an acceptable time frame.

1.2.1 Mass Spectrometry

The MS instrument consists of three components: an ionization source, a mass analyzer and an ion detector. The source permits samples to be introduced to the instrument by converting molecules (typically present in solution) into gas-phase ions. These gaseous ions can be separated (in space or in time) through the mass analyzer, according to the respective mass to charge ratio (*m/z*) of each individual ion. The ion detector permits measurement of the number of ions of a given *m/z*, which is ultimately presented as a mass spectrum.

There are several types of MS instruments, differing in the source and mass analyzer (the ion detector is chosen according to the analyzer). Various instruments tailor the detection system to the class of molecule (large *vs* small molecules, solid *vs* liquid *vs* gas phase samples or polar *vs* non polar molecules) being investigated. Proteomic studies generally employ ESI as a source, coupled to linear ion trap mass analyzers [17,18]. Currently, nano-electrospray ionization (NSI) [19], similar to ESI, is a more sensitive and efficient method for LC-MS experiments due to a lower flow rate (200–800 nL/min comparing to 1-100 µL/min for ESI) and a smaller inner diameter (ID) spray tip (2-10 µm *vs* 75 µm). NSI has become the most popular form of ESI for proteome analysis [20]. Given that the ThermoFisher 'LTQ'® linear ion trap mass spectrometer NSI, was exclusively used throughout this research, a focused description of this instrument, in the context of protein identification, is presented below.

One of the main merits of ESI or NSI for proteomic research is the generation of multiply charged ions. With the analysis of large protein molecules, which can be considered as polyprotic acids, generation of multiple charged ions is typical of the electrospray process. An example of an NSI mass spectrum for horse heart myoglobin is shown in Figure 1.1A. The number of charges seen in the mass spectrum of a protein depends on the number of basic amino acid residues, the solution pH, as well as the folding state of the protein. Basic sites of proteins include the side chains of lysine, arginine, histidine, and the N-terminus, whose structures are provided in Figure 1.1B. Protons are transferred to these sites from solvent during the ionization process. A precise

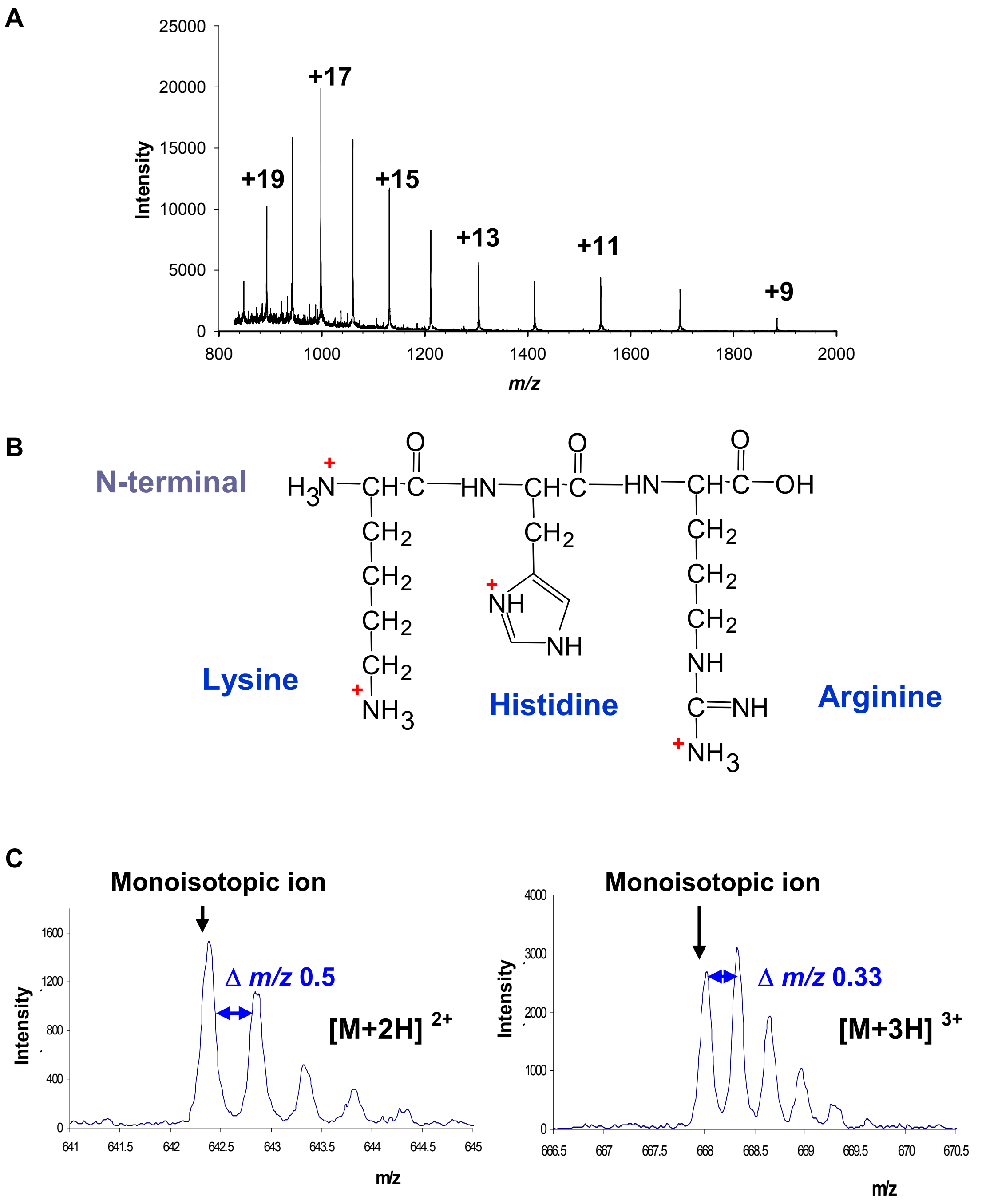


Figure 1.1 Mass spectra produced by ESI or NSI. (A) A mass spectrum of the intact protein Horse Heart Myoglobin, showing the charge envelope containing the protein with multiple charge states (up to +20). (B) The structures of the three basic amino acids lysine, histidine and arginine which can be protonated to give rise the positive charges during ESI or NSI (positive mode). (C) Mass spectra of a doubly and triply charged peptide. The mass difference of a doubly charged peptide between isotopic ions is 0.5 *m/z*, the mass difference of a triply charged peptide is ~0.33 *m/z*.

molecular mass of large proteins/peptides can be obtained through a deconvolution calculation. As shown in Fig 1.1C, peptides formed by trypsin digestion can be found with doubly or higher charge states. Because trypsin cleaves peptides at the basic amino acid residues lysine and arginine, these peptides are easily ionized and fragmented. Peptide fragment patterns are full of information which can be used for peptide identification and the important roles of these patterns will be explained in detail in Section 1.2.2.2. Due to the mechanism of detection (i.e., detection based on mass to charge ratio) generation of multiply charged ions through NSI or ESI have become favorable ionization methods for obtaining accurate molecular masses of large biological molecules (i.e., proteins/peptides) whose masses exceed that of the upper mass range of the mass analyzer.

Formation of molecular ions by ESI/NSI is followed by introduction into the mass spectrometer where a mass analyzer is used to separate them based upon their *m/z*. Thus, the mass analyzer is another critical part of the mass spectrometer. Commonly used mass analyzers in proteomics include linear quadrupoles, quadrupole ion traps (QIT), linear ion traps, quadrupole time-of-flight (TOF), Orbitrap and Fourier transform ion cyclotron resonance mass analyzers [17]. These instruments differ in their physical principles, their performance standards, and their ability to support specific analytical strategies. All work presented in this thesis is conducted using a LIT mass analyzer. The instrument provides medium mass accuracy, has tandem mass capabilities (ion fragmentation), a high scanning speed, and a reasonable cost [17]. Fragmentation ($MS^{/}MS$ or MS^2) are very useful for providing additional structural information, thus is commonly used for peptide and subsequent protein identification.

1.2.2 Tandem Mass Spectrometry

As early as the 1960s, MS was used for protein sequencing [21,22], mostly for identifying proteins with peptide mass fingerprinting [23,24]. Today, because of its ability to elucidate the amino acid sequence of peptides, MS/MS has become almost indispensable in high throughput MS-based protein identification studies [25,26]. In this thesis, proteins are identified by LC-MS/MS. Thus, the platform of MS/MS will be explained in detail in the following sections.

MS/MS was developed to derive structural detail, and originally was realized through the use of two mass analyzers between the ion source and the detector. Selected precursor ions are isolated or transmitted by the first mass analyzer, fragmented by collision induced dissociation (CID), and fragment ions are scanned and detected by the second mass analyzer. In a LIT, the first two steps occur in the ion trap. Detailed structural features of the peptides (i.e., amino acid sequences of peptides) can be deduced from the resulting fragments generated in an MS/MS experiment.

Fragment ions can be produced and detected by tandem MS in either space or time. Tandem MS “in space” refers to the configuration described above, referring to the selection of parent ion and separation of fragment ions being performed in two distinct regions of the mass spectrometer. Examples of modern MS instrumentation which perform tandem MS in space are the triple quadrupole and TOF-TOF instruments. Tandem MS “in time” refers to parent ion isolation, fragmentation, and detection in the same mass analyzer. The LIT is a good example of an instrument which employs this method of MS/MS.

1.2.2.1 Collision Induced Dissociation

When an ion collides with a neutral atom or molecule, a portion of the ion's kinetic energy is lost as its translational velocity decreases. The lost translational energy of the ion is converted to internal energy causing the ion to fragment. Collision materials are chosen to be inert gases such as helium, argon and xenon, as they, by definition, do not react with ions. Proteomics studies employing triple quadrupole and ion trap instruments use low energy CID which allows predictable fragmentation of peptides in the MS [27].

For protein identification, peptides are typically ionized through protonation in ESI or NSI and fragment ions generated by CID and analyzed by MS/MS. Peptide ions fragmented by CID predominantly produce 'b' and 'y' ions. A diagram depicting b and y ion nomenclature for peptides [28] is shown in Figure 1.2. The mechanism of b and y ion generation by CID is explained through the "mobile proton model" [29,30] and charge-site-directed fragmentation of amide bonds [31,32] as shown in Figure 1.3. Briefly, a single peptide is isolated from a peptide mixture trapped into a LIT and fragmented by following CID. The mass spectrum of a series of b and y ions generated by the same precursor peptide is used to deduce its peptide sequence. The interpretation of MS/MS spectra will be described in detail in next section.

1.2.2.2 Peptide Sequencing

Product ion spectra can be manually interpreted to determine the amino acid sequence of a peptide by recognizing the b and y ion series and calculating the residue masses of amino acids from adjacent ions in the same series. An example of the peptide

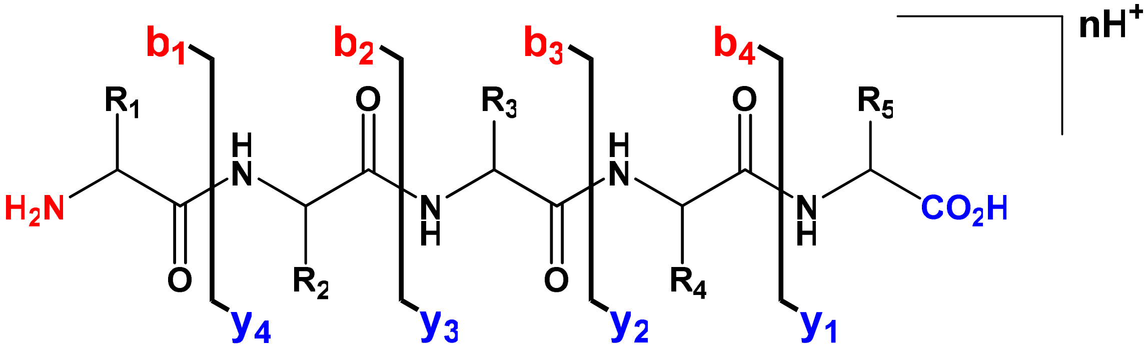


Figure 1.2 A schematic diagram of b and y ion nomenclature for a peptide. The common fragment ions for a dissociated peptide by CID are b ions that contain charge in the original N-terminus and y ions that contain charge in the original C-terminus. The numeral subscript represents the number of amino acid residues from either the N-terminus or C-terminus to the cleavage site. A series of b and y ions are generated by cleavage of different amide bonds.

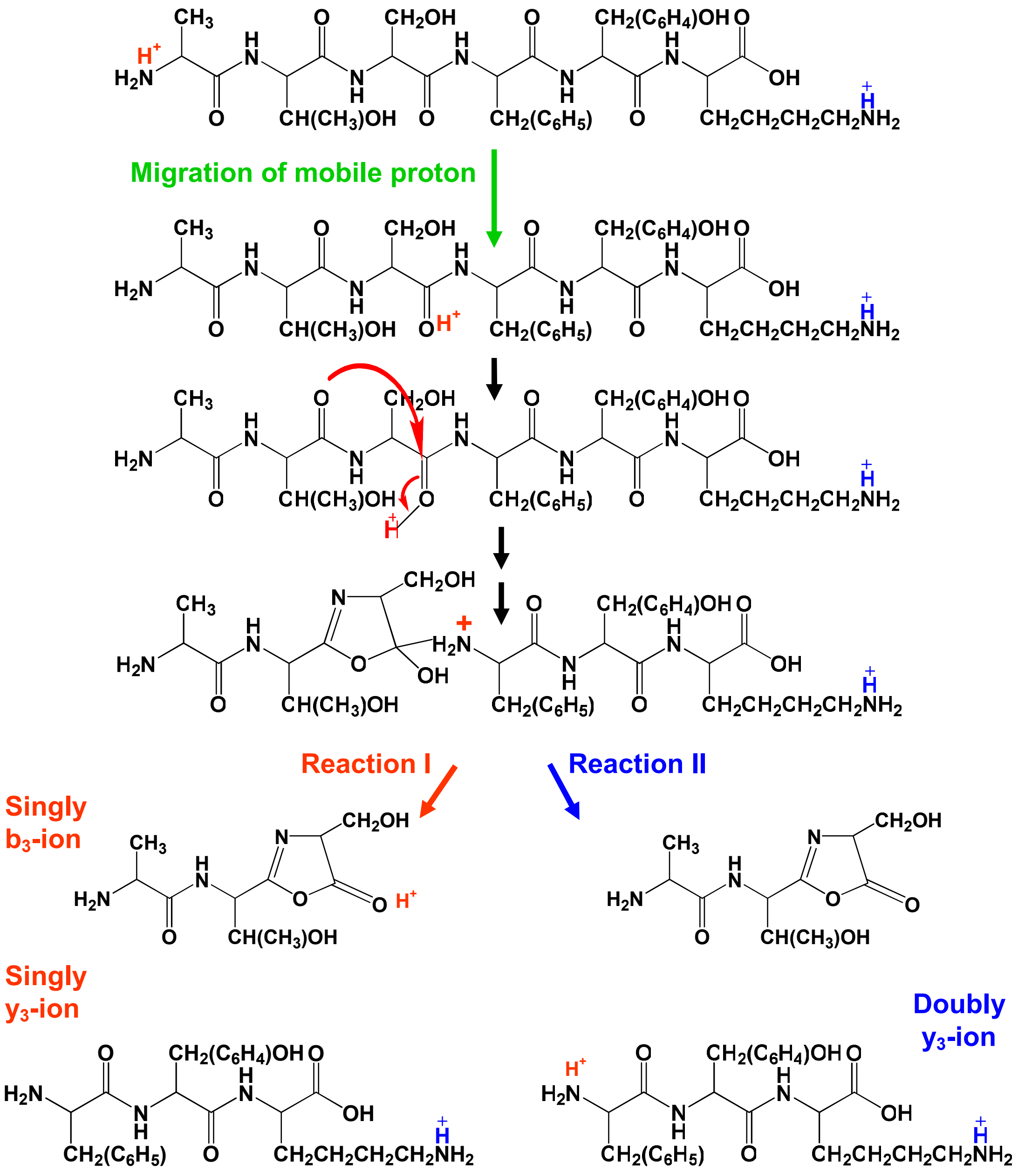
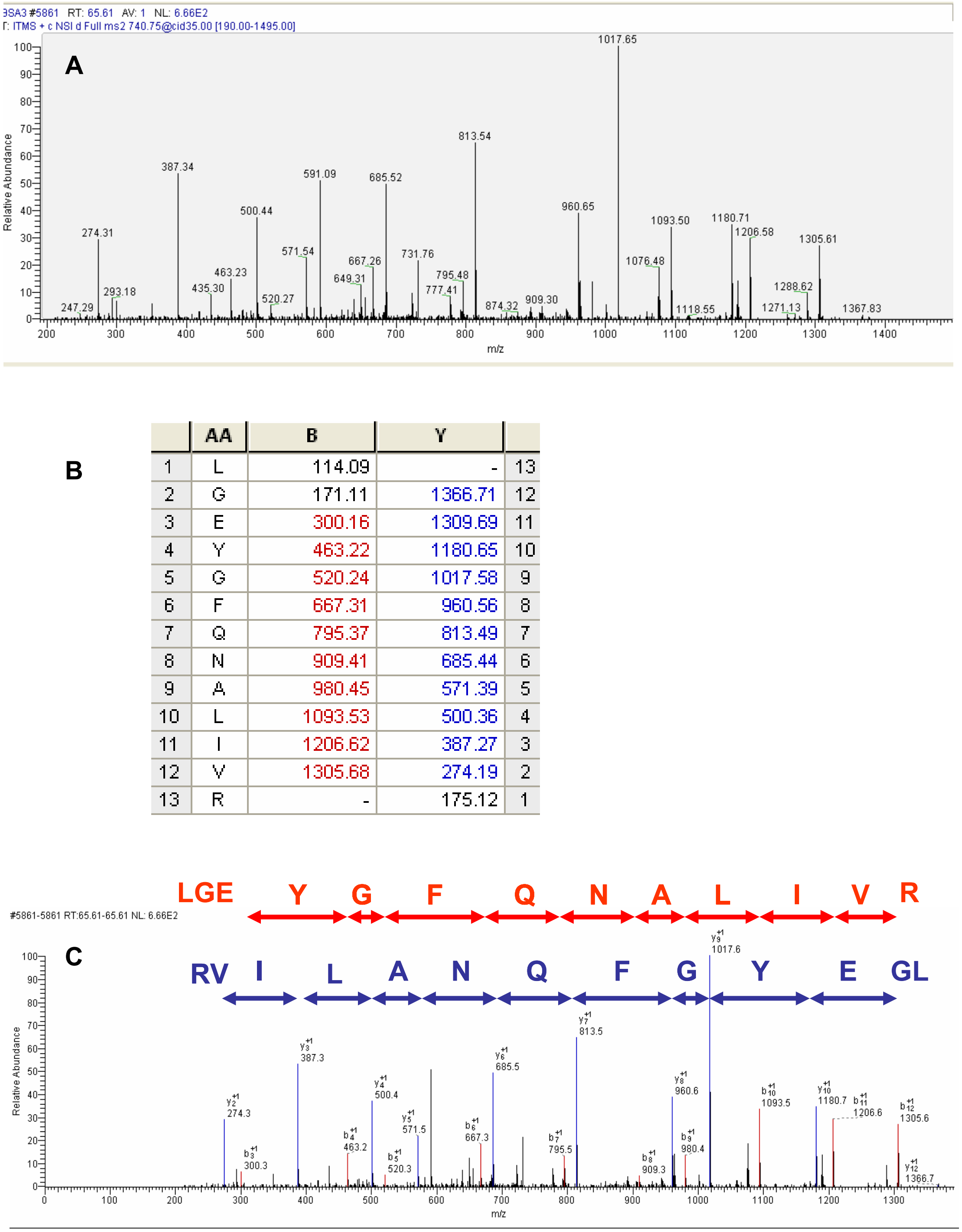


Figure 1.3 The mechanism of CID. A doubly charged tryptic peptide is used as an example. The peptide is initially protonated at two basic sites (N-terminus and lysine) by ESI or NSI. The proton attached to N-terminus can migrate down the length of the peptide according to "mobile proton model" [ref]. Following CID, this fragmentation pathway is initiated by the site of protonation using the newly formed peptide as an example, proceeds through a cyclic intermediate that subsequently fragments by one of two reactions. A singly charged b ion is formed by reaction I while the charge or the proton remains at the C terminus and a singly charged y ion is formed because the proton is originally attached at lysine. A doubly charged y ion is formed where reaction II occurs. A series of b and y ions are formed as different amide bonds cleave from the sub-population of the parent ion during CID.

sequencing by MS/MS spectrum interpretation is shown in Figure 1.4. In obtaining the correct peptide sequences through interpretation of MS/MS spectra, the enzyme used for generation of the peptide and peptide modification (see Section 1.2.2.3) should be taken into consideration. Commonly, tryptic peptides are used for sequence determination due to ease of ionization and spectral interpretation, given that a basic lysine or arginine is held at the C-terminus (see Section 1.3.3.3). Manual interpretation of the spectrum allows the mass difference between adjacent ions to be determined, and the corresponding amino acid to be determined (Fig 1.4C). A list of mass differences is compared with the mass values of amino acid residues [33], which are shown in Table 1.1. Interpretation is complicated by the fact that both b and y ions are present, but not always observed in the spectrum (see Fig. 1.4B & C). The loss of neutral molecules such as water and ammonia from b and/or y ions further complicates the spectrum [34,35]. Finally, some residues have the same *m/z* (e.g., leucine and isoleucine) and the combination of two residues can have the same *m/z* as one residue (shown in Table 1.2) [33,36]. Because of the extreme complexity of each MS/MS spectra, and the incredible number that are generated in a proteomics experiment (tens to even hundreds of thousands), computer-based interpretation algorithms are essential for peptide sequencing.

1.2.2.3 Database Searching for Protein Identification

Database searching involves correlating the experimental fragment ion spectra with theoretical spectra predicted for each peptide contained in a protein sequence database. Multiple protein sequence databases are available, including the comprehensive protein sequence databases derived by nucleotide sequences (e.g., the Protein DataBank), the curated protein databases built selectively from nucleotide sequences (e.g., Swiss-Prot



B

	AA	B	Y	
1	L	114.09	-	13
2	G	171.11	1366.71	12
3	E	300.16	1309.69	11
4	Y	463.22	1180.65	10
5	G	520.24	1017.58	9
6	F	667.31	960.56	8
7	Q	795.37	813.49	7
8	N	909.41	685.44	6
9	A	980.45	571.39	5
10	L	1093.53	500.36	4
11	I	1206.62	387.27	3
12	V	1305.68	274.19	2
13	R	-	175.12	1

Figure 1.4 The interpretation of an amino acid sequence of a doubly charged BSA peptide (740.75 *m/z)* obtained from an experiment. (A) MS/MS spectrum of the peptide “K.LGEYGFQNALIVR.Y”. (B) the list of b and y ions fragmented by the peptide. (C) The illustration of the peptide sequence deduction (note: red represents b ions and blue is for y ions)

Table 1.1 Residue masses of the 20 amino acids. Obtained from [33]

Amino	One-letter code	Residue mass (Da)
Glycine	G	57.02
Alanine	A	71.04
Serine	S	87.03
Proline	P	97.05
Valine	V	99.07
Threonine	T	101.05
Cystine	C	103.01
Leucine	L	113.08
Isoleucine	I	113.08
Asparagine	N	114.04
Asparate	D	115.03
Glutamine	Q	128.06
Lysine	K	128.09
Glutamate	E	129.04
Methionine	M	131.03
Histidine	H	137.06
Phenyalanine	F	147.07
Arginine	R	156.10
Tyrosine	Y	163.06
Tryptophan	W	186.08

Table 1.2 Amino acid combinations that are equal to a single amino acid residue mass. Obtained from [33]

Amino acid combination	Residue mass (Da)	Equivalent amino acid
GG	114	N
GA	128	Q, K
GV	156	R
GE	186	W
AD	186	W
SV	186	W
SS	174	Acrylocysteine

database) [37,38] and the Universal Protein Resource (UniProt) [39,40]. The protein database for *S. cerevisiae* as used in this thesis was downloaded from UniProt knowledgebase (UniProtKB).

Two of the most common search algorithms used for spectral interpretation are SEQUEST [41] (a mathematical correlation method) and MASCOT (a probabilistic approximation) [42]. Given that SEQUEST was used throughout this thesis, it will be explained in detail.

The process begins by generating a set of candidate peptides contained in the database which match the mass of the experimental parent ion. The highest accuracy for peptide identification is obtained by narrowing searches based on species identity, the enzyme used (including possible missed cleavages), possible peptide modifications (i.e., oxidation), and the mass tolerance for peptide and fragment ions. SEQUEST generates a theoretical tandem mass spectrum, calculating the b and y ions for each candidate peptide ion from the initial list and compares them against the experimental MS/MS spectrum, arriving at the candidate peptide which best matches the experimentally obtained spectrum. Each 'hit' from a SEQUEST search is provided as a possible match only, with the term cross-correlation (X_{Corr}) used for comparing the similarity between the theoretical and experimental fragment ion spectrum. X_{Corr} is the numerical value of the cross-correlation function, with a higher X_{Corr} meaning a stronger fit between the experimental and theoretical spectrum [43]. Additional parameters used in SEQUEST to rank the "goodness" of the fit include: the Delta Correlation value (ΔCN), showing the difference of the X_{Corr} values between the first and second hit in a search result; S_p, showing the preliminary score considering the number of matching ions between the theoretical and experimental MS/MS spectra; and R_{sp}, which ranks the particular matching peptide during the preliminary score. Setting the values of the individual parameters allows filtering of the list of matched peptides, providing results with a higher degree of confidence.

Typically, one peptide is sufficient to positively identify a given protein. However, proteins with similar amino acid sequences are often difficult to distinguish, unless the correct peptide is found. Translating a peptide list to a protein list is therefore a source of error in protein identification. Also, all search programs have errors at the peptide level as well, and represent false matches to the experimental MS/MS spectra. The false positive rate is a major parameter used to minimize the errors of a search result, and can be obtained through a decoy database search [44]. Here, the raw MS data are searched against a randomized amino acid database, or more typically a reversed sequence database, using the identical parameter set for the true (forward) database search. The false positive rates are calculated by doubling the number of hits by decoy searching (known to be false) and dividing this by the total number of hits by the forward search. Peptide filters are considered to be sufficient when this rate is below 1%.

1.2.3 Separation Methods

1.2.3.1 RPLC for Peptide Separation Prior to MS

In a MS experiment, ESI and NSI are subject to ion-suppression effects. Ion-suppression can be caused by highly concentrated samples, or samples containing non-peptide ions (e.g., salts). Weaker signals are observed in these instances due to signal saturation above a concentration of ~l0^{-5} M. Also, considering the complexity of a proteome, overlapping signals from peptides with identical or near-identical *m/z* values would prevent ions from being observed. For these reasons, a complex peptide mixture must be simplified (i.e., separated) prior to MS/MS analysis. Reversed phase liquid chromatography is a high resolution peptide separation tool that simultaneously separates and desalts samples by employing MS-compatible solvents. Therefore, coupling RPLC to NSI-MS enhances sensitivity by generating higher analyte fluxes as peptides elute as narrow chromatographic peaks and are electrosprayed into the MS detector.

Reversed phase refers to the fact that the column contains a non-polar stationary phase and uses relatively polar mobile phase solvents (i.e., water and acetonitrile with 0.1% (*v/v*) formic acid) to elute analyte molecules. Traditionally, a C_{18} column is used in reversed phase peptide separation. This octadecyl (C_{18}) chain is bonded to silica beads which constitute the stationary support, allowing separation of analytes based on hydrophobicity. To improve the efficiency of separation and ionization of peptides, capillary-size columns, which have internal diameters ranging from 50 to 300 microns, are commonly used. Not only is this favored for coupling to NSI, but it is also often the case that sample is extremely limited and smaller diameter columns eliminate the need for sample dilution. A capillary column is formed by directly packing the beads in a pulled capillary tube leading to the nanospray emitter tip. This eliminates post-column peak broadening by removing all dead volume ahead of the nanospray tip.

For high-efficiency peptide separation, two or more orthogonal dimensions of LC separation are used in whole-proteome analysis (the so-called shotgun proteomics). A well-known example of multi-dimensional LC separations is known as Multidimensional Protein Identification Technology (MudPIT) [45] where a mixture of peptides directly by strong cation exchange (SCX), followed by RP microcapillary LC (µLC) prior to MS analysis. Although an effective means of peptide separation, differentiation of protein isoforms or homologues is not possible with MudPIT. Another disadvantage of MudPIT is a masking effect of highly abundant proteins over of low-abundance proteins due to ion suppression can result in missed identifications of these low abundance proteins [46]. Thus, many identified peptides belong to a small group of highly-abundant proteins which makes the interpretation of results of shotgun proteomic experiments at the protein level difficult. Masking effects can be decreased by separation at the protein level [25]. Separation of proteins prior to digestion can provide some constraints for protein identification (see Section in 1.3.3.3), improving the confidence of proteins identified and also improving the number of proteins identified. Protein separation prior to digestion has become more widely used in proteomics studies [47,48].

1.2.3.2 Gel-Based Protein Separation

An effective means of protein separation employs two dimensional gel electrophoresis (2DGE) in which proteins are separated based on their isoelectric point (pI; the point where a protein experiences no net charge) in a process known as isoelectric focusing (IEF) and then by molecular weight (MW) using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) [49,50]. This method has unrivalled resolution and a theoretical separation capacity of 10,000 spots [51]. Separated proteins can be visualized using various gel staining techniques, with coomassie dyes [52] or silver stains [53] being most common [54]. To some extent, visualization by staining makes the proteome profile easy to see. Next, the separated proteins can be selectively excised from the gel and sequenced by MS/MS. Although effective for analysis of the majority of the proteome, this standard method is not compatible with proteins with extreme pI values or masses and IMPs. These extreme proteins are lost by 2DGE due to the problems with IEF such as solubility and the limited pI range [55].

Limitations of 2DGE are addressed by coupling one-dimensional (1D) SDS-PAGE with nanospray RPµLC MS/MS, so-called GeLC MS/MS [56,57]. SDS is an ionic detergent that is extremely efficient at solubilizing and denaturing proteins, allowing SDS-PAGE separation of proteins with a wide range of pI values, MWs, and hydrophobicity [58,59]. A schematic of the GeLC MS/MS workflow is shown in Figure 1.5. This separation and identification method is an unbiased platform for complex protein mixtures [60].

Despite optimization of the process, complete recovery of peptides from gel slices is still a challenge [61]. In response to variable recoveries from gel slices, another method has been developed known as Gel Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE) [62]. GELFrEE separates proteins by MW through a tube column containing a polyacrylamide gel, similar to 1D SDS-PAGE, but proteins elute into solution phase, allowing more complete recovery of protein samples. Because SDS interferes with trypsin digestion, removing or decreasing SDS from the prefractionated proteins is necessary [59]. SDS is effectively removed by organic solvent precipitation. Precipitated proteins can then be resolubilized and digested, and then analyzed by traditional LC-MS/MS. Although the gel-based platforms are proper for intact protein prefractionation, these platforms are difficult to automate. Therefore, solution-based separation methods have recently attracted more attention.

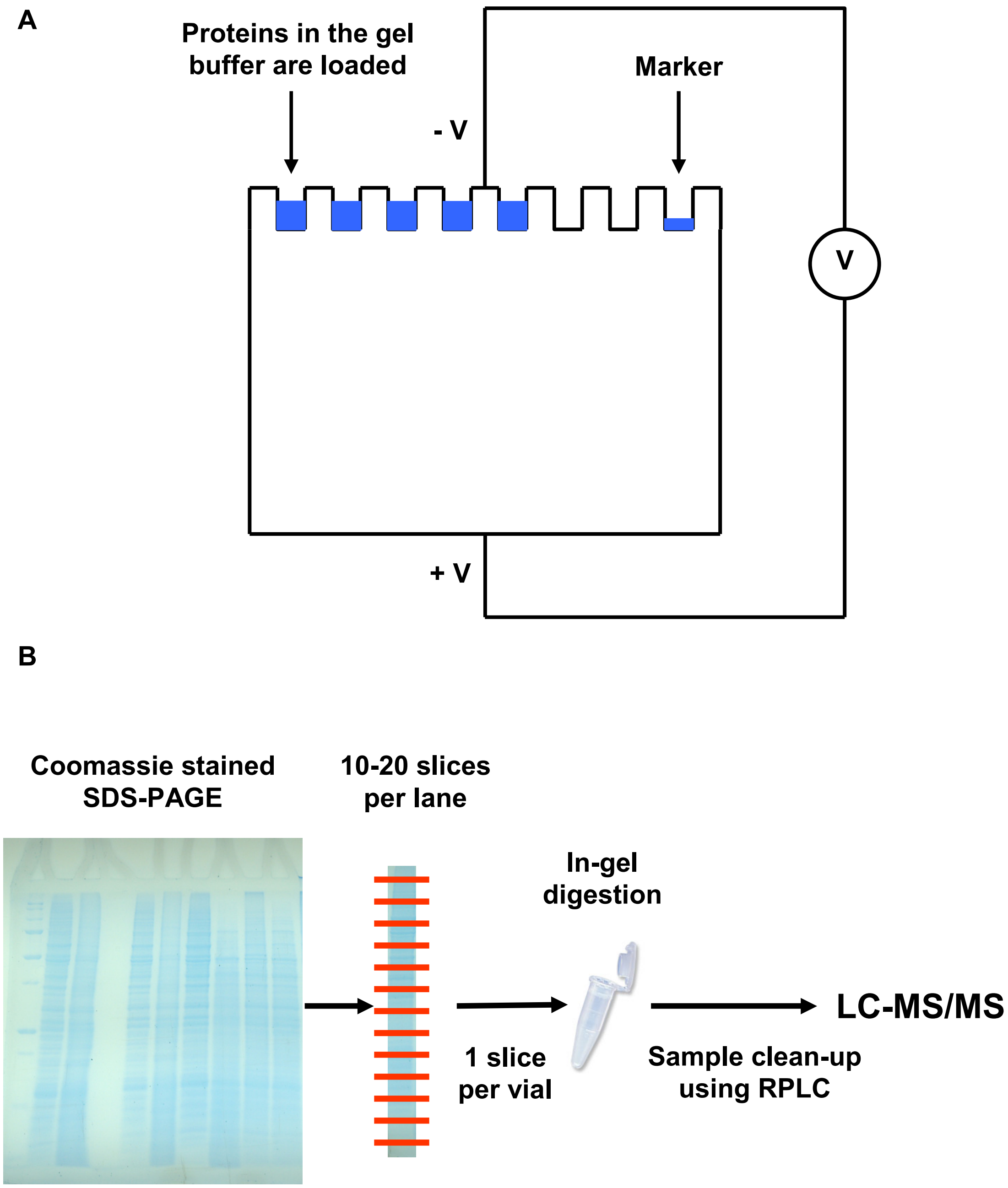


Figure 1.5 Schematic of GeLC experimentation. (A) The method of SDS-PAGE is shown. Proteins in the gel buffer are loaded into a SDS-PAGE lane and separated by applying voltage (i.e., 240 V). (B) The proteins in the gel are stained using Coomassie dye or silver. Each gel lane is cut into 10-20 slices, and then each slice is further diced into smaller pieces (~1mm^2). All the pieces are placed into a vial and digested by adding trypsin according to the in-gel digestion protocol [61]. The extracted peptides are cleaned by RPLC and then analyzed through LC-MS/MS analysis.

1.2.3.3 Solution-Based Protein Separation

The use of LC for protein separation has the advantages of high loading capacity, speed, reproducibility, ease of automation, and direct coupling to ESI-MS. Of the possible LC separation methods, RPLC is preferred as the solvents are compatible with ESI or NSI. Solvents used in RPLC normally are the mixture of water and a water miscible organic solvent such as acetonitrile, both containing ion-pairing reagents such as trifluoroacetic acid (TFA). RP columns for protein separation are generally made up of C_4, C_8, C_{18}, or RH solid phase beads [63,64]; however, protein separation by RP is not as effective as peptide separation due to the size and solubility of proteins. Because IMPs, especially those containing high number of transmembrane alpha-helices, are hydrophobic, they are insoluble in the water/acetonitrile mobile phase and cannot be effectively separated. Consequently, for proteomic RPLC separation including transmembrane proteins, alternative solvents are needed.

Previous studies have shown formic acid and isopropanol solvent systems to be compatible with membrane protein separation by RPLC [59,65]. Because of the good solubility of hydrophobic proteins in formic acid, this solvent system was selected for large-scale analysis of the yeast proteome by RPLC. Reversed phase materials generally used for membrane protein separation are silica–based, such as C_4 and C_{18} [66-68], or polymer resin (poly styrene-divinyl benzene (PSDVB) copolymer beads), such as R1, R2 and Polymer beads [65,69]. PSDVB polymers are similar to reversed phase beads in that they have similar selectivity characteristics to C_8 or C_{18}-bonded silica columns. An advantage of these columns is that there are no pH limitations, as the beads are polymers, allowing a wider range of pH to be used. Protein prefractionation methods employing the formic acid/isopropanol solvent system and different RP columns are investigated for whole yeast proteome separation.

Although one dimensional separation can be effective for the resolution of simple protein mixtures, complex mixtures require an additional dimension to reduce the complexity of the sample and improve the efficiency of separation of peptides. Here, yeast proteins prefractionated by RPLC are digested with trypsin or CNBr (cyanogen bromide) and trypsin. The peptides are further separated and identified by RPμLC MS/MS or LC-MS/MS.

1.3 Biological Sample Preparation

Sample preparation is one of the most crucial processes in proteomics research because the results of experiments are dependent on the proper handling of the starting material. For different types of biological samples, different methods of preparation are required. Samples for protein analysis can be classified as three types: tissues, cell populations (from cell culture), and biological fluids. These biological samples not only have high variability in abundance, charge, size and hydrophobicity of proteins, but also in their cellular and subcellular distribution, expression level, and post-translational modifications. Thus, a raw proteome is an incredibly complex system even considering a relatively simple organism such as yeast, *S. cerevisiae* (model used in the thesis), which needs to undergo numerous manipulations to prepare the sample for analysis. Before explaining the preparation methods, the background of the biological sample will be described first.

1.3.1 Basic Biological Background of *S. cerevisiae*

S. cerevisiae is a well-studied organism, being the first eukaryotic genome to be completely sequenced [1]. The genome is composed of 12,156,677 base pairs and 6,275 genes, compactly organized on 16 chromosomes. Only about 5,000 of these are believed to be true functional genes [70].

A yeast *S. cerevisiae* cell is generally ellipsoidal in shape, ranging from 5 to 10 μm in length, with subcellular structures consisting of a cell wall, peripheral plasma, plasma membrane, cytoplasm, nucleus, mitochondria and other organelles. The structure of a yeast cell is shown in Figure 1.6. All the proteins in the cell can be grouped into cytosolic (soluble) and membrane (hydrophobic) proteins.

Membrane proteins, especially IMPs, have important roles in regulating yeast nutrition (i.e., uptake of necessary compounds or ions and extrusion of molecules hazardous to the cell) and other essential cellular processes, such as signal transduction, vesicle trafficking, energy generation, molecule transport, cell adhesion, and intercellular communication [2]. A well known example is an active ion transporter membrane protein, Na^+/K^+ ATPases. This protein transfer of K^+ into the cellular membrane and transfer of Na^+ out is critical to cell survival [71]. Thus, IMPs receive considerable attention in proteomic research for their critical roles in life [72,73].

The proteins associated with biological membranes fall into two main categories [74]: extrinsic (peripheral) and intrinsic (integral) proteins (Figure 1.7). Peripheral membrane proteins are associated with the surface of the membrane, temporarily adhering to the membrane *via* non-covalent interactions with phospholipid head groups or IMPs embedded in the membrane. The majority of these proteins are actually water

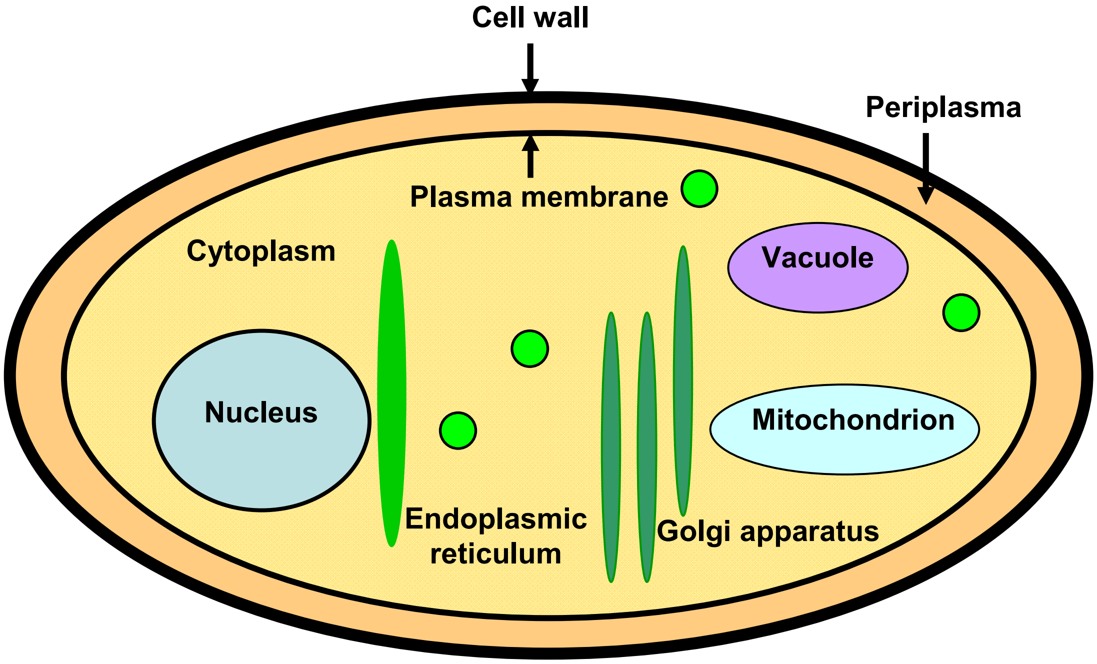


Figure 1.6 Structure of yeast *S. cerevisiae* cell. *S. cerevisiae* cells have shape that are round to ovoid, 5 to 10 µm in diameter. A cell consists of a cell wall, peripheral plasma, plasma membrane, cytoplasm and organelles such as nucleus, vacuole, mitochondria, endoplasmic reticulum, Golgi apparatus, and others (green balls). The wall of a yeast cell is a remarkably thick (100 to 200 nm) envelope. Inside the cell, a salty cytoplasm takes up most of the cell volume, in which soluble proteins are dissolved. The plasma membrane itself is composed of a phospholipid bilayer and includes a variety of membrane proteins in approximately equal proportion to the phospholipids based on mass.

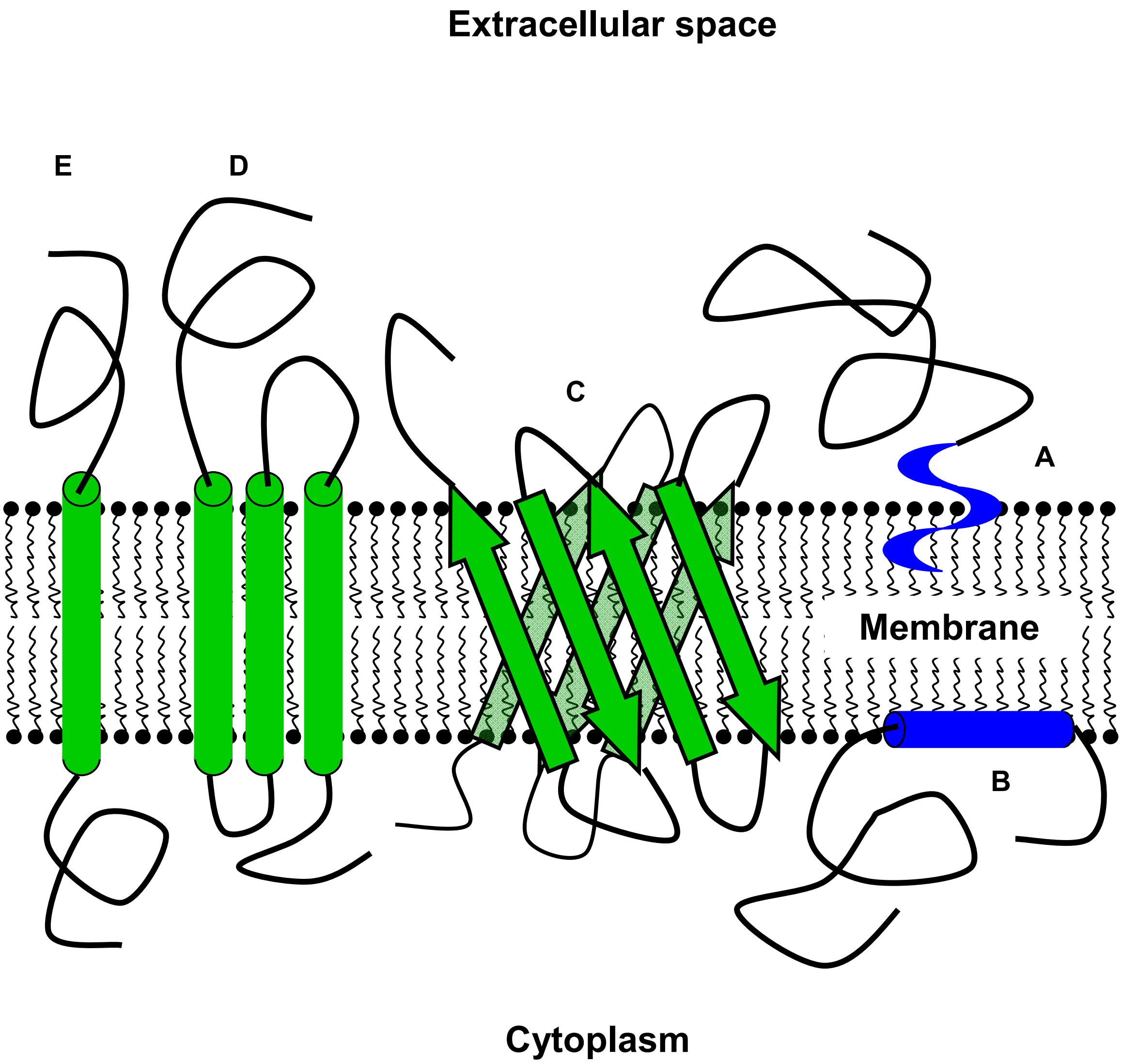


Figure 1.7 Schematic presentation of the types of integral membrane proteins. Anchored membrane proteins (A, B) are permanently associated with the membrane and do not have any segments that fully span the membrane. Transmembrane proteins contain segments across the membrane. They are cataloged as beta-barrel transmembrane proteins (C) and alpha-helical transmembrane proteins (D: one helix domain; E: multi helices spanning across the membrane). Peripheral membrane proteins are associated with those proteins (not shown).

soluble, unlike IMPs which are permanently attached to the membrane and form part of it. IMPs in particular are among the most challenging to manipulate for proteome analysis.

Transmembrane proteins make up the majority of IMPs and are characterized by two main secondary structures: transmembrane β-barrel porin-type domains and one or more transmembrane α-helices (TMH) (Fig. 1.7C-E). The alternate polar and hydrophobic amino acid residues of the transmembrane β barrels are present in the central pore and lipids, respectively. Only a few percent of the total proteome are β-barrel proteins and their overall hydropathy values are similar to that of soluble proteins, making them easier to identify than TMH containing proteins or transmembrane proteins [59,75].

TMH containing proteins provide the greatest challenges for proteome analysis. They have one or more α-helical transmembrane domain composed of hydrophobic or apolar amino acid residues which spans the membrane. Due to these hydrophobic transmembrane domains, IMPs have a tendency to aggregate and precipitate in purely aqueous solutions after being removed from the bilayer. These physico-chemical properties of transmembrane proteins result in the need for special sample preparation protocols and separation methods for them compared to soluble proteins.

1.3.2 Protein Properties

From a separation standpoint, protein properties are typically sorted into several groups according to their physical and chemical characteristics rather than their biological functions. These properties are pI, MW, abundance, and hydrophobicity. The proteome of yeast has wide ranges of these properties such as molecular mass (from >180 kDa to <10 kDa), pI (acidic <4.3 to basic >11), abundance (high to low), and hydrophobicity [45,76]. It should be mentioned that 75% of the yeast proteome belongs to the low abundant proteins (fewer than 5,000 molecules per cell) [10] and 30% of the proteome are IMPs (hydrophobic proteins) [77]. For whole proteome analysis, an unbiased separation method is needed to systematically identify proteins, meaning that all classes of proteins should be identified (e.g., hydrophilic and hydrophobic, small and large proteins).

Among these protein properties, poor solubility of proteins presents a unique challenge for MS analysis. Unlike pI and MW which can be determined by amino acid composition (obtained from the ExPASy website: ca.expasy.org/tools/pi_too.html), hydrophobicity of IMPs are obtained by using the grand average of hydropathy (GRAVY) score as an estimated measurement [78]. The GRAVY score is an important value to qualitatively differentiate soluble proteins from membrane/transmembrane proteins as it considers not just the hydrophobicity of each residue in the protein but also the overall size of the protein. This provides some indication of the physical state of the protein. A positive GRAVY score means that the protein is hydrophobic while a negative score indicates it is hydrophilic.

A GRAVY score of a protein is calculated using equation 1.1, in which N stands for total number of residues in the protein/peptide and N_i is for the number of one residue in the protein/peptide. The hydropathy values of 20 amino acids (presented as $A_{\text{hydropathy}}$) can be obtained from the literature reported by Kyte *et al.* [78], and the values are shown in Table 1.3.

$$\text{GRAVY} = \Sigma\,(N_i \times A_{\text{hydropathy}})/N \tag{1.1}$$

GRAVY scores for *S. cerevisiae* were obtained from ‘Sequence Manipulation

Table 1.3 Hydropathy scale used in GRAVY calculation. Obtained from [78]

Amino	One-letter code	Hydropathy index
Glycine	G	-0.4
Alanine	A	1.8
Serine	S	-0.8
Proline	P	-1.6
Valine	V	4.2
Threonine	T	-0.7
Cystine	C	2.5
Leucine	L	3.8
Isoleucine	I	4.5
Asparagine	N	-3.5
Asparate	D	-3.5
Glutamine	Q	-3.5
Lysine	K	-3.9
Glutamate	E	-3.5
Methionine	M	1.9
Histidine	H	-3.2
Phenylalanine	F	2.8
Arginine	R	-4.5
Tyrosine	Y	-1.3
Tryptophan	W	-0.9

Suite: Protein GRAVY' (www.ualberta.ca/~stothard) by submitting the yeast protein sequences in FASTA format obtained from UniProt. Values for peptides are easily calculated using an in-house generated Excel program.

1.3.3 Sample Preparation

1.3.3.1 Protein Extraction

The first stage of proteome analysis begins with the isolation of the protein from the sample. If proteins are already contained in an aqueous medium, as is the case for analysis of biological fluids such as plasma and urine, then the proteins can be isolated by centrifugation, ultrafiltration, spin column purification, or solid phase extraction. Proteins from cells or tissues must first be removed from the cell. As seen in the structure of a yeast cell (Figure 1.6), the yeast cell has a hard cell wall, which requires physical disruption in order to extract proteins from the cell [79]. Homogenization allows extraction of proteins from yeast cells. There are five homogenization methods: mechanical, ultrasonic, pressure, freeze-thaw, osmotic and detergent lysis [80]. Pressure homogenization, as seen with the French Press, was used for sample homogenization in this thesis.

The French Press [81] is an effective method for disrupting eukaryotic cells which have a resilient plasma membrane and/or cell walls. Cells are suspended in a lysis buffer, which can include detergents or other additives to aid protein solubilization (e.g., SDS or urea). The cell suspension is then placed into a cylinder which is brought up to an extremely high pressure (~10,000 to 20,000 psi) and squeezed past a needle valve. Disruption of the cells occurs as the pressure bomb is opened, rapidly releasing the pressure on the cells causing lysis. Cellular debris can be removed by low speed centrifugation. At this point, assuming no detergents have been added, the membrane fraction can be separated from the soluble fraction through ultra-centrifugation (130,000×*g*).

For a comprehensive proteome analysis, all proteins (including hydrophobic proteins) should be isolated and brought into solution phase. With the aid of chaotropes or detergents, solubility as well as extraction efficiency of proteins is improved. SDS is an effective solubilizing agent consisting of an anionic organosulfate with a 12-carbon tail attached. Transmembrane proteins are solubilized by disruption of non-covalent bonds in the proteins by the hydrocarbon tail, and aggregation is prevented by the anionic sulfate portion of the molecule. SDS is a compatible reagent to be used on downstream separations such as SDS-PAGE.

1.3.3.2 Protein Precipitation and Solubilization

The presence of non-protein material such as lipids and salts in the biological sample can cause problems in the downstream analysis of the sample. Addition of sample additives, including SDS, can also interfere with the detection of proteins, as well as interfering with protein digestion. Isolation of proteins from the remainder of the biological matrix is accomplished by two main methods [82]: solid phase extraction, and protein precipitation. Solid phase extraction is virtually identical in concept to reversed phase separation, except that all proteins are released from the column as a single fraction. Solid phase extraction is widely applied to isolate proteins, but comes at a risk of biased protein elution. Not all proteins will bind to the solid support, nor are all proteins released from the column during the extraction phase. As an alternative strategy, the precipitation of proteins with organic solvent can be used to isolate or enrich the analyte and eliminate interfering components. Two commonly used precipitation methods are acetone precipitation [82,83] and chloroform/methanol/water (CMW) [84].

Precipitation methods work by displacing water from the solvation layer which surrounds proteins in solution, increasing protein-protein interactions and causing proteins in solution to aggregate and precipitate. Cold acetone with a solvent ratio of 4:1 (cold acetone:sample; *v:v*) [85] precipitates proteins while lipids, salts, and detergents, mostly remain in solution. Methanol/chloroform/water/sample (3:1:4:1 *v:v:v:v*) [84], will cause lipids to partition into the chloroform layer, while salts and detergents remain in the aqueous/methanol layer, and proteins precipitate at the chloroform and methanol aqueous interface. The resulting pellets, made up of precipitated proteins, can be resolubilized using various reagents (e.g., formic acid, 8M urea, or 1-4% SDS), with the help of sonication and/or mechanical agitation. The reagent used is dependant on the downstream application (i.e., 60% formic acid for RPLC, SDS for SDS-PAGE).

1.3.3.3 Digestion

The most common detection strategy in proteomics is currently known as a “bottom up” approach, referring to the LC-MS/MS analysis of peptide fragments generated by protein digestion [25,86]. Trypsin is the most widely used proteolytic enzyme which cleaves proteins at the peptide bond on the carboxyl side of lysine and arginine residues (unless followed by proline). Trypsin requires a slightly alkaline environment in order to maintain its activity (pH ~8). Knowing that all peptides produced by tryptic digestion (except the C-terminal peptide) have either a lysine or an arginine at the N- and C-terminus (see Figure 1.8A), it is easier to identify proteins by constraining the database searches in the interpretation of the product ion spectrum.

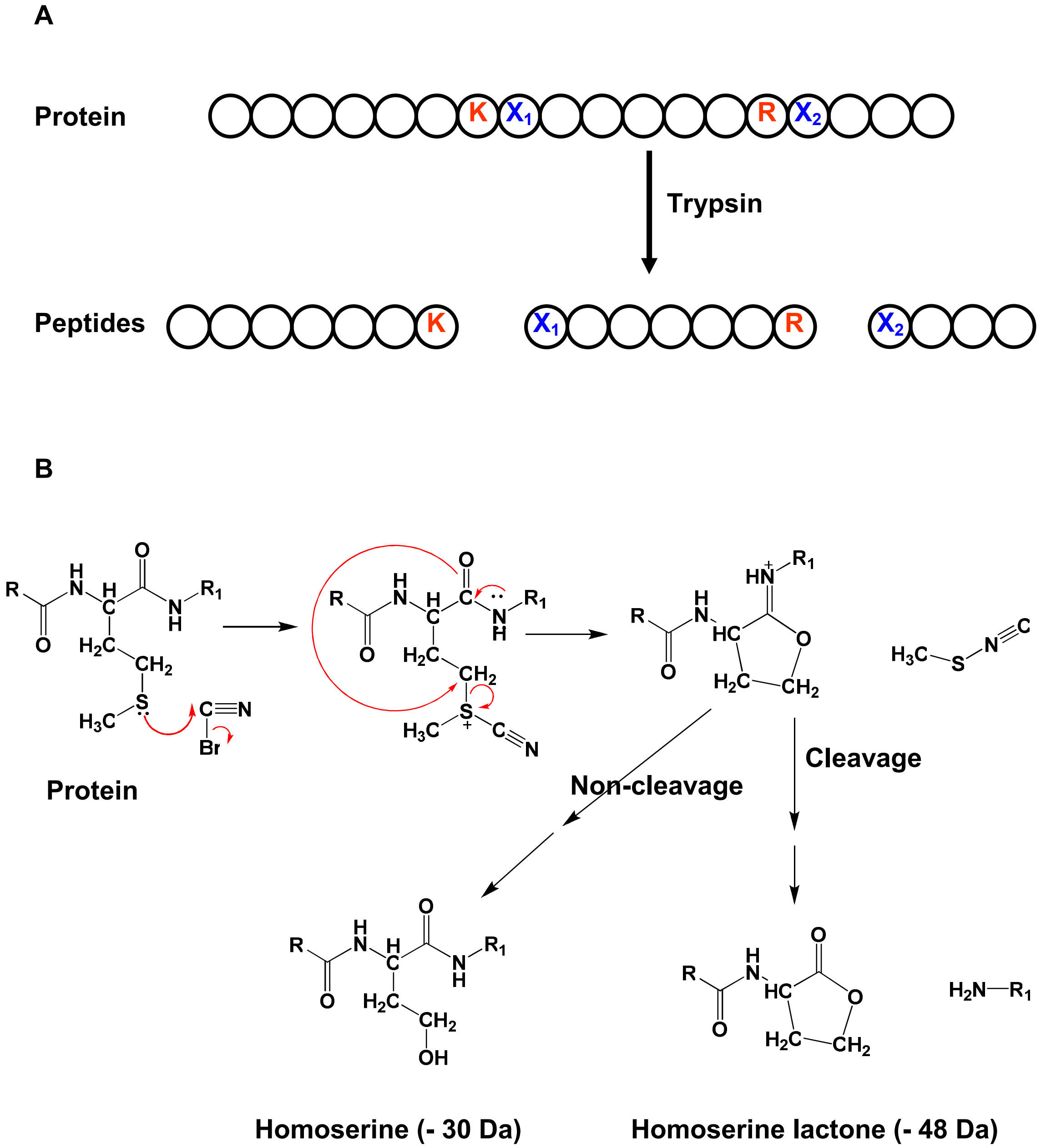


Figure 1.8 Schematic diagrams of trypsin (A) and CNBr (B) digestions. Trypsin cleaves peptide bonds at the C-terminus of lysine (K) or arginine residues (R). X represents other amino acid residues. (B) CNBr hydrolyzes peptide bonds at the C-terminus of methionine residues through a formation of a five member ring intermediate. Homoserine lactone of peptides are formed, but modification of proteins is formed as homoserine due to non-cleavage effect.

The chemical reagent CNBr is capable of chemical digestion of proteins [87]. The reagent can be used for all protein samples, but is most common with membrane protein digestion due to scarcity of the basic residues (i.e., lysine and arginine) within the transmembrane domains. For CNBr digestion, proteins are dissolved in a high concentration of acid (e.g., 70% formic acid). CNBr selectively reacts with methionine residues so the amide bond on the carboxyl side of methionine residue is cleaved and peptidyl homoserine lactone and aminoacyl peptide fragments are formed [45,88]. The schematic diagram of this reaction is shown in detail in Figure 1.8B. Thus, proteins are chemically cleaved into peptides which may be further digested into smaller peptides using trypsin digestion, so-called CNBr/trypsin digestion. Using the CNBr/trypsin digestion, identification of peptides produced from transmembrane proteins is improved.

1.4 Research Proposals

The research presented in this thesis represents efforts toward assessment of universal protein separation and identification methods. Due to the complexity of biological samples, an unbiased proteome analysis protocol is critical to completely profile a biological system. Among all classes of proteins, integral membrane or transmembrane proteins provide a significant challenge, particularly for protein separation owing to their poor solubility in aqueous phase. However, these proteins are very important for the biological systems because they are essential in fundamental intracellular processes. Thus, to asses a universal proteomic prefractionation method, unbiased identification of membrane proteins, especially transmembrane proteins are used as an important criterion.

SDS-PAGE separation platforms, including GeLC and GELFrEE, are suggested to be compatible techniques for proteome characterization due to their use of SDS during separation. This detergent has an unrivalled ability to assist in the solubilization of hydrophobic proteins, and GeLC MS/MS is well characterized to separate hydrophobic proteins [59,89]. GELFrEE MS/MS has the potential to match the performance of GeLC MS/MS, given that a similar separation method is involved, however, the two platforms differ greatly post-separation, including the SDS removal and digestion methods used between them. As of yet, the GELFrEE MS/MS platform has not been assessed in terms of its ability to characterize hydrophobic proteins. In Chapter 2, these two methods are evaluated for comprehensive proteomics analysis. For evaluation, a reliable dataset of yeast integral membrane /transmembrane proteins must first be established. While computational approaches are available to classify proteins (membrane *vs* cytosolic), there is considerable disagreement between the varying programs. Thus, an additional objective of Chapter 2 is to develop a more reliable computational approach to classify the yeast proteome dataset.

When coupled to MS, gel-based platforms provide a useful means of protein identification, however coupling these techniques to MS is difficult to automate, and is limiting in certain applications. Therefore, easy to automate solution-based separation platforms are preferred for high-throughput proteome analysis. Because of the incompatibility of SDS to RPLC techniques, these platforms present entirely new demands on sample preparation. Chapter 3 investigates alternative strategies to extract, purify and solubilize proteins. Sample preparation generally entails the following: (1) cell disruption and proteome extraction; (2) protein cleanup; and (3) final resolubilization in a

RPLC-compatible solvent. Three stages of sample preparation are investigated to establish a proper (compatible) sample preparation protocol for proteome prefractionation by RPLC (objective of Chapter 4).

Finaly, Chapter 4 investigates a solution-based separation platform for profiling of the yeast proteome. An alternative solvent system is proposed to improve protein retention and recovery from reversed phase gradient separation. The solution separation and identification platform is compared with gel-based separation analysis in terms of the ability to identify a representative distribution of yeast proteins, according to size, isoelectric point, and hydrophobicity.

Chapter 2

Assessment and Distribution of Proteins Identified through GELFrEE MS/MS *vs* GELC MS/MS

2.1 Introduction

Proteomic research focuses on identifying proteins expressed in a cell, tissue or organism. However, traditional proteomics methods are problematic when applied to integral membrane proteins (IMPs), in particular alpha helical transmembrane proteins (herein referred to as transmembrane proteins in this chapter), are hydrophobic, and thus are generally insoluble in aqueous solutions, which cause problems with solution phase separations such as reversed phased liquid chromatography (RPLC). Hydrophobic proteins also have reduced ionization efficiency making detection by mass spectrometry (MS) challenging.

While difficult to analyse, IMPs (especially transmembrane proteins) are among the most important classes of proteins, making up approximately 30% of the proteome in living organisms [77]. The helix bundle class of IMPs is essential for fundamental cellular processes such as electron transport, cell-to-cell communication, and transportation of molecules across cell membranes. IMPs also account for a large percentage (70%) of all identified drug targets [90,91] making identification of IMPs especially important in comprehensive protein identification studies.

The work in this chapter is divided according to two major themes. The ultimate goal of this chapter is to perform a comparison between analytical methods (i.e., GeLC and GELFrEE MS, see Section 2.3) for proteome analysis, in terms of their ability to identify IMPs and transmembrane proteins. To enable such a comparison, a detailed assessment of bioinformatic approaches to classify proteins must first be performed (Section 2.2).

2.2 Transmembrane Protein Prediction

Prior to performing a comparison of IMPs, transmembrane proteins or otherwise hydrophobic yeast proteins identified through various analytical workflows, the reference yeast proteome database needs to be classified accordingly. Unfortunately, no dataset exists which accurately classifies all yeast proteins according to membrane proteins and cytosolic proteins. While all yeast protein sequences are provided by the literature from their gene sequences, specific cellular locations and structures of membrane proteins are incomplete [75]. X-ray crystallography is the preferred method for obtaining structural information of proteins, but it has limited applicability for integral membrane protein structure prediction. The limitations are due to inherent hydrophobic properties of IMPs or the difficulties in crystallization of them. The structures of large membrane proteins are also difficult to obtain by other means, such as nuclear magnetic resonance methods [92]. Thus, structural classification of membrane proteins has been hindered due to the limited amount of available structural data [93]. Of the 60,369 protein structures currently deposited in the Protein Data Bank, only 1,062 (1.7 %) are for membrane proteins [94]. As a consequence, computational methods remain the only viable alternative for integral membrane/transmembrane protein prediction.

For IMP prediction, gene ontology (GO) is a commonly used approach. Gene classification terms were first applied to genomic studies of *Saccharomyces cerevisiae* in

1998. The GO terminology was developed and then established in 2000 [95]. In brief, GO terms provide specific annotations which classify proteins according to biological process, cellular components, and molecular functions. The GO term specifying IMPs is “GO: 0016020”. The original classification of a given protein according to a particular GO term can be obtained through direct experimental evidence, inferred evidence, or purely through prediction software. Thus, while GO terms are generally considered reliable, protein classification according to these terms must be taken with a certain level of discretion.

There is presently no GO term to isolate helical transmembrane proteins from a proteome. However, a variety of algorithms (over 30) have been established for topology prediction of transmembrane alpha helices (TMHs) in IMPs [59], which predicts not only the number of TMH, but also the location of the TMH in the protein sequences. Most of these algorithms claim extremely accurate results based on assessment of various training sets, referring to libraries of experimentally known transmembrane proteins. Although no single prediction algorithm consistently outperforms the other when compared within different training sets [96], the training set does provide an artificial means of assessing the reliability of the method. Compared to the whole proteome of an organism, the training set is a far simpler dataset with lower sample diversity. To date, a systematic comparison of TMH prediction tools has not been performed using a complex proteome dataset. Therefore, a reliable dataset of transmembrane proteins (distinguished from soluble proteins) of the yeast *S. cerevisiae* cannot depend on a single prediction method [72]. Nilsson *et al.* suggested a strategy using a consensus of multiple prediction methods to establish the most accurate list of transmembrane proteins from a given proteome [97], therefore a reliable dataset of yeast transmembrane proteins might be obtained by a combination of the predicted results from several existing prediction algorithms.

Prediction algorithms dependant on a hidden Markov model (HMM) [98,99] tend to predict the structure of integral membrane proteins with greater accuracy than non-HMM [59]. The HMM is a statistical model based on classification of different parts of the protein (shown in Figure 2.1) that has been widely used for TMH prediction since 1989 [75]. Each protein consists of various main structures states, including the helix caps (intracellular or extracellular tails are closely attached to a helix), helix core, loop (cytoplasmic side or non-cytoplasmic side), and a globular domain in a loop (see Fig. 2.1). Each state of the protein displays different hydrophobic character or characteristic distribution of amino acids that HMM models use to align protein sequences within these states. For instance, a transmembrane alpha helix consists of an unusually long stretch of hydrophobic residues, with positively charged residues (arginine and lysine) being mainly found in non-transmembrane loops on the cytoplasmic side. TMHs can be predicted based on the probabilities of amino acids being distributed as clusters along the protein sequence.

The three prediction algorithms known as TMHMM2 [100], HMMTOP2 [101,102], and Phobius [103] employ the HMM and are reported to rank highly in terms of their ability to predict helical transmembrane proteins [59,96]. Therefore, these algorithms were selected for prediction of transmembrane proteins in this thesis. Algorithms based on different principles can help to improve the accuracy of transmembrane protein prediction. Thus, SOSUI [104] (meaning “hydrophobic” in

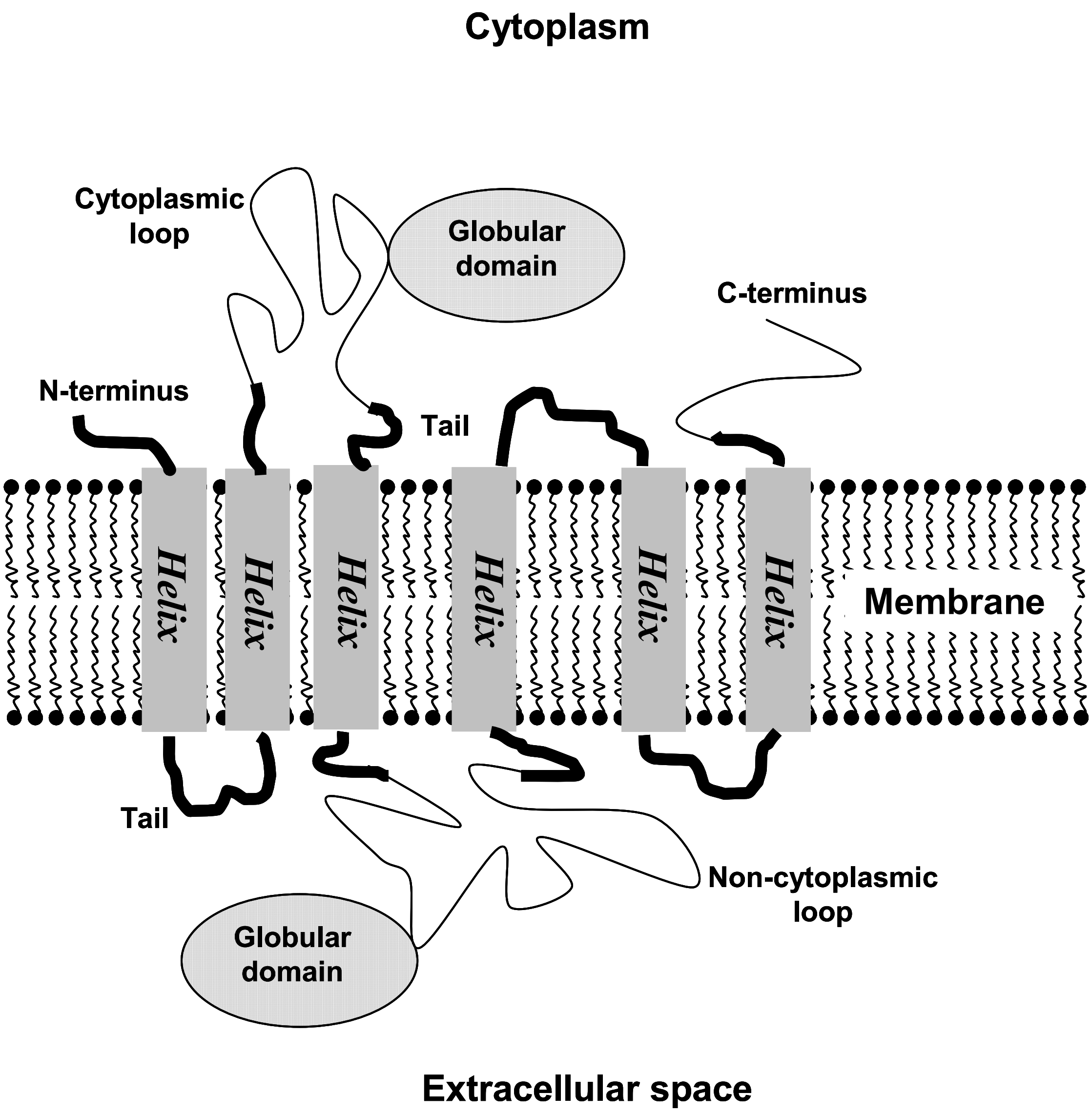


Figure 2.1 The detailed structural states of a typical helical transmembrane protein used in the hidden Markov model. A common topology of a transmembrane protein, which has N and C terminal tails in a cytoplasm side, is chosen as an example. Adapted from [104]. The main states include: cytoplasmic loop (thin line), non-cytoplasmic loop (thin line), tails (thick lines), membrane-spanning alpha helices (grey box), and globular domains (oval shape). Tails (caps) represent a part which interact directly with the membrane, and two tails between helices form a short loop. Longer loops are formed by tail-loop-tail sequences, in which a loop does not interact with the membrane.

Japanese), a prediction algorithm that is not dependant on HMM, is chosen for investigation because it is highly ranked in terms of TMH prediction [59]. SOUSI predicts transmembrane proteins based on four parameters: the hydropathy index (non-polar or hydrophobic residues) [78], an amphiphilicity index (polar or hydrophilic residues), an index of amino acid charges, and the length of each sequence. More accurate transmembrane protein prediction may be obtained by combining SOSUI with the three other algorithms in a multiconsensus fashion.

Transmembrane proteins predicted from a yeast proteome are first determined through optimization of the results from the four algorithms. To further estimate reliability of this dataset of transmembrane proteins, the grand average of hydropathy (GRAVY) scores [78] and the transmembrane proteins in common with the IMPs obtained from GO annotation are used as evaluation parameters. This dataset of the yeast transmembrane proteins is distinguished significantly from the soluble proteins and is used for the comparison of GeLC and GELFrEE analysis of the whole yeast proteome.

2.3 GELC *vs* GELFrEE MS/MS for Proteome Analysis

Considering the hydrophobicity of transmembrane proteins, sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is a widely used protocol for the separation of all proteins. All proteins can be dissolved in an aqueous solution containing SDS and separated based on molecular weight (MW) in SDS-PAGE. Separated proteins are cut from the gel and subjected to a digestion/extraction step using trypsin [61]. Extracted peptides can then be separated by RPLC and analyzed through electrospray ionization coupled with tandem mass spectrometry (LC-MS/MS) in a conventional ‘bottom-up’ approach. This method, known as GeLC, yields identification of a representative sampling of the proteome [57,60].

Recently, a similar method for proteome separation and analysis known as Gel Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE) has been developed [62]. GELFrEE relies on the same separation mechanism as SDS-PAGE, separating proteins based on MW; however, proteins are electroeluted from the gel and directly collected intact in the solution phase. An advantage of GELFrEE is that separated proteins can either be analyzed using top-down (intact), or bottom-up (digested) approaches. After separation with GELFrEE, SDS must be removed, usually through solvent precipitation using acetone [85] or chloroform/methanol/water (CMW) [84]. Removal of SDS is necessary due to interference with MS analysis. Protein pellets must then be re-dissolved in an MS-compatible buffer digested with trypsin, and analyzed by LC-MS/MS.

Given the similarities between GELFrEE and GeLC, a systematic comparison between the numbers of proteins identified through each method has been performed by Botelho *et al.* [105]. This report demonstrated that GELFrEE followed by MS/MS identified a similar number of yeast proteins and peptides, with similar distribution in terms of the protein MW. Given that GeLC is reported to be essentially unbiased to the whole proteome in terms of the abundance, size, charge, and hydrophobicity of identified proteins [57], this chapter aims to extend the comparison of the two platforms, looking specifically at the distribution of proteins according to hydrophobicity and develop a platform with the ability to characterize IMPs/transmembrane proteins. Proteins identified from three protocols employing GeLC, GELFrEE with acetone precipitation and GELFrEE with CMW precipitation will be assessed according to the number of IMPs and transmembrane proteins identified, as well as their hydrophobicity distribution as measured through GRAVY score established by Kyte and Doolittle [78]. The raw data for these comparisons were obtained from previously published experiments performed in the Doucette group [105].

2.4 Methods

2.4.1 The Theoretical Yeast Proteome

Protein sequences of all proteins (5,802) from the species *Saccharomyces cerevisiae* (herein referred simply as ‘yeast’) were downloaded from the UniProt website (http://www.uniprot.org/). Masses, pIs and GRAVY scores of the 5,802 proteins were obtained from freely accessible websites as described in Section 1.3.2 of Chapter 1.

The dataset of 1,832 yeast integral membrane proteins was predicted according to a Gene Ontology annotation term (GO: 0016020) through the UniProt website (www.uniprot.org/).

The number of alpha-helical transmembrane proteins for the yeast proteome was predicted *via* the four automatic algorithms: TMHMM2 [100], HMMTOP2 [102], and Phobius [103] and SOSUI [104]. These algorithms are all freely accessible web-based tools and their websites are shown in Table 2.1.

Detailed procedures for obtaining datasets of the total yeast proteins, IMPs and transmembrane proteins are described in the Appendix.

Table 2.1 The four selected transmembrane protein prediction algorithms.

Program	Website	Comments[(A)]
TMHMM2	http://www.cbs.dtu.dk/servies/TMHMM/	Very good at distinguishing soluble from membrane-integral proteins and TMH prediction
Phobius	http://ww.phobius.cgb.ki.se/	TMH prediction with N-terminal signal peptides distinguished from TMHs
HMMTOP2	http://www.enzim.hu/hmmtop/html/adv_submit.html	Good at distinguishing soluble from membrane-integral proteins
SOSUI	http://www.bp.nuap.nagoya-u.ac.jp/sousui	Very good at distinguishing soluble from membrane-integral proteins

[A] Comments are taken from the literature reviewed by Whitelegge *et al*. [59]

2.4.2 Data for GeLC and GELFrEE

The proteins and peptides identified by GeLC, GELFrEE with acetone precipitation, and GELFrEE with CMW precipitation have been published [105]. Permission was obtained from the authors, being members of our research laboratory, for use in the present work.

2.5 Results and Discussions

To evaluate the efficiency of GELFrEE MS/MS with acetone or CMW precipitation *vs* GeLC MS/MS for a comprehensive proteome analysis method, the datasets of yeast IMPs/transmembrane proteins must first be established. Computational approaches using bioinformatics (e.g., GO term, TMHMM) are the best way to obtain these datasets. Each platform can then be evaluated by comparing the number of identified IMPs/transmembrane proteins, and the number of TMHs in identified transmembrane proteins. As a comprehensive proteome analysis method, these two gel-based platforms are analyzed in terms of mass, isoelectric point (pI) and hydrophobicity (GRAVY score) and compared to the distributions of the entire yeast proteome. The relative difference (chi-square test) can be used as a criterion to evaluate the efficiency of proteome analysis of these two platforms.

2.5.1 Established Datasets of Integral Membrane & Transmembrane Proteins

A dataset of 1,832 yeast IMPs was established using GO annotation as described in Section 2.4.1.1, making the percentage of predicted IMPs in the yeast proteome 31.6% (1,832 out of 5,802). GO annotation not only incorporates genetic information (i.e., cellular component), but also includes experimental and computational prediction information [95] making this dataset reliable for evaluation of comprehensive proteome fractionation methods.

Because there is no GO term for transmembrane proteins, the dataset of yeast transmembrane proteins needs to be established through computational approaches. Four TMH prediction algorithms, TMHMM, SOSUI, Phobius and HMMTOP, were selected.

2.5.1.1 Comparison of the Four Algorithms

Based on the 5,802 protein sequences of yeast *S. cerevisiae*, the number of predicted transmembrane proteins using each prediction algorithm is summarized in Table 2.2. While the reliability of the four prediction methods are ranked at high level, disagreement between predicted numbers of transmembrane proteins among the four methods is clearly shown. As shown in Table 2.2, HMMTOP predicts a significantly greater number of proteins than the other three algorithms, in which similar numbers of transmembrane proteins are predicted. Based on the 20-25% theoretical percentage of transmembrane proteins (gene codes) [77], each of the three algorithms (TMHMM, 19.6%; SOSUI, 21.9% and Phobius, 22.8%) appears to yield reasonable transmembrane protein predictions, and HMMTOP (34.1%) needs to be further examined.

Discrepancy between the predicted transmembrane proteins among the algorithms is presented. Figure 2.2A shows many transmembrane proteins uniquely predicted by each algorithm. The accuracy of each prediction algorithm is not 100% (approximately 70%), meaning that an error of mis-prediction is inherent [72,75]. The prediction accuracy of algorithms are evaluated by the training sets, however, these sets are limited to small sizes because the transmembrane proteins with known structure are relatively

Table 2.2 Comparison of the four selected prediction algorithms.

Program	**# predicted TMPs[A]**	**# unique predicted TMPs**	**#TMPs common *vs* IMPs[B]**	**% overlap *vs* predicted TMPs**
TMHMM	**1,136**	**3**	**1,053**	**92.7**
SOSUI	**1,271**	**104**	**1,037**	**81.6**
Phobius	**1,325**	**84**	**1,121**	**84.6**
HMMTOP	**1,973**	**625**	**1,267**	**64.2**

[A] Transmembrane proteins (TMPs) represent predicted proteins containing α-helical transmembrane domain by each algorithm.

[B] The number of TMPs represents the overlap between the predicted TMPs from each algorithm and the 1,832 IMPs (predicted using the GO term 0016020).

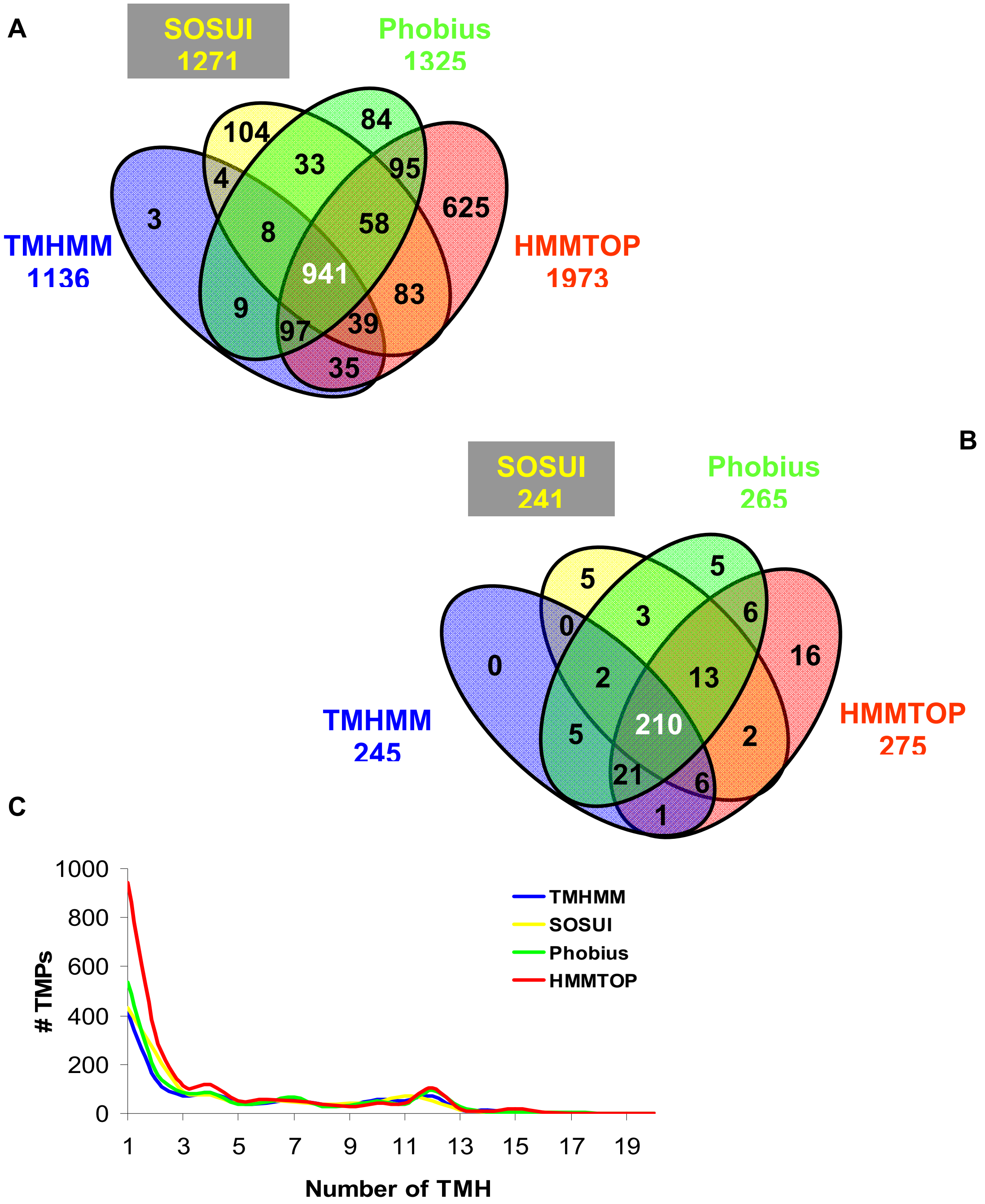


Figure 2.2 Comparison of the four prediction algorithms. (A) A Venn diagram showing total transmembrane proteins (TMPs) predicted from each of the four algorithms. Number of TMPs from each of the four algorithms as shown beside a label are different from each other. (B) A Venn diagram displaying the TMPs with TMHs ≥9 predicted by each algorithm. Number of these TMPs for each algorithm are close and the predicted TMPs are mostly overlapped with each other. (C) A line plot illustrating the distribution of the number of TMPs *vs* the number of TMH clearly shows the significant difference of TMP prediction by the four algorithms at TMPs with lower number of TMHs (i.e., one helix). The number of TMPs with TMHs higher ≥ 3 are close between each of the four prediction datasets.

few [72]. Additionally, a prediction algorithm is modified to have high accuracy of prediction for the training sets, in which the number and location of TMHs are predicted by aligning protein sequences to that of the training sets. Comparing amino acid similarity between the training set and the actual proteome, the results predicted by the algorithms are slightly biased toward proteins similar to the training sets.

In order to obtain a more reliable prediction algorithm, the overlapped proteins between predicted transmembrane proteins and the 1,832 GO annotated IMPs are considered. One would expect that all the predicted transmembrane proteins would completely overlap with the IMPs (GO annotated), but there are differences in the degree of overlap between each algorithm. Overlapping numbers of predicted transmembrane proteins for each of the four methods are 92.7% (TMHMM), 81.6% (SOSUI), 84.6% (Phoubis), and 64.2% (HMMTOP). Thus, it seems reasonable to expect that a more accurate prediction can be obtained from all the four algorithms. A total of 941 transmembrane proteins (16.2%) are consistently predicted by each algorithm, therefore one might simply take these proteins to represent the list of "confident" transmembrane proteins. However, such a list might under-represent the transmembrane proteins present in yeast (16.2% is less than 20-25%). Furthermore, as shown in Figure 2.2A, HMMTOP predicts the most unique transmembrane proteins (625), identifying 3, 104 and 84 unique transmembrane proteins from TMHMM, SOSUI, and Phobius, respectively. No protein structure data bank can confirm whether these uniquely predicted transmembrane proteins are true or false, so a reliable and simple strategy is needed to obtain the dataset of transmembrane proteins.

2.5.1.2 Comparison of Number of Predicted TMHs

An important aspect of the prediction algorithms is their ability to predict transmembrane topology, which provides a transmembrane protein with a number of TMHs. Thus, transmembrane proteins can be differentiated from soluble proteins using the prediction algorithms and the proteins containing number of predicted TMHs are used to assess a comprehensive protein analysis method. To obtain a reliable dataset of yeast transmembrane proteins, the four algorithms are compared according to the number of the transmembrane proteins obtained from each algorithm *vs* the number of predicted TMHs per protein (see Figure 2.2B). The majority of transmembrane proteins are predicted from only 1 or 2 TMHs, although some proteins contain up to 21. The greatest difference between prediction algorithms is noted when only a single TMH is predicted in a protein, accounting for the majority of differing numbers of transmembrane proteins predicted with each algorithm. Figure 2.2C shows that the predicted transmembrane proteins with high TMHs (≥9) are similarly predicted by each of the four algorithms. However, the number of TMHs predicted for a given protein will differ depending on the program used. Using HMMTOP, 35 proteins are predicted to have 3 to 6 of TMHs, but none of these proteins are seen in the three other predictions.

As suggested by Nilsson *et al*. [97], combining prediction results may provide a higher degree of accuracy than any single prediction method. Ameico *et al*. [106] designed a website to provide a variety of algorithms available, thus, users make their own judgment in deciding a consensus prediction. Based on this principle, a combination dataset of 2,218 transmembrane proteins generated from all four algorithms is more advanced than any one prediction algorithm. However, this number of transmembrane proteins (38.2%) may over-represent the actual number of transmembrane proteins in yeast (between 20 and 25%) [77], therefore a simple consensus approach is needed for rapid transmembrane protein prediction.

Each prediction algorithm is based on different principles. For example, TMHMM and HMMTOP are dependant on the HMM, but they assess different numbers of defined states of an integral membrane proteins (see Fig. 2.1) such as the five states for TMHMM and the seven states for HMMTOP. Although the Phobius algorithm is based on HMM as TMHMM and HMMTOP, it incorporates a signal peptide prediction model. The hydrophobic region of a signal peptide (SP) is similar to that of a transmembrane helix, but the region of a SP (approximately 7-15 residues) is shorter than that of a TMH (15-30 residues). Identification of these regions improves the accuracy of TMH prediction. In contrast, SOSUI is based on a different model as described in Section 2.2 from the HMM. As a consequence, a more confident dataset of the yeast transmembrane proteins, which is discriminated from soluble proteins or non-transmembrane proteins, can be generated from a consensus of all the results from the four algorithms. By averaging the number of TMHs in a given transmembrane protein and using only those proteins that contained greater than (or equal to) one TMH domain per protein, a dataset of 1,136 transmembrane proteins was quickly obtained. The new dataset improves discrimination between non-transmembrane proteins and transmembrane proteins by incorporating all parameters used in the four algorithms. The percentage of transmembrane proteins, 19.6% (1,136 out of 5,802), is also agreement with the 20% predicted helical transmembrane proteins from gene prediction in the literature [77].

2.5.1.3 Evaluation of the New Dataset of Transmembrane Proteins

The newly generated dataset of 1,136 yeast transmembrane proteins by a simple averaging strategy is further shown to be reliable by determining the number of predicted TMH domains per protein. Transmembrane proteins with single-spanning TMH domain have distinct properties from multi-spanning proteins due to the fact that all of the residues in the spanning regions are highly hydrophobic [100]. A great number of transmembrane proteins (i.e., small-molecule transporters) with higher number of TMH domains (six or seven) were observed in bacterial genomes [75]. Figure 2.3A shows that the distribution of the number of TMH domains is similar between the new dataset and that predicted from each of the four algorithms. The percentages of transmembrane proteins containing one-single TMH for this new dataset is slightly lower than that of the four predictions, meaning that these proteins are constrained in order to be more accurate through the average strategy. The percentages of the transmembrane proteins with a number of TMH domains higher than 3 for this new dataset are slightly higher than the numbers predicted from the other algorithms. This result is in agreement with the real biological system as mentioned above.

The reliability of the new dataset is further examined by an effect of length of a transmembrane protein. Figure 2.3B shows that predicted transmembrane proteins with length higher than 500 account for approximately 43% of total predicted transmembrane proteins in the four predictions. Although no literature reports prediction of transmembrane helices affected by size of proteins, proteins with large soluble domains (globular domains as seen in Fig. 2.1) tend to blur the “positive-inside rule” (arginine and lysine mainly found in cytoplasmic loops) used in the algorithms. While the

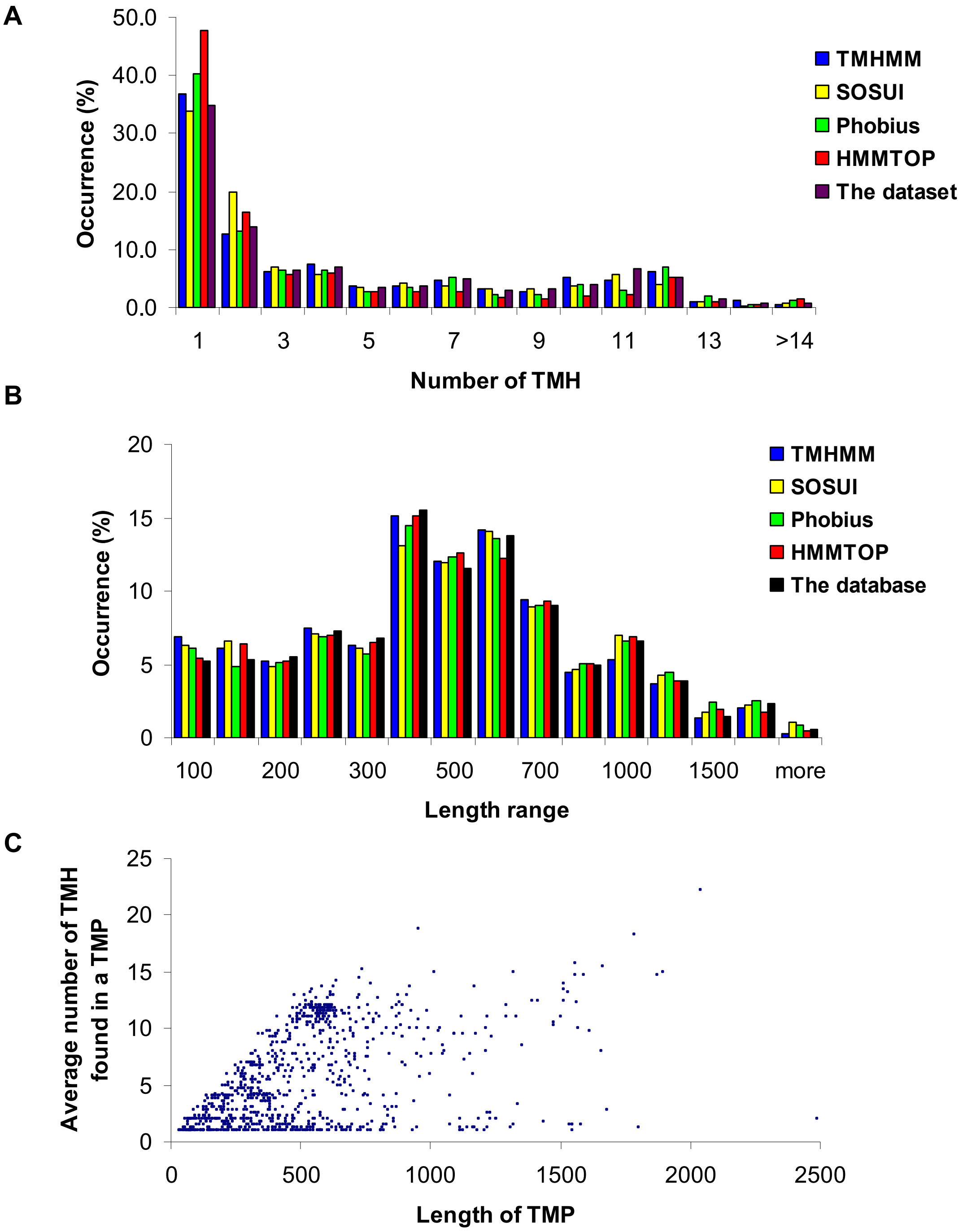


Figure 2.3 Evaluation of the new data set. (A) The distribution of the number of predicted TMH *vs* the % occurrence of total transmembrane proteins (TMPs) from the four prediction algorithms is compared to the new dataset of the TMPs established. (B) The distribution of length range vs the % occurrence of predicted TMPs from each prediction. (C) Relationship between the length and the average number of TMH found in a TMP within the dataset of 941 TMPs commonly predicted from the four algorithms.

transmembrane proteins are large, the globular domains in loops on the non-cytoplasmic side may contain a great number of positively charged residues [107], making accurate prediction of transmembrane proteins difficult.

The number of TMH domains is plotted against protein length in Figure 2.3C, where the dataset of 941 'reliable' transmembrane proteins commonly predicted by the four prediction algorithms methods was selected as a model. Figure 2.3C indicates that the number of predicted TMH domains partially corresponds to the overall size of the protein, where larger proteins are predicted to contain more TMH domains. This size effect is limited using the simple average strategy.

2.5.1.4 Test on Hydrophobicity Distribution

A hydrophobic scale (i.e. GRAVY) is useful to test this new dataset for accuracy. Hydropathy index assigns positive values to hydrophobic amino acids, therefore lending more positive GRAVY scores to transmembrane proteins over soluble proteins due to hydrophobic TMH domains. Comparing the scores for the newly generated dataset with those predicted from the four individual algorithms, the GRAVY distribution of each dataset is shown in Figure 2.4A. These datasets show a significant difference from them and the new dataset based on percentage distribution of transmembrane proteins to GRAVY range. Higher percentages of hydrophobic proteins (positive GRAVY scores) are predicted from this new dataset than any individual algorithm.

Interestingly, GRAVY scores of proteins do not shown correlation to the length of proteins, especially for soluble proteins [78]. Figure 2.4B confirms this result by plotting the relationship between length and GRAVY score of the dataset of the 941 transmembrane proteins. Large proteins do not represent hydrophobic proteins, in

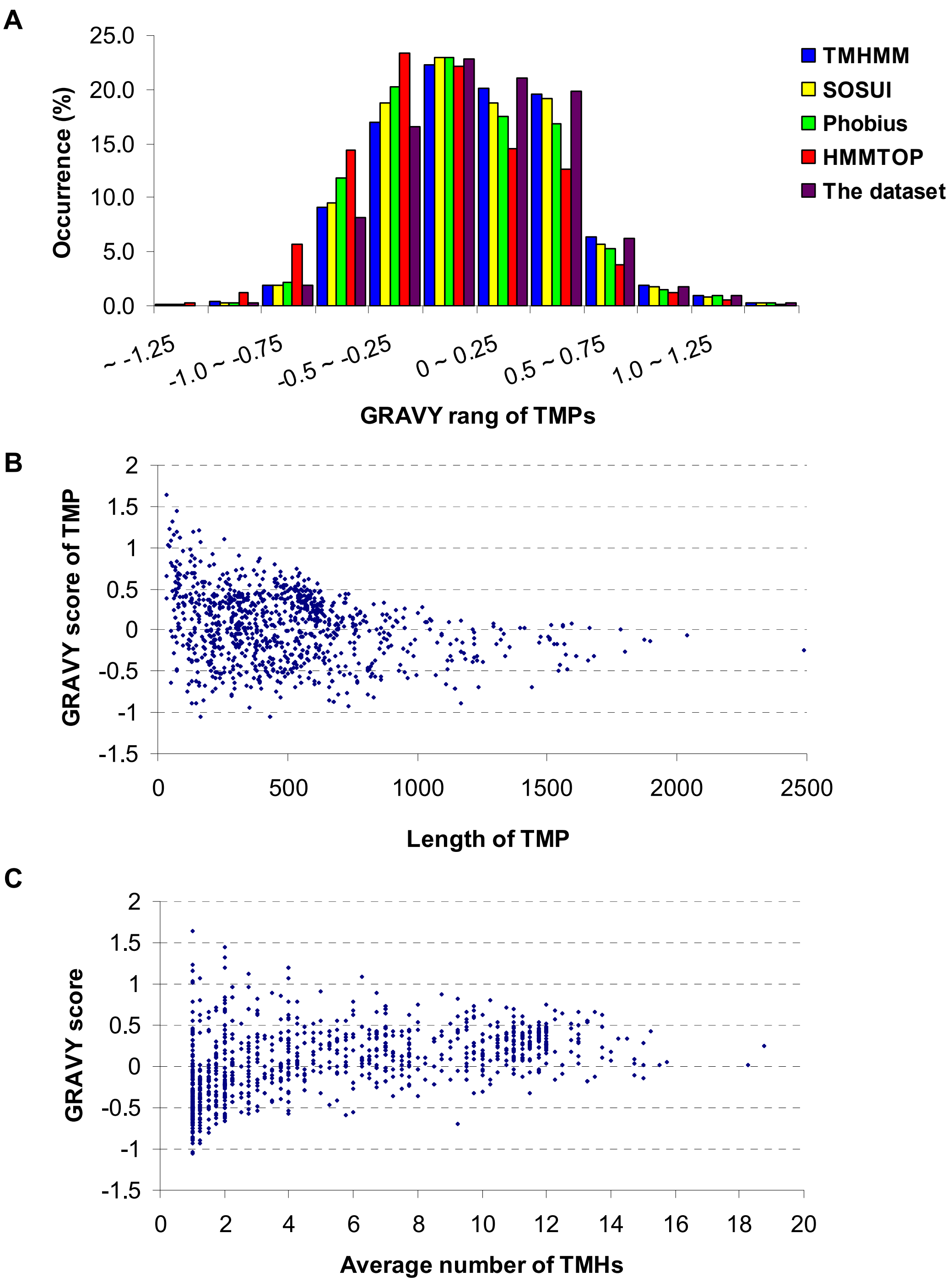


Figure 2.4 Test of the new dataset on GRAVY distribution. (A) The dataset of 1,136 transmembrane proteins (TMPs: black column) obtained from the averaging strategy is compared to the TMPs predicted from each of the four algorithms on the distribution of GRAVY range *vs* the % occurrence of total predicted TMPs. (B) Relationship of GRAVY score to length of a TMP within the dataset of the 941 TMPs. (C) Relationship of GRAVY score *vs* number of TMH of a TMP within the same dataset as (B).

particular, they have a tendency towards hydrophilicity. Figure 2.4B indicates that large transmembrane proteins tend to be hydrophilic. Positive GRAVY scores of proteins are mainly determined by their hydrophobic domains or number of hydrophobic amino acid residues (i.e., leucine, methionine) in a given protein. The GRAVY score of a transmembrane protein is affected by its number of TMHs. Figure 2.3C shows that large proteins have a potential to obtain high number of TMHs, hence, they may be to more hydrophobic than small proteins. Accuracy on large transmembrane protein prediction should take account for the size effect.

To evaluate the effect of number of TMH domains on hydrophobicity of a transmembrane protein, the GRAVY score to number of TMHs of a given transmembrane protein is plotted in Figure 2.4C. The partial correlation between GRAVY scores and number of TMHs is shown. Transmembrane proteins containing more TMH domains tend to be more hydrophobic. The high percentage of transmembrane proteins in positive GRAVY ranges implies more hydrophobic proteins to be predicted, meaning that the dataset of predicted transmembrane proteins is reliable. The average approach will decrease false prediction by concerning all effects through the four different prediction algorithms. Therefore, the new dataset of 1,136 transmembrane proteins is more reliable than the dataset predicted from each of the four algorithms due to slightly higher percentages of hydrophobic proteins (positive GRAVY scores). Using this newly generated dataset of 1,136 yeast transmembrane proteins as a reference, evaluation of proteome analysis platforms can be accomplished.

2.5.2 Comparison of GeLC and GELFrEE Platforms on Identified Proteins

2.5.2.1 Comparison of Identified Integral Membrane/Transmembrane Proteins

The two platforms or three methods (GeLC, GELFrEE/acetone and GELFrEE/CMW) were compared. The data for the three methods are summarized in Table 2.3. The numbers of identified total proteins/IMPs from GeLC (1,327/358) are similar to the numbers of identified proteins from the GELFrEE with acetone precipitation (1,320/ 356) and slightly higher than those identified from the GELFrEE with CMW precipitation (1,213/313). It is noted that the numbers of the total identified proteins from the three methods are much smaller than the yeast proteome (5,802). As shown in Table 2.3, the percentages of identified IMPs are similar between the GeLC (27.0%) and GELFrEE with Acetone (27.0%), but the percentage of identified IMPs from the GELFrEE with CMW (25.8%) is slightly lower than the other two methods. This indicates that the GELFrEE/acetone method is more comparable with the GeLC than the GELFrEE/CMW method. To evaluate whether the three methods are comprehensive or not, the percentages of the identified IMPs from the three methods (27.0%, 27.0%, and 25.8%) are compared to the theoretical percentage of IMPs (31.6% as shown in Section 2.5.1). On the basis of protein identification from the complex yeast proteome using these platforms, it is estimated that the three protein analysis methods are compatible to be used for whole proteome identification, but the GELFrEE/acetone approach is slightly better than GELFrEE/CMW for IMP identification.

Unlike IMP identification, a slight difference of identified transmembrane proteins is shown between the GeLC and GELFrEE/acetone methods (180 *vs* 166). The

Table 2.3 The numbers of proteins identified using three methods (GeLC, GELFrEE/Acetone, and GELFrEE/CMW).

Protocols	# Identified proteins (peptides)	# IMPs	# TMPs	% IMPs	% TMPs	# ID unique TMPs
GeLC	**1,327 (4,333)**	**358**	**180**	**27.0**	**13.6**	**47**
GELFrEE/Acetone	**1,320 (4,898)**	**356**	**166**	**27.0**	**12.6**	**25**
GELFrEE/CMW	**1,213 (4,502)**	**313**	**152**	**25.8**	**12.5**	**16**

Notes: TMPs represent alpha-helical transmembrane proteins.

slight difference between the GELFrEE/acetone and GELFrEE/CMW methods is mainly caused by differential protein recovery of the two precipitation methods, which is investigated in Chapter 3. The percentage of the identified transmembrane proteins (13.6% of total protein) from the GeLC is closer to the theoretical percentage of transmembrane proteins (19.6% as described in section 2.5.1.1) than that from the GELFrEE/acetone (12.6%) or the GELFrEE/CMW (12.5%). On the basis of these numbers, the GeLC is slightly biased toward proteome analysis under the conditions used for peptide identification (i.e., LC-MS/MS). GeLC is a comprehensive proteome analysis method, which is confirmed by Schirle *et al.* [56] and de Godoy *et al*. [57]. The GELFrEE platform has no significant difference from the GeLC while considering variance of instruments (i.e., LC-MS/MS) and sample handling (i.e., protein precipitation). As a result, the GELFrEE platform is comparable to GeLC and has the potential to be compatible for proteome analysis.

2.5.2.2 GRAVY, Mass & pI Distributions

Whether the two platforms under the current conditions employed are proper to be a comprehensive proteome analysis method or not was further examined using GRAVY distribution. Relative to the yeast theoretical proteome/IMPs/transmembrane proteins, the percentages of identified total proteins/IMPs/transmembrane proteins from the three methods are distributed against the GRAVY range. As shown in Figure 2.5, the distribution of the percentages of the three types of proteins (i.e., total proteins, transmembrane proteins) identified from the three methods to GRAVY range is indeed very similar, but a difference between the experimental and the theoretical results is presented.

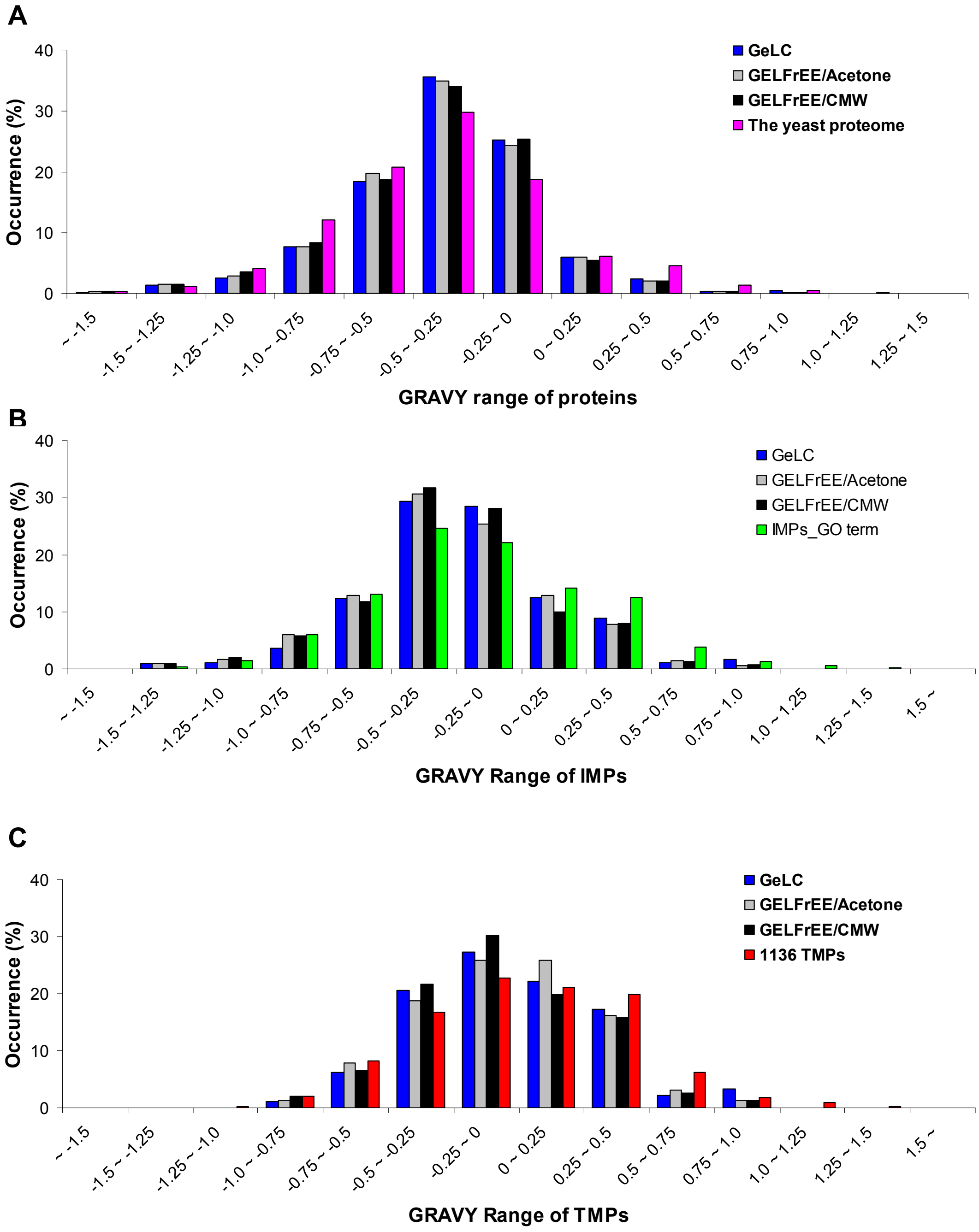


Figure 2.5 GRAVY distributions of the % occurrence of identified proteins from the three methods. (A) Totally identified proteins; (B) integral membrane proteins; and (C) transmembrane proteins (TMPs). The experimental results are compared to the theoretical ones: the yeast proteome (pink), the dataset of the 1,832 IMPs (green), and the dataset of 1,136 TMPs (red). The GRAVY distribution of identified TMPs shows a difference between GeLC and GELFrEE.

The GRAVY distribution of the totally identified proteins is not significantly different within the three methods, however, a difference of this distribution is shown between the experimental (three methods) and theoretical distributions, meaning that under the current conditions of LC-MS/MS, both of the platforms are biased toward hydrophilic protein identification. Interestingly, the GRAVY distribution of the identified transmembrane proteins from GELFrEE/acetone is more similar to this theoretical distribution than GeLC, using chi-square (χ^2) as a criterion, GeLC has a relatively better ability to identify transmembrane proteins than GELFrEE/acetone. The comparisons of the two gel-based platforms are further investigated on mass and pI distributions.

The mass and pI distributions of totally identified proteins from the three methods are compared using the same χ^2 test as the GRAVY distribution. The mass distribution of the identified proteins from the GELFrEE/acetone is still closer to the theoretical distribution than is GeLC. However, on the pI distribution, GeLC follows more closely to the theoretical distribution than GELFrEE. Further comparison of GELFrEE and GeLC for proteome analysis is studied.

2.5.2.3 Number of TMH Domain Distribution

To test the bias degree of hydrophobic protein identification for the two platforms, the number of TMH domains distributed throughout identified transmembrane proteins from the three methods is plotted in Figure 2.6. A similar trend of this distribution is shown between the experimental and the theoretical result (the dataset of 1,136 transmembrane proteins), but a difference between them is observed. Using a χ^2 test, the number of TMH domains distributed throughout identified transmembrane proteins from the GELFrEE/acetone (16.5) is more similar to the theoretical distribution than that from

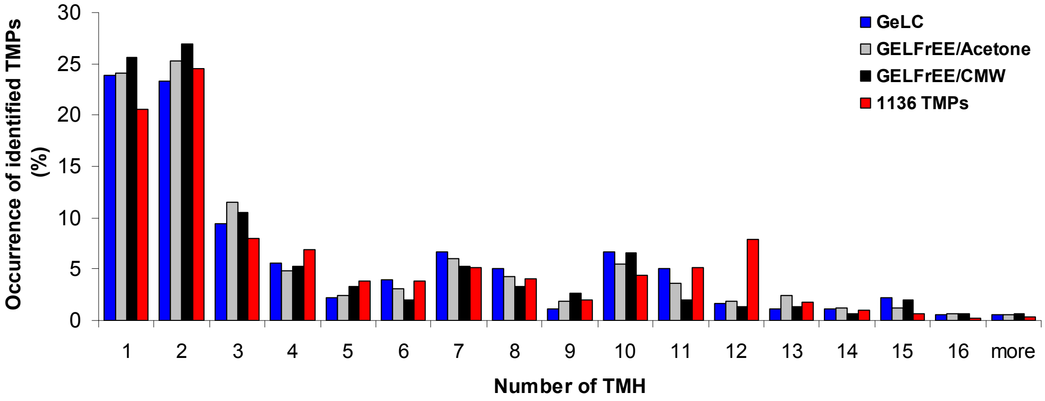


Figure 2.6 Comparison of the three methods on the identified transmembrane proteins. The distribution of the number of TMH domains versus the % occurrence of the identified TMPs from GeLC, GELFrEE/Acetone, and GELFrEE/CMW methods are compared, and also compared with the new dataset of 1,136 TMPs obtained from the average approach.

the GeLC (18.5) and GELFrEE/CMW (22.5). The χ^2 of GELFrEE/CMW is higher than GELFrEE/acetone and lower than GeLC. Although numbers of IMP/transmembrane proteins identified from GELFrEE/CMW are fewer than that from GeLC and GELFrEE/acetone, GELFrEE/CMW seems to prefer very hydrophobic protein identification. This result indicates that the two platforms have their own merits while they are applied in a complex sample analysis.

2.5.3 Comparison GeLC *vs* GELFrEE Platforms on Identified Peptides

The difference between the experimental and the theoretical numbers are likely due to limitations of LC-MS/MS analysis. Ionization efficiency and peptide fragmentation of hydrophobic peptides are less efficient than hydrophilic peptides, limiting the scanning capabilities of MS and separation efficiency of RPμLC [29,108]. Avoiding these limitations will result in high percentages of identified integral membrane/transmembrane proteins. Therefore, the two platforms have the potential to be totally unbiased proteome analysis approaches.

Referring to Figure 2.7A, approximately 50% of all identified peptides are common from all three methods. By contrast, 1,585, 1,179, and 584 unique peptides were detected from the GeLC, GELFrEE/acetone and GELFrEE/CMW methods, respectively. Investigation of the three methods was performed in terms of length and GRAVY score on total identified peptides. Neither the length distribution (data not shown), nor the GRAVY distributions of the percentage of identified peptides (see Fig. 2.7B) indicate differences between the three methods. The slight variance of transmembrane protein identification between the three methods as mentioned in the previous section is not caused by the conditions of LC-MS/MS.

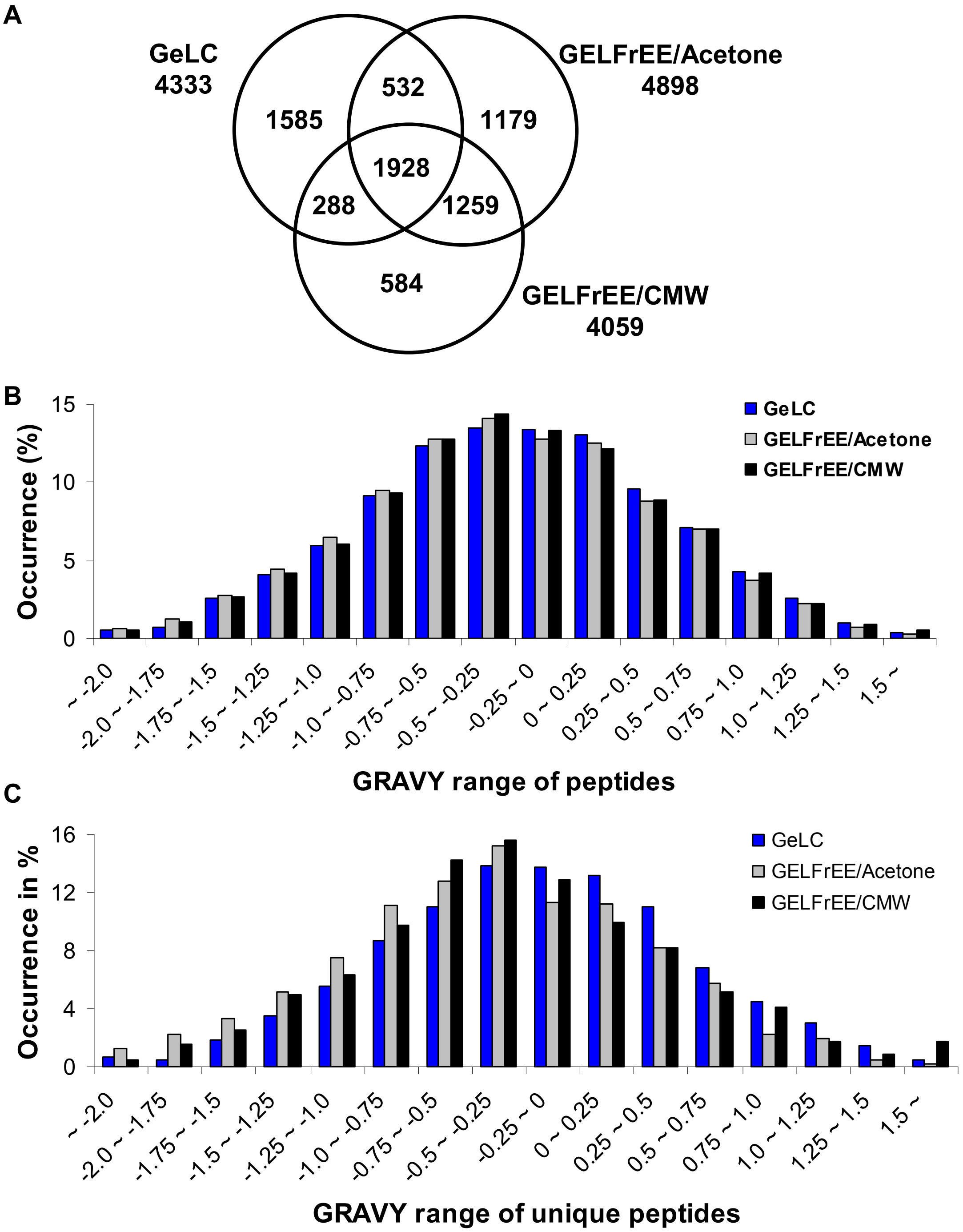


Figure 2.7 Comparison of identified peptides from the three methods. (A) A Venn diagram showing overlap of the peptides identified from each of the three methods (GeLC, GELFrEE/acetone, and GELFrEE/CMW). The number of the total identified peptides is shown along with the label (e.g., GeLC 4333). The GRAVY distributions of the % occurrence of totally identified peptides (B) and uniquely identified peptides (C) from the three methods present a similar trend.

Minor differences in the length (data not shown) and GRAVY distributions of peptides uniquely identified (see Fig. 2.7C) from the three methods indicate that this variance within the three methods is presumably due to the method itself (i.e. digestion conditions). The GeLC is slightly preferred toward hydrophobic peptide identification, thus, the percentages of identified integral membrane/transmembrane proteins for the GeLC are slightly higher than the two GELFrEE methods.

A detailed comparison between the two platforms at either the protein or peptide identification shows that the two platforms are comparable, and also capable to be used as a comprehensive proteome analysis while the conditions (i.e., peptide separation) of LC-MS/MS are improved. The benefit of the GELFrEE MS/MS (solution-based digestion) platform is compatible with top-down approaches (intact protein identification).

2.6 Conclusions

Following a multiconsensus method that employs four highly ranked helical transmembrane protein prediction algorithms, a reliable dataset of transmembrane proteins has now been established and proven to be reliable for the yeast proteome. Based on this protein dataset, the GELFrEE platform (GELFrEE with acetone or CMW) has been shown to be comparable in performance as the GeLC platform. The distribution of identified proteins for each method by mass and pI shows no significant difference when compared to the whole yeast proteome. When these two platforms are applied to a complex yeast proteome, they are shown to be slightly biased toward soluble protein identification.

The numbers of identified peptides are very similar between the GeLC and GELFrEE protocols. Also, the length and GRAVY distribution patterns of identified peptides are similar for the three methods. Furthermore, the two precipitation methods have little difference on protein identification and both are slightly biased for transmembrane protein precipitation. Under the conditions of current peptide identification, all of the results demonstrate that the two platforms are comparable with each other, but slightly biased on the full proteome analysis.

The distribution of mass, length, and hydrophobicity of the identified peptides are limited by the conditions of LC-MS/MS. By overcoming the limitations of LC-MS/MS or coupling one more dimensional separation to the two protocols, numbers of identified transmembrane proteins or the percentage of them may increase. Overall, the GELFrEE platform has high potential to be used for the whole proteome analysis.

Chapter 3

Optimization of Proteome Extraction Methods for Solution-Based Proteome Separation and Analysis

3.1 Introduction

The objective of the work presented in this chapter is to determine the most effective method to extract and solubilize the complete proteome (including transmembrane proteins) from biological samples for further separation and analysis using mass spectrometry (MS). The combined pairing of protein separation with MS detection can be considered the end stage in a lengthy series of manipulations which bring a "raw" biological sample into a concise set of data depicting the protein content of a sample. Thus, for separation and MS analysis to be successful, proper sample preparation (ahead of instrumental analysis) is crucial to the experiment [109].

For proper sample preparation, the greatest concern is that all proteins must first be extracted from cells. Proteins must be brought into solution in an unbiased fashion, meaning that all classes of protein (e.g., hydrophilic/ hydrophobic) are fully solubilized. Any loss of protein in this first step will lead to incomplete profiling of the proteome. Once extracted, proteins must be purified to remove non-protein contaminants which may interfere with subsequent sample manipulation or analysis. These contaminants may consist of naturally occurring material such as lipids, nucleic acids, carbohydrates, and salts. Other contaminants include additives such as detergents which assist in cell disruption, maintain protein solubilization, or inhibit enzyme activity. Solid phase extraction, dialysis and protein precipitation represent the most popular approaches for protein purification. Maintaining high and unbiased protein yield throughout the purification protocol is an important consideration for comprehensive proteome analysis.

Sodium dodecyl sulfate (SDS) is a powerful solubilizing detergent which is routinely used in biochemical studies [58,110]. However, SDS, even at low concentrations, can interrupt enzymatic digestion [111] and suppress ion signals in MS [85,112]. In practice, SDS is widely used for intact protein separation using one dimensional (1D) SDS polyacryamide gel electrophoresis (PAGE) [56]. The coupling of gel-based separation platforms to mass spectrometry (GeLC MS/MS [56,57] or Gel Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE) MS/MS [62]) have been successful for bottom-up MS approaches, as demonstrated in Chapter 2. Although the gel-based platforms can be used to obtain an unbiased profile of a proteome, they are fairly laborious and difficult to automate. Therefore, alternative methods (gel-free or solution-based protein separation methods), which do not rely on SDS, yet still maintain membrane protein solubility, would be beneficial for MS-based protein analysis.

Liquid chromatography (LC) is a popular approach to separate proteins [11,64,113]. This solution-based platform permits a facile method for high-throughput analysis or automation of instruments. Reversed phase (RP) LC in particular provides high resolution for separation of peptides, and is also easy to couple to MS through electrospray ionization (i.e., LC/MS) [86]. RPLC is also used for protein-level separation [107,114,115], though with greater difficulty than peptide-level separation. Here, one must consider the greater range of hydrophobicity encompassing intact proteins. In particular, the solubility of alpha-helical transmembrane proteins or hydrophobic proteins is difficult to maintain in RPLC-compatible solvents [4,116]. It has also been noted that high molecular weight proteins have poor recovery in RPLC [114,117]. These challenges will be addressed in Chapter 4, in which a solvent system (60% formic acid in water (*v*/*v*) and isopropanol) is used for protein separation [65]. The goal of this chapter is to devise a strategy for the isolation of total proteins from cellular systems ahead of RPLC separation. These front-end manipulations are considered a vital component related to the development of a comprehensive solution-based platform for proteome analysis.

For successful proteome separation through RPLC, the front-end sample manipulation methods can be characterized as follows: (1) proteome extraction, wherein all the proteins (including hydrophobic proteins) are isolated from a cellular system and brought into solution phase through cell disruption, (2) protein precipitation, which is used to remove interfering compounds (salts, lipids, etc.) and additives (such as SDS) from the extracted proteins, and (3) protein resolubilization, wherein the purified and enriched proteins or precipitated proteins must be brought back into solution. The solvent system used to resolubilize all the proteins should be fully compatible with RPLC.

Yeast (*Saccharomyces cerevisiae*) was selected as a model system to test the extraction, precipitation and resolubilization protocols investigated in this chapter, with emphasis on maintaining high recovery for the IMP/transmembrane protein fraction.

3.2 Experimental

3.2.1 Materials

Formic acid, trifluoroacetic acid (TFA), ammonium perfluorooctanoate (APFO), ammonium bicarbonate, iodoacetamide, bovine serum albumin (BSA), bovine trypsin (TPCK treated, cat. T8802), protease inhibitor cocktail (cat. P2714), and yeast extract-peptone-dextrose (YPD) were purchased from Sigma (Oakville, Ontario). Chloroform, acetone, HPLC (MS)-grade acetonitrile, HPLC-grade methanol, and acetic acid were purchased from Fisher Scientific (Ottawa, Ontario). SDS, urea, Tris, dithiothreitol (DTT), and other reagents used for casting and running SDS-PAGE were obtained from Bio-Rad (Mississauga, Ontario). All other chemicals were obtained from Sigma and were used without further purification. Milli-Q grade water was purified to 18.2 MΩ/cm.

3.2.2 Yeast Strains

Fresh yeast cells of *Saccharomyces cerevisiae* (strain BY4741) streaked on an agar plate were generously supplied by Dr. Melanie Dobson (Department of Biochemistry & Molecular Biology, Dalhousie University). Two isolated colonies from the plate were scraped using a sterile inoculation loop and added to 10 mL of autoclave-sterilized YPD medium (Sterilmatic Autoclave, Market Forge). The solution was vigorously shaken to disperse the yeast cells, followed by incubation at 28ºC with shaking at 120 rpm on an orbital shaker (Environmental Shaker, Queue) for 16-18 hrs. Two milliliters of this suspension were transferred to 500 mL of sterile YPD medium (contained in 1L flask), with 4 flasks inoculated at one time. The cells were grown to mid-log phase (OD_{600} 0.6-0.7) at 28ºC with agitation (120 rpm) in the orbital shaker (approximately 7 hours). Harvested cells were collected by centrifugation at 3500×g (Sorvall RC-5B Refrigerated Superspeed Centrifuge, Du Pont Instruments) for 15 min at 4ºC, and then washed three times with ice-cold sterile water. The instruments for cell culturing were generously supported by Dr. Robert L. White (Department of Chemistry, Dalhousie University). The collected cells were divided into 8 aliquots and stored overnight at -20ºC. Cells were lysed using different lysis solutions, as described in Section 3.2.3.

Harvested yeast cells (strain MT302/28B) were provided by Dr. Lois Murray (Department of Biochemistry & Molecular Biology, Dalhousie University). These cells were similarly grown to mid-log phase at 30ºC and aliquots were stored at -20ºC until lysis.

Harvested yeast cells (strain BY4741) were generously supplied from Dr. Christopher McMaster (Department of Biochemistry & Molecular Biology, Dalhousie University), grown to mid-log phase (OD_{600} 0.5) at 30ºC. The cell aliquots were stored at -20ºC until lysis.

3.2.3 Cell Lysis

Aliquots of yeast strain BY4741 (from Dr. Dobson) were lysed by either French press at 20,000 psi (three passes) or by boiling for 10 minutes. In each case, 250 µl of protease inhibitor cocktail (general use), combined with 3.5 mL of a lysis solution was added to the yeast cells prior to disruption. For French press, cells were suspended in one of five lysis solutions: (1) 50 mM Tris-HCl pH 7.5, (2) 50 mM Tris buffer with 8 M urea at pH 7.5, (3) 50 mM Tris buffer with 4% SDS at pH 7.5, (4) 50 mM Tris buffer with 2% APFO at pH 7.5, or (5) 60% formic acid in water. For boiling, cells were suspended in solutions (3), (4), or (5) as described above. Following cell disruption, the lysate was subject to a low speed centrifugation at 3200×g for 15 min at 4ºC to remove unbroken cells. A BCA protein assay kit (Pierce) was used to measure protein concentration in the extracts according to the manufacturer's instructions.

For testing protein resolubilization in different solvent systems (Section 3.3.1, Fig 3.1 & 3.2), proteins extracted from yeast strains BY4741 (from Dr. McMaster) and yeast strain MT302/28B (from Dr. Murray) were used. Both of the strains were suspended in 25mM Tris-HCl pH 7.65 and subjected to French press with three passes at 20,000 psi. Unbroken cells were removed from the lysate by centrifugation as described above. This extraction procedure was facilitated with the help of Mr. Ken Chisholm (NRC-IMB Halifax, NS). Bio-Rad Bradford and Dc protein assay kits (Bio-Rad) were used to measure protein concentration in the two extracts according to the manufacturer's instructions.

3.2.4 Preparation of Subcellular Fractions

The extract of yeast strain MT302/28B lysed using a French press in 25 mM Tris-HCl pH 7.5 was ultra-centrifuged at 103,000×*g* (Optima L-90 K Ultracentrifuge, Beckman Coulter) for 65 min at 4ºC with the help of Mr. Elden Rowland (Proteomics and Mass Spectrometry Center, Halifax, NS). The resulting supernatant was collected as the cytosolic protein fraction (soluble proteins) and the pellets were saved as the membrane protein fraction. The protein concentration of the supernatant was measured to be 3 μg/μL using a Bio-Rad Bradford protein assay.

3.2.5 Protein Precipitation Methods

3.2.5.1 Cold Acetone

This procedure was adapted from the acetone protein precipitation protocol described by Lemaire *et al.* [83]. Briefly, 200 μL of cold acetone (-20ºC) was mixed with 50 μL of lysate (4:1 (*v:v*) ratio) then incubated overnight (~16 hrs) at -20ºC. The lysate was centrifuged for 15 min at 20,000×*g* (accuSpin Micro, Fisher Scientific) and the supernatant discarded. If SDS was added in the lysate, pelleted proteins were washed an additional two times through addition of 200 µL of cold acetone (-20ºC). The pellets were air dried in the fume hood to remove residual acetone.

3.2.5.2 Chloroform/Methanol/Water

This procedure was adapted from the protocol described by Wessel *et al.* [84]. A final 1:4:3 (*v*:*v*:*v*) ratio of chloroform:methanol:water was used. To do so, 200 µL of methanol was mixed with 50 µL of lysate, and then 50 µL of chloroform was added to the mixture. After brief vortexing, 150 µL of deionized water was added to the solution, resulting in a phase separation. Following gentle shaking, the sample was centrifuged for 15 min at 20,000×*g* (accuSpin Micro, Fisher Scientific), and the top layer (water/methanol) was removed, being careful not to disrupt the protein pellet at the interface of the two solvent layers. A 200 µL portion of methanol was then added with gentle shaking to form a single miscible solution. The sample was centrifuged a further 15 min at 20,000×*g* to pellet the protein precipitate and the supernatant was decanted. The protein pellets were finally washed with 200 µL of methanol, and then air dried in the fumehood to remove residual methanol.

3.2.6 SDS-PAGE

Following protein extraction, 50 µL of each lysate was dried in a SpeedVac (Savant). A Laemmli buffer system [118], comprising 50 mM Tris pH 6.8, 2% SDS, 5% Glycerol, and 0.002% bromophenol blue, was added to each dried pellet to make the final protein concentration 2.5 µg/µL. With brief vortexing, the mixed solutions were heated at 95ºC for 5 min, briefly centrifuged at 10,000×*g* (accuSpin, Fisher Scientific), and 10 µg of yeast proteins for each sample was loaded into a 1 mm well, 15% T (2.67%C) SDS-

PAGE gel and run at 240 V for ~45 min. The procedure for running SDS-PAGE was obtained from the instructions for the Mini-PROTEAN 3 (Bio-Rad). Protein bands were visualized using Coomassie staining [52] according to the instruction of Coomassie blue (G-250 stain) or silver stained [53] according to standard protocols. For protein identification through mass spectrometry (Section 3.2.9) following in-gel digestion (Section 3.2.7), 50 µg of sample was loaded into SDS-PAGE and stained using Coomassie for visualization before digestion.

3.2.7 In-gel Digestion

Following SDS-PAGE separation and visualization, each lane was cut into 15 gel slices and subjected to in-gel digestion as described by Mann *et al.* [61]. Peptide extraction was performed twice to insure maximum recovery and all extracts were pooled and dried to completion using a SpeedVac. Samples were stored at -20ºC until cleanup.

3.2.8 Sample Cleanup

Prior to MS, all samples are subject to cleanup using RPLC on an Agilent 1200 HPLC system with UV detection at 214 nm. The 1×50 mm RP column was packed in-house with Waters (Milford, MA, USA) Spherisorb S5ODS2 C_{18} beads (5 µm, 80 Å pore size). Water with 0.1% (*v*/*v*) TFA (Solvent A) and acetonitrile with 0.1% (*v*/*v*) TFA (Solvent B) were used as the solvent system at a flow rate of 100 µL/min. In all cases, dried peptide samples were redissolved in 100 µL of 5% solvent B and 95 µL was injected onto the column. The solvent was initially set at 5% B, and held for 10 min following sample injection. A linear gradient was set as follows: 5-95% B over 9 min, 95% B held over 3 min, 95-5% B in 1 min. This gradient was repeated a second time to thoroughly clean the column prior to a subsequent sample injection. Eluted peptides were collected over an 8 min period from time 20 to 28 min and divided into three aliquots. All aliquots were dried to completeness in the SpeedVac and stored at -20 ºC until LC-MS/MS analysis.

3.2.9 LC-MS/MS Analysis

3.2.9.1 Conditions for LC-MS/MS

An Agilent (Palo Alto, CA, USA) 1200 LC nanopump/autosampler was coupled to a nanospray ionization source (2.5 kV) and MS data was collected using a linear ion trap (LTQ) mass spectrometer (ThermoFisher, San Jose, CA, USA). All fractions obtained from the in-gel digestion were subjected to LC-MS/MS analysis with duplicate runs. Each sample was re-dissolved in 12.5 µL of 5% acetonitrile/water with 0.1% formic acid and 10 µL of the solution was loaded onto a frit (New Objective, Woburn, MA, USA) capillary column (75 µm i.d. with a 10 µm spray tip) packed with 25 cm of 3 µm Phenomenex C_{18} beads (Torrance, CA, USA). For LC-MS/MS analysis, Solvent A was water with 0.1% (*v*/*v*) formic acid and solvent B was acetonitrile with 0.1% (*v*/*v*) formic acid. The column was equilibrated with 5% B for 30 min prior to analysis. The gradient was set as follows: 5%B, 0 min; 10% B at 0.1 min; 30%B, 100 min (0.2% per min); 80% B, 105 min (5% per min); 5% B, 105.1 min, and held for 13 min to re-equilibrate the column. The flow rate was 0.25 µL/min.

3.2.9.2 Conditions for Recording MS/MS Spectra

A "tripleplay" dependant scan was used to record MS spectra. A full MS scan (400-1700 *m/z*) was used to select the three most intense doubly or triply charged ions for MS/MS analysis. The ion's charge state was determined *via* a high resolution "zoom scan" described in Chapter 1. Selected ions were subjected to MS/MS fragmentation through collision induced dissociation (CID). To validate the quality of data obtained from the LC-MS/MS experiment, 50 fmol of digested BSA was analyzed by LC-MS/MS, requiring at least 40% amino acid sequence coverage prior to proceeding with sample analysis.

3.2.10 Database Searching

MS/MS spectra were searched against a yeast *S. cerevisiae* database containing 5,802 unique protein entries, using the Bioworks 3.2 software package (ThermoFisher), which uses the SEQUEST search engine. The following search parameters were used to identify proteins: precursor ion tolerance 1.5 amu and fragment ion mass tolerance 1.0 amu; tryptic peptides with no more than two missed cleavages; chemical modifications including fixed carboxyamidomethylation at cysteine, differential oxidation at methionine, and differential carbamylation at lysine or arginine for the proteins extracted from the buffer solution containing 8 M urea.

The protein search results were filtered according to the following settings: Xcorr >2.2 (+2 ions) and >3.75 (+3 ions), ΔCn>0.1, and Rp <4. Duplicate peptide hits were also removed. The false positive rate, which was calculated according to the number of peptides identified in a reverse yeast (decoy) database search [44], was controlled to be less than 1% by adjusting the peptide probability score to 10^{-4}.

3.3 Results and Discussion

Our workflow for solution-based proteome separation and MS analysis can be described in three stages: (1) sample preparation (ahead of LC separation); (2) reversed phase proteome fractionation (subject of Chapter 4); (3) post-column coupling and MS analysis of fractions (subject of Chapter 4). Since all classes of proteins must be recovered, each stage of sample preparation, being dependent on those preceding it, must be properly optimized. Considering now the first stage of the workflow, sample preparation generally entails the following: (1) cell disruption and proteome extraction; (2) protein cleanup; and (3) final resolubilization in LC-compatible solvents. Using yeast (*S. cerevisiae)* as a model, sample preparation ahead of solution-phase proteome fractionation is herein evaluated. Proteome profiling is conducted with SDS-PAGE along with LC-MS/MS and the percentage of integral membrane/transmembrane proteins identified is used to evaluate the sample preparation procedure.

3.3.1 Comparison of Proteome Extraction using SDS-PAGE

Two lysis methods, French press [80] and simple boiling [110,119], were selected to disrupt the yeast cells. The French press was chosen over other lysis methods such as sonication/beads [120] and freeze-thaw [80], in that French press is one of the more effective approaches to disrupt cells, while boiling, if effective, would represent a simple and rapid lysis method.

Based on the data summarized in Table 3.1, one observes that the extraction efficiency of the simple boiling protocol is close to that of the more laborious French press protocol. Three lysis solvents, 50 mM Tris-HCl with either 4% SDS or 2% APFO, and 60% formic acid/water, were selected for comparison of the two lysis methods. The highest concentration of extracted protein was achieved with French press, using the Tris HCl solvent system (3.9 μg/μL). Although on average the boiling method yielded lower concentrations of extracted protein than French Press extraction in the same lysis

Table 3.1 Comparison of proteome extraction efficiency of five lysis solutions and two lysis methods [A]

Lysis method	Tris/water (µg/µL)	Tris/urea (µg/µL)	Tris/SDS (µg/µL)	Tris/APFO (µg/µL)	60% FA (µg/µL)
French press	**3.9**	**3.2**	**2.9**	**3.3**	**3.1**
Boiling	**N**	**N**	**1.9**	**3.1**	**1.8**

[A] Comparison is based on protein concentration of an extract measured using a BCA protein assay. Five lysis solutions are described in Section 3.2.2.
N indicates no experimental data is available.

solutions, it still provides an acceptable protein yield for MS analysis. Optimal lysis through boiling occurred using a Tris-HCl/APFO solution as extraction buffer, wherein the protein concentration from boiling was 3.1 μg/μL, which is in the range of protein concentration of the five extracts using French press (from 2.9 to 3.9 μg/μL). Compared with French press, this rapid and simple lysis method is useful to manipulate a large number of samples simultaneously. The French press approach requires ~30 min per sample, whereas boiling is complete in 10 min, regardless of the number of samples. Another benefit of boiling is that sample preparation does not require an expensive instrument for cell lysis, as is the case with the French press. Also, low volumes of sample can be lysed by boiling, whereas French press requires ~ 1 mL. Finally, sample loss is avoided with boiling, as some sample inevitably remains in the French press cell.

As is the goal of this study, the complete (unbiased) extraction of all proteins in a proteome has yet to be assessed. The total protein concentration in the extract does not provide any information on the nature of the extracted proteins (i.e. the percentage of hydrophilic or hydrophobic, high or low mass, acidic or basic proteins). Therefore, to further compare the two lysis methods on the basis of the distribution of extracted proteins, SDS-PAGE with Coomassie staining was used to assess possible differences between the two extraction protocols.

Figure 3.1 demonstrates that boiling is in fact not as efficient as French press for complete proteome extraction by showing greater numbers of protein bands from the French press samples than boiling. The effect is especially apparent in the high mass region, demonstrated by the lack of protein bands >100 kDa in boiled samples. The trend of poor recovery for large molecular weight (MW) proteins is consistently observed in

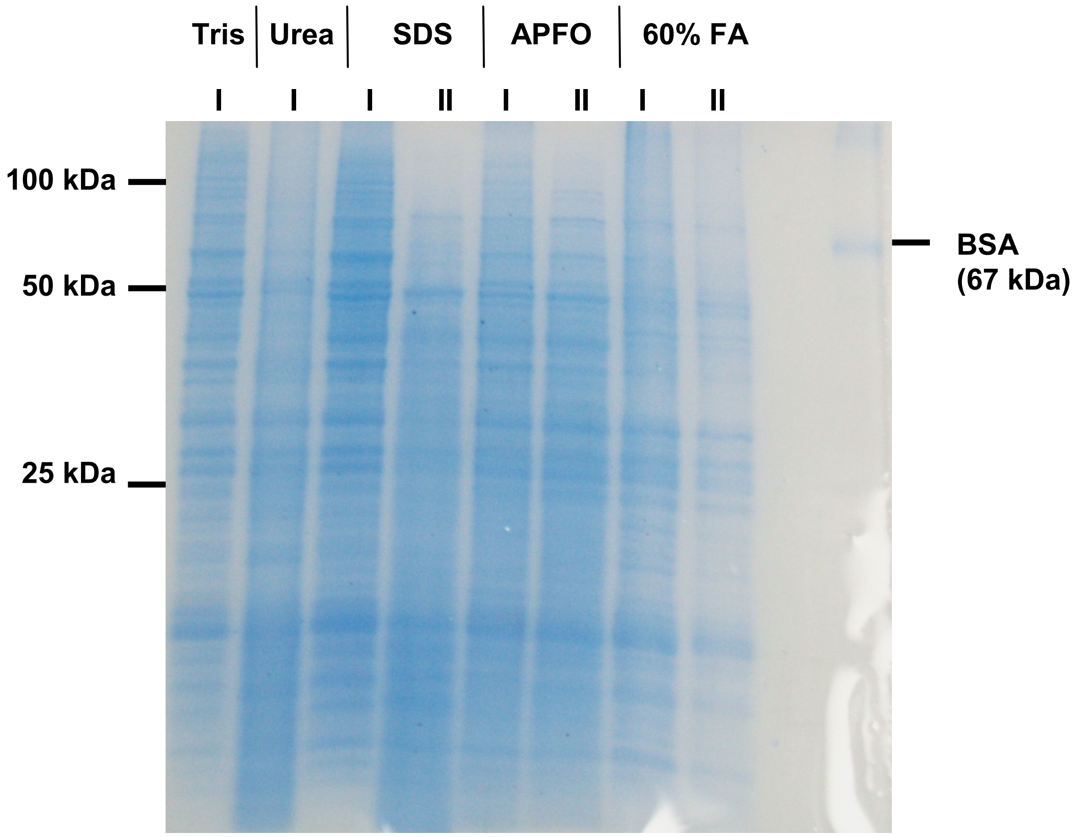


Figure 3.1 Coomassie stained SDS-PAGE showing comparison of protein extraction methods. Five lysis solutions, 50 mM Tris-HCl buffer pH 7.5(Tris), 50 mM Tris-HCl/8M urea (Urea), 50 mM Tris-HCl/4% SDS (SDS), 50 mM Tris-HCl/2% APFO (APFO), and 60% formic acid in water (60% FA), were tested. Yeast protein was extracted using two lysis methods, French press (I) and boiling at 100ºC for 10 min (II).

boiled samples from each lysis solution, indicating the inefficiency of boiling for complete proteome extraction. This result is consistent with the study reported by Kushnirov *et al.* [121] where yeast cells were lysed by boiling in a sample buffer containing SDS. A possible explanation for the lack of high MW proteins from boiled samples could be that the cell wall may be acting as a barrier, preventing extraction of large MW proteins. Another explanation could be due to heat degradation, meaning that, while they may be solubilized, they might simply not appear in the expected region of the SDS-PAGE gel.

The extraction efficiency of the Tris-HCl/APFO lysis solution is noted to be similar using both boiling and French press and provides better extraction efficiency over the Tris-HCl/SDS lysis solution following both extraction protocols (see Table 3.1 and Figure 3.1). APFO is a volatile ionic detergent that has greater extraction efficiency than SDS, but can be removed by evaporation using SpeedVac following protein extraction [122]. These properties of APFO make this lysis solution recommended for whole proteome extraction.

Another lysis solution, 60% formic acid/water is found to be ineffective for whole yeast proteome extraction in either of the lysis methods. Protein extraction with 60% formic acid shows significantly fewer protein bands than Tris-HCl containing lysis solutions (Figure 3.1). A hypothesis for this ineffective extraction is the interaction of the formic acid with the cell membrane or wall, leading to structural changes which could cause incomplete extraction. Although this solution has excellent protein solubility [65,67,116] and is free of detergents and buffers, 60% formic acid can not be used for proteome extraction because of incomplete proteome extraction.

High yields of proteins extracted using the French press from all lysis solutions indicate that this is the most effective method for protein extraction. However, protein banding patterns (Figure 3.1) from French Press extracted samples show different patterns between each extraction solution, implying variable protein recovery between them. Because SDS-PAGE only provides a rough mass distribution of separated proteins, the gel image is not an effective means of determining the optimum protocol for proteome extraction. To address this limitation of SDS-PAGE, proteins from each extract should be identified by GeLC MS/MS [56,57] (a comprehensive protein analysis method) in order to gain intrinsic information of identified proteins such as mass, pI or hydrophobicity.

3.3.2 Characterization of Identified Proteome using LC-MS/MS

A portion of each French press-extracted yeast sample was subjected to LC-MS/MS analysis. Identified proteins are displayed in Figure 3.2 in Venn diagrams. An identical amount of protein (50 μg) was analyzed from each extraction condition, therefore, any variability in the total number of identified proteins are due to differences among the extracts. The greatest number of identified proteins, 1,204, was found in the SDS-containing extract. However, this number only represents 21% of the entire yeast proteome (5,802 proteins). This is a primary consequence of the limitations of proteome profiling through LC-MS/MS [25,86]. In any proteomic experiment, it is the most abundant proteins which are first identified. The identified proteins of the extracts are taken as a representative sampling of the yeast proteome contained in the extract, and can thus be used to assess the effectiveness of the extraction, as discussed below.

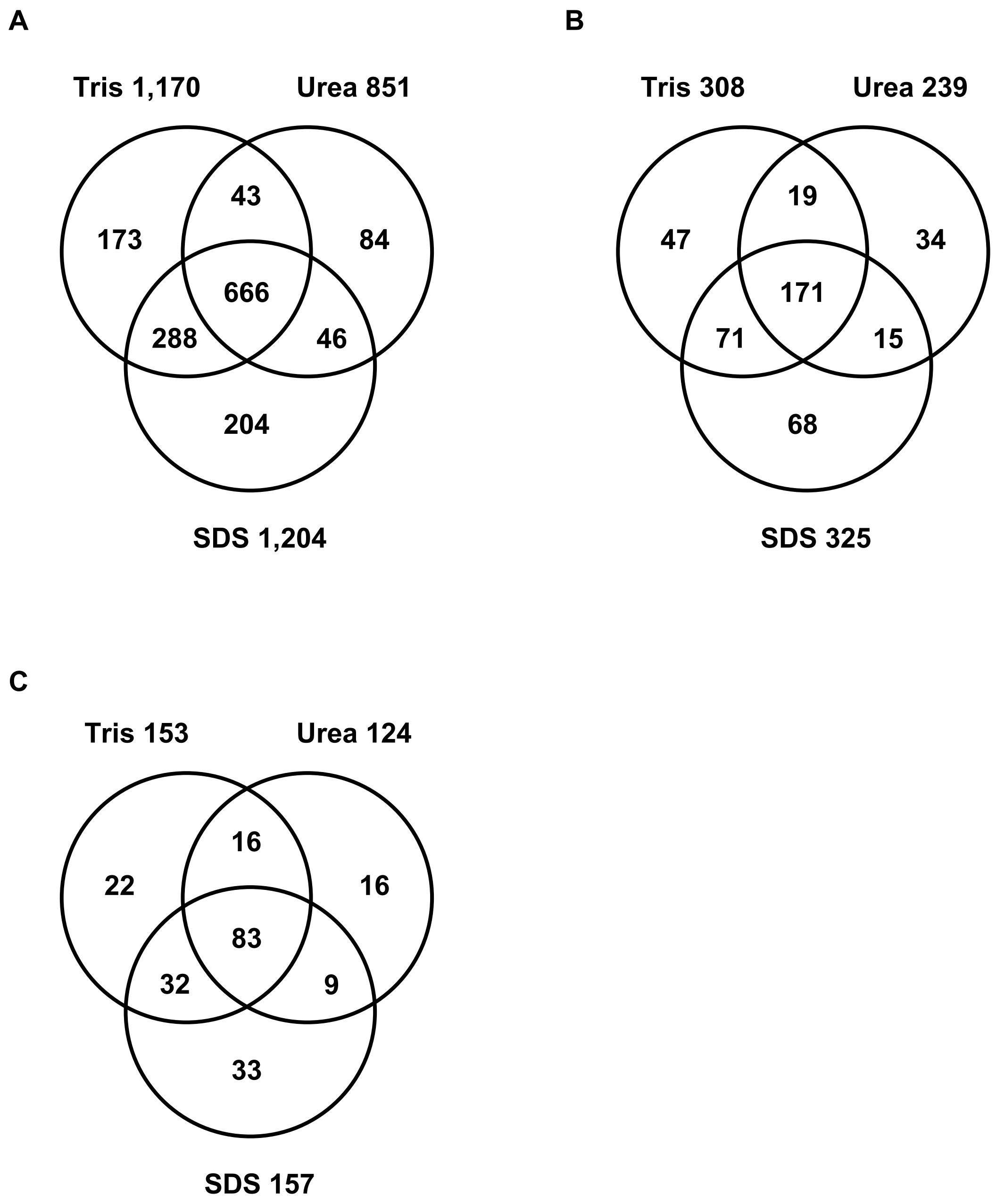


Figure 3.2 Venn diagrams showing overlap of identified proteins from the three lysates. The extraction efficiency of the three lysis buffers, 50 mM Tris-HCl pH 7.5 solution (Tris), Tris-HCl with 8M urea (Urea) or Tris-HCl with 4% SDS (SDS), are determined using number of totally identified proteins (A), integral membrane proteins (B) and helical transmembrane proteins (C). Number of identified proteins for each lysate is indicated with a label (i.e., Tris 1,170).

The Tris-HCl/urea extract yielded a significantly lower number of identified proteins than Tris-HCl/SDS and Tris-HCl/water extracts. This unexpected observation is possible due to heating. The conventional protocol for running SDS-PAGE requires the sample to be heated, typically at 95ºC for 5 min. Thus, urea will hydrolyze to cyanate upon heating, which may result in modification of proteins (so-called carbamylation) at N-terminus, lysine or arginine residues, and hence evokes artifactual charge heterogeneity [50,123]. This modification possibly affects the separation of protein in the gel because SDS can not associate with positively charged side chains (lysine or arginine), and also prevents protein digestion by trypsin. These may explain why a lower number of proteins in the Tris-HCl/urea extract are identified and missing protein bands with high mass are clearly shown in lane I of the Tris-HCl/urea extract (see Figure 3.1).

The total number of identified proteins between Tris-HCl/SDS and the Tris-HCl/water extracts are very similar, however nearly 20% of the identified proteins are unique to each extract. The variation in the total number of identified proteins, together with a large number of non-overlapping proteins between the various extraction buffers points to clear differences in the effectiveness of each extraction process. To quantify the effectiveness of each extraction, in terms of their degree of bias toward complete proteome extraction, the properties of the identified proteins such as mass, hydrophobicity, and isoelectric points (pI) are compared.

3.3.2.1 Mass Bias

The complete yeast proteome contains a distribution of proteins over a wide mass range, from 5 to over 200 kDa (the database of yeast proteome obtained from UniProtKB). As shown in Figure 3.3, the majority of these proteins are found between 20

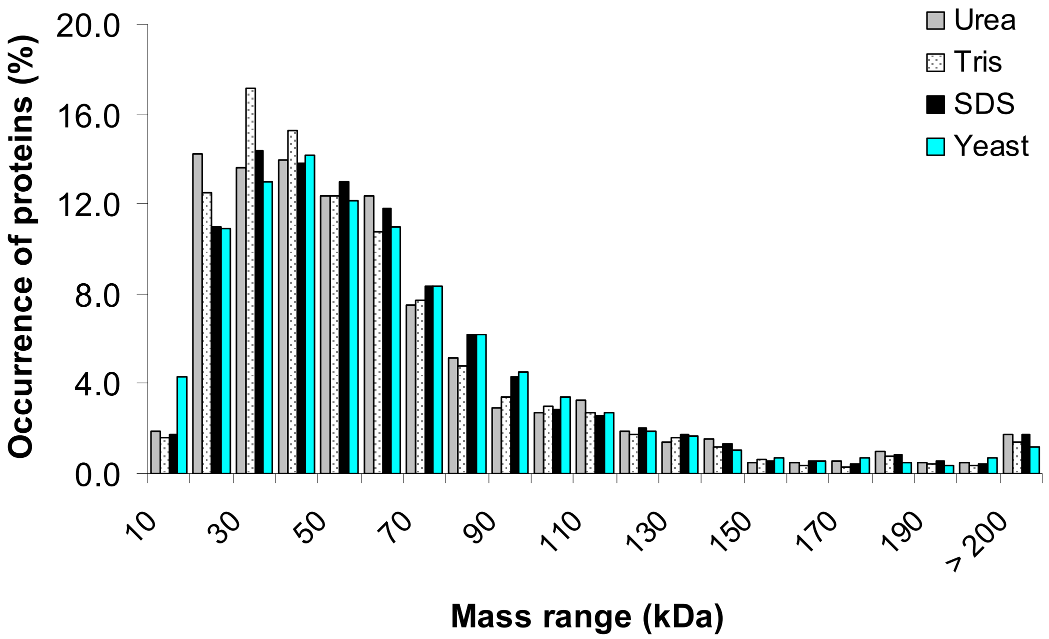


Figure 3.3 Mass distribution of the % occurrence of identified proteins from the three lysates. Relative to the entire yeast proteome (Yeast), the distributions of mass range versus the % occurrence in the total identified proteins from 50 mM Tris-HCl with 8 M urea extract (Urea), 50 mM Tris-HCl extract (Tris), and the 50 mM Tris-HCl with 4% SDS extract (SDS).

and 100 kDa. The mass distribution of the complete yeast proteome is compared to that of the identified proteins from each of the three extraction conditions, which are also plotted in Figure 3.3. A qualitative examination of these results reveals only minor differences in mass distribution of extracted and analyzed proteins relative to those of the theoretical distribution. The frequency of identified proteins with low mass (≤30 kDa) increases slightly with the 8 M urea and the Tris-HCl buffer systems, however the SDS solvent system appears to provide a mass distribution which most closely resembles that of the theoretical yeast proteome distribution.

Chi-square ($\chi 2$) analysis of the mass distribution profile of each of the three lysates shows that the extract proteins from the Tris-HCl/SDS lysis solution is the least bias towards mass distribution compared to the mass distribution of the entire proteome. Because this χ^2 test only revealed the difference on mass distribution for the three lysates, the additional information (pI and GRAVY distributions) can be used to select a suitable lysate for further comprehensive proteome analysis.

3.3.2.2 pI Bias

Yeast proteins display a wide range of pI values, with approximately 50% of which are above 8 [76]. Figure 3.4A compares the pI distributions of the identified proteins from each lysate to the theoretical distribution. A similar trend of the pI distribution is shown for all three lysates, displaying a slight bias towards identification of acidic proteins (pI values lower than 7*)*. According to a χ^2 test, the Tris-HCl/SDS lysate gives the most similar distribution of identified proteins to that of the whole yeast proteome based on pI. Thus, on the basis of protein pI, the most effective solvent system for the least unbiased proteome extraction is the buffer containing SDS.

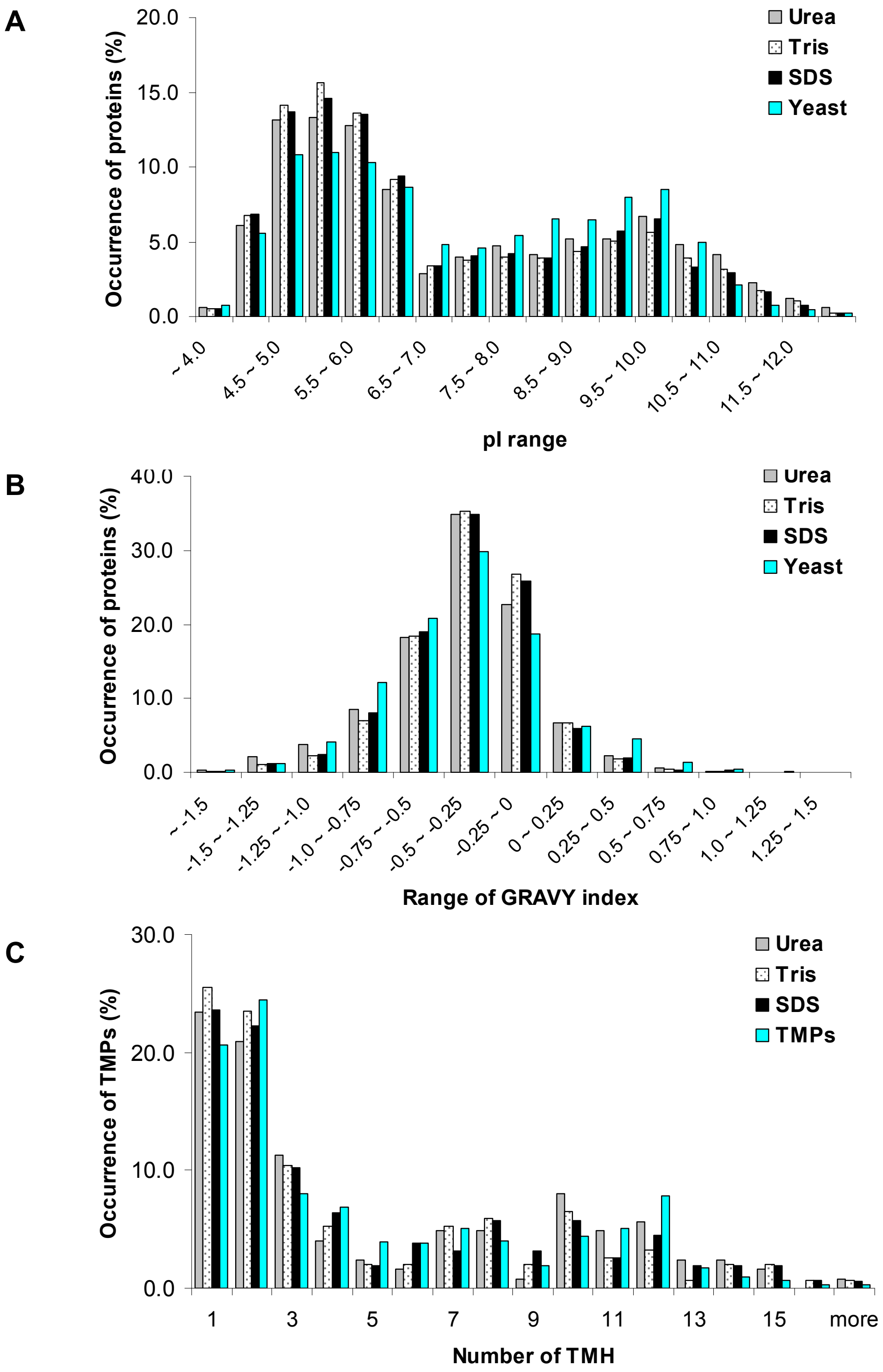


Figure 3.4 Comparison of identified proteins from the three lysates on pI (A) and GRAVY distribution (B). The two distributions are also compared to those of the entire yeast proteome. (C) The number of TMH distribution of the % occurrence of identified transmembrane proteins (TMPs) from the three lysates is compared to the dataset of 1,136 TMPs.

3.3.2.3 Hydrophobicity Bias

The hydrophobicity (GRAVY) distributions of identified proteins are shown in Figure 3.4B. Proteins are classified into four groups according to their GRAVY range: hydrophilic (lower than -0.5), mildly hydrophilic (-0.5 to 0), mildly hydrophobic (0 to 0.5), and hydrophobic (higher than 0.5) [124]. The majority of identified proteins in all extraction solutions are hydrophilic or mildly hydrophilic proteins. This is consistent with the samples where 70% of a whole yeast proteome are hydrophilic [77]. For a representative sampling of the yeast proteome, hydrophobic proteins or transmembrane proteins would be expected to be identified. As seen in Figure 3.4B, mildly hydrophobic and hydrophobic proteins are identified in the three lysates even though percentages of those proteins are low. Also, hydrophobic protein identification is consistently observed by the number of identified IMPs/transmembrane proteins for the three lysates (see Fig. 3.2B & C).

Comparing the three lysates on GRAVY distribution of identified proteins, a χ^2 testing show that the Tris-HCl/urea lysate is least biased compared to the entire yeast proteome. Additionally, the percentages of IMPs/transmembrane proteins identified from the Tris-HCl/urea lysate are the highest (28.1% and 14.6%) among the three lysates, and are slightly lower than the theoretical ones (31.6% and 19.6%). Using visualization of SDS-PAGE analysis, Soulié *et al.* [123] demonstrated that the aggregation of IMPs can be reduced by adding urea into the Laemmli gel buffer [118], and Horvath *et al.* [119] also stated that the efficient extraction of yeast proteins would inprove through using the urea-containing gel buffer. However, overheating should be avoided. Here, the study confirms this observation. Although the highest of percentages of IMPs/transmembrane proteins are identified in the Tri-HCl/urea lysate using the GeLC analysis, the number of identified total proteins for this lysate is the lowest within the three lysates. The Tris-HCl/urea lysate is therefore not recommanded.

A closer assessment of the distribution of identified proteins according to hydrophobicity is shown in Figure 3.4C where the occurrence of transmembrane proteins identified is displayed according to the number of TMH domains identified in those proteins. According to χ^2 testing, extraction with Tris-HCl/SDS is shown to be the most efficient for hydrophobic protein extraction, comparing to the other two lysis buffers.

Membrane proteins tend to be of lower abundance in the cell, and are thus less likely to be identified [10,76]. Thus, under the current technique for protein identification as mentioned earlier, the least degree of bias toward complete proteome extraction is a proper lysate to represent a sampling of the yeast proteome. As a conclusion, the Tris-HCl/SDS lysate is taken as the best solvent system for yeast proteome extraction, providing the highest number of identified proteins, without significant compromise to the mass, pI and hydrophobicity distribution of the identified proteins.

3.3.3 Comparison of CMW and Acetone Precipitation Methods

The optimal extraction buffer for 'complete' proteome extraction has been shown to be Tris-HCl/SDS. Prior to loading on a reversed phase column, this detergent must be removed, along with other interfering substances (such as nucleic acids, lipids, salts, etc.). Some of the most common protein purification methods are precipitation, dialysis, gel filtration, and solid-phase extraction [50,80]. Among them, protein precipitation has been demonstrated as an effective strategy for SDS removal, wherein the concentration of SDS can easily be reduced by over 3 orders of magnitude [85]. Dialysis on the other hand will not efficiently remove SDS and may also cause large sample loss [80,85]. Solid-phase extraction is also not as effective for SDS, and also may suffer poor protein recovery [85]. Therefore, precipitation methods will be used for protein purification [59,125].

Among the many available precipitation methods such as ammonium sulfate [126] and trichloroacetic acid (TCA)/acetone [126], chloroform/methanol/water (CMW) and acetone precipitation are chosen here for SDS removal at high protein recovery after protein extraction. Puchades *et al.* [85] reported the protein recovery from a two protein standard mix for CMW and acetone precipitation to be 50% and 80%, respectively. In a purely qualitative study, Jiang *et al.* [126] determined that the two solvents yielded 'satisfactory' recovery of human plasma proteins, through visualization by SDS-PAGE. Using extracted yeast protein from Tris-HCl/SDS, total protein recovery (assessed from BCA protein assay) was found to be 97 ± 4% from cold acetone and 79 ± 3% for CMW precipitation (four replicates were used and shown in Figure 3.5). These results suggest that high recovery and reproducibility are possible with precipitation, so long as extreme care is taken to avoid protein loss as the supernatant is decanted (described in experimental Section 3.2.5.2).

Protein banding patterns are observed to be nearly identical between the control and precipitated samples, but it should be noted that a certain degree of protein loss is inevitable through each precipitation approach. To assess the types of proteins lost during precipitation, a yeast proteome extract was fractionated into membrane and soluble fractions using ultracentrifugation and each fraction was precipitated. Any proteins remaining in the supernatant of the precipitate were dried by a SpeedVac to permit visualization on SDS-PAGE. Referring to Figure 3.6, protein bands with low mass are

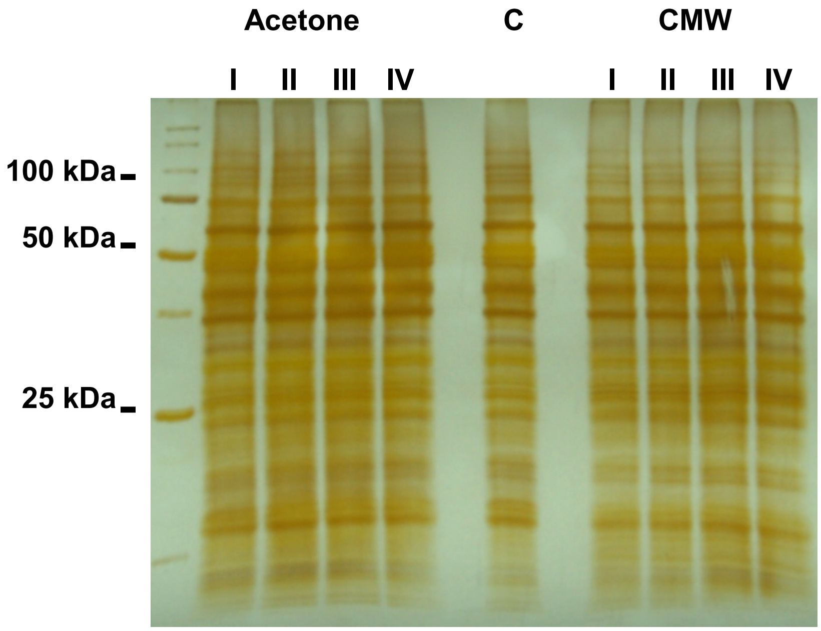


Figure 3.5 SDS-PAGE silver stained gel image displaying protein recovery after Acetone or CMW precipitation. Four replicates (I-IV) of each precipitation method show their reproducibility. As a control (C), 10 µg of unprecipitated yeast protein extract was similarly resolved on the gel. Both precipitation methods show similar banding patterns as the control, however acetone precipitation resulted in slightly darker, and therefore more concentrated bands.

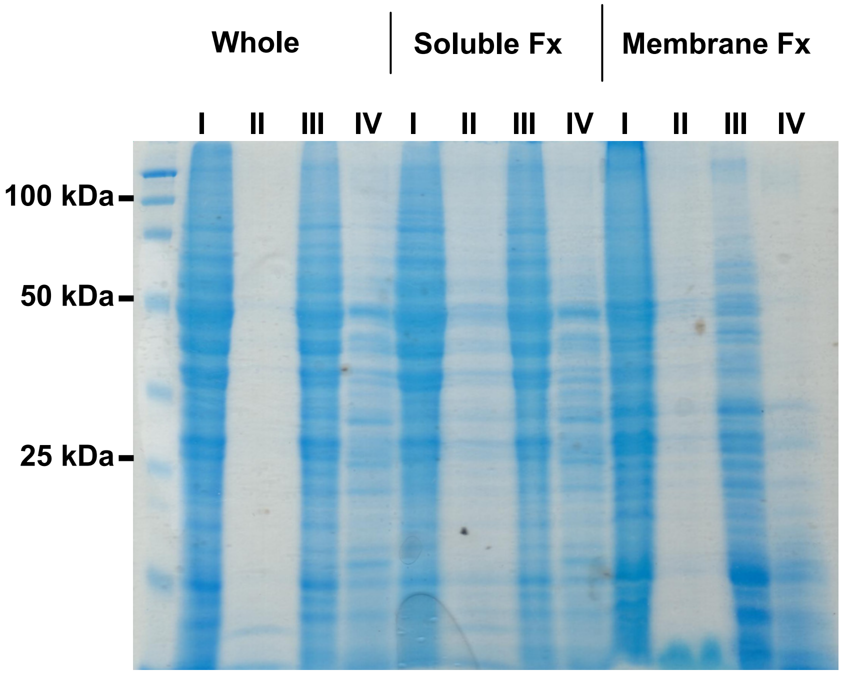


Figure 3.6 Coomassie stained SDS-PAGE gel showing protein recovery of acetone precipitation. The protein solubilizing efficiency of 60% formic acid (III) and 0.1% TFA (IV) solvent system was tested through dissolving precipitated protein pellets from whole cell lysate, the soluble protein fraction (Soluble Fx), and membrane protein fraction (Membrane Fx). Lanes I and II represent the original sample without precipitation and the precipitation supernatant (acetone + solution), respectively.

shown in the acetone supernatants of the whole cell extraction, with more protein bands observed in the soluble protein fraction. These results suggest that the small amounts of yeast protein which are lost upon the acetone precipitation are low molecular weight proteins.

A much higher degree of sample loss was observed with the CMW precipitation (23% loss). Figure 3.5 provides a qualitative assessment of the recovered proteins, which appears to be very similar to the control lane (unprecipitated extract). A more thorough assessment is therefore provided through LC-MS/MS analysis of the proteins recovered from the SDS-PAGE separation (i.e., a GeLC MS/MS experiment).

As seen in Table 3.2, the total number of identified proteins/IMPs/transmembrane proteins from the CMW precipitation is slightly higher than that of the acetone precipitation. This follows an adjustment for protein recovery, wherein equal amounts of sample (50 μg) were taken for MS analysis. A large number of overlapping proteins are observed between methods (about 80% in common), suggesting that the two methods have similar efficiency for protein precipitation. However, to fully assess the effectiveness of these precipitation methods, the distribution of protein properties (mass, pI and GRAVY) were determined as described in the previous section.

3.3.3.1 Assessment of Bias from Protein Precipitation

Figure 3.7 indicates that the mass, pI or GRAVY distribution of identified total proteins presents a similar trend between acetone and CMW precipitation methods. For a 'complete' proteome precipitation, the CMW precipitation method is slightly better than the acetone method using χ^2 testing. Additionally, the percentage of identified

Table 3.2 Comparison of the two precipitation methods.

Protocol	# total proteins	# unique proteins	# IMPs	# TMPs	% IMPs	% TMPs	Recovery (%)
Acetone	**1,363**	**205**	**365**	**175**	**26.8**	**12.8**	**97 ± 4**
CMW	**1,451**	**293**	**402**	**197**	**27.7**	**13.6**	**79 ± 3**

Notes: TMPs represents alpha-helix containing transmembrane proteins. Recovery of proteins is calculated through difference of protein masses before and after precipitation.

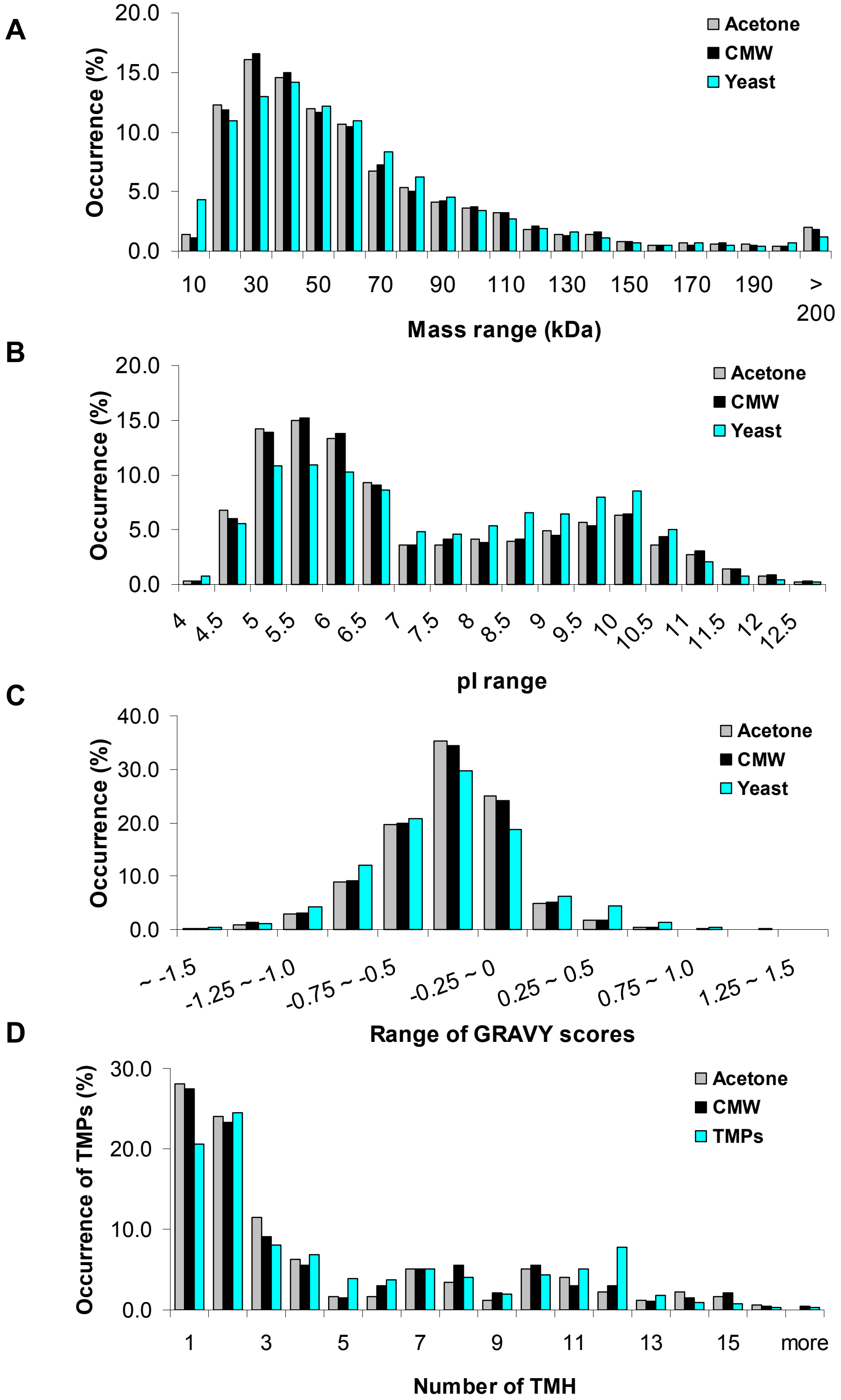


Figure 3.7 Assessment of Acetone and CMW precipitation methods. Relative to the entire yeast proteome (Yeast), the % occurrence of proteins identified following Acetone and chloroform/methanol/water (CMW) precipitation are displayed according to (A) mass, (B) pI, and (C) GRAVY range. (D) The number of TMH domain distribution of the % occurrence of identified transmembrane proteins (TMPs) from the two precipitation methods are compared to that of the dataset of 1,136 TMPs.

IMPs/transmembrane proteins for the CMW (27.7%/13.6%) is slightly higher than that of acetone precipitation (26.8%/12.8%). It is noted that those percentages are identical to that of the Tri-HCl/SDS lysate (before precipitation), implying that the two precipitation methods are unbiased toward the yeast proteome.

This result is further examined on the distribution of identified transmembrane proteins in terms of number of TMH domains per protein. The CMW precipitation method is slightly better than the acetone method using χ^2 testing. Furthermore, the TMH distribution of occurrence on each number is similar between the un-precipitates (from the Tris-HCl/SDS lysate) and the precipitates (CMW precipitation). Therefore, although the MS detection platform (i.e., LC-MS/MS) is likely biased towards hydrophilic proteins, the CMW precipitation methods do not appear to contribute any greater degree of protein loss for IMPs/ transmembrane proteins.

In summary, the CMW and acetone precipitation methods are considered to be acceptable approaches for protein purification, but the CMW method is slightly better. The CMW precipitation has also been shown to be a better approach for removal of SDS as well as lipids [59], therefore, the CMW method is selected as optimal for sample preparation.

3.3.4 Comparing Protein Resolubilization

Following precipitation, proteins must be resolubilized in a solvent system suitable for RPLC separation. Aqueous solutions with low levels of a miscible organic solvent and traces of acid (water, <10% acetonitrile with 0.1% TFA) are commonly used for protein separation on RPLC [67]. However, such a solvent system may be problematic for complete resolubilization, particularly for hydrophobic proteins.

Therefore, a higher concentration of acetonitrile may increase initial protein resolubilization efficiency [59], which could then be diluted to levels suitable for RPLC injection. To test the efficiency of this method, 30% acetonitrile with 0.1% TFA is selected as a potential solvent to improve protein resolubilization. A 60% formic acid and isopropanol solvent system was established by Heukeshoven *et al.* [67] to separate hydrophobic proteins with RPLC. Formic acid has an unrivaled ability to solvate hydrophobic proteins, thus 60% formic acid [65] and 98% formic acid [4]) were used to dissolve protein pellets.

A significant difference of protein resolubilization is presented between the water/ acetonitrile and the formic acid solutions (see Fig. 3.8). Increasing the concentration of acetonitrile from 5% to 30% results in darker bands, consistent with the protein recovery measured by using the Bradford protein assay. The recovery of protein resolubilized increased from 9% to 12% when the concentration of acetonitrile increased from 5% to 30%. These results show low protein resolubilization in the water/acetonitrile solutions. Conversely, the formic acid/water solutions demonstrate high protein resolubilization. The banding pattern of proteins resolubilized using formic acid show equal resolubilization of proteins across a wide mass range, making the formic acid/water solutions optimal for resolubilizing precipitated protein pellets.

It is interesting to note that no difference of protein solubilzation is evident between 60% [65] and 98% formic acid [4] solutions. Thus, 60% formic acid is selected to dissolve the precipitated proteins because of lower extent of protein formylation in lower formic acid concentration [4]. The high solubilization efficiency of the 60% formic

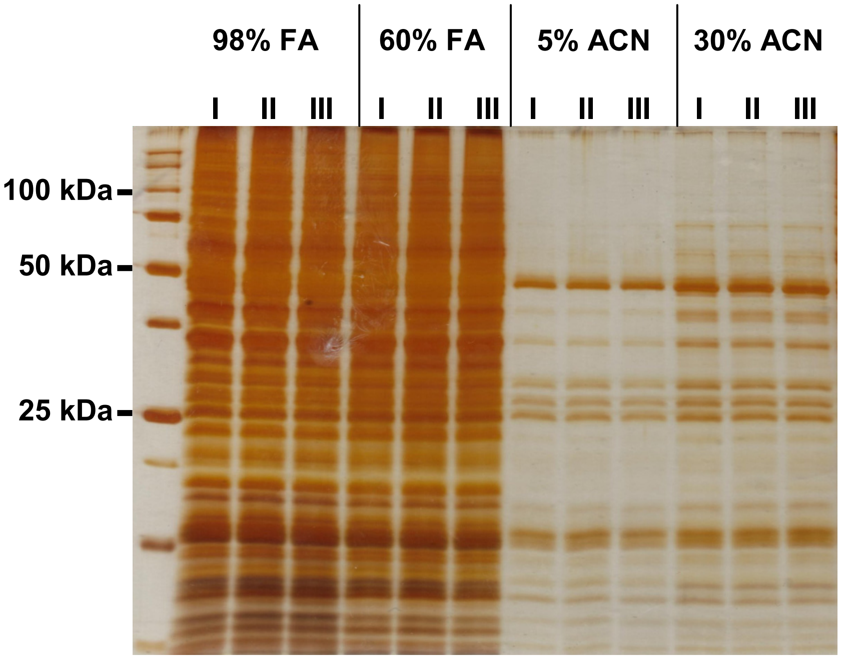


Figure 3.8 SDS-PAGE silver stained gel image displaying the resolubilizing efficiency of the four solvents. The protein solubility of 98% formic acid (98% FA), 60% formic acid (60% FA), 5% acetonitrile with 0.1% TFA (5% ACN) and 30% acetonitrile with 0.1% TFA (30% ACN) were tested by redissolving protein pellets, which were precipitated from the whole cell extract using chloroform/methanol/water precipitation. I, II and III indicate replicate resolubilization procedures.

acid solution is further demonstrated in Figure 3.6 where 60% formic acid shows better protein solubilization than water with 0.1% TFA on proteins from the whole cell, membrane fraction and soluble proteins. Because 60% formic acid is compatible with RPLC, it chosen for redissolving the protein pellets.

3.4 Conclusion

The objective of this chapter was to optimize the initial sample preparation step of proteome extraction and purification ahead of RPLC separation. The study has shown that yeast proteins can be extractedin a French press using 4% SDS in 50 mM Tris-HCl as lysate. The optimum procedure for purification was shown to be using CMW precipitation and resolubilization in 60% formic acid.

Chapter 4

Proteome Prefractionation by RP-HPLC

4.1 Introduction

A complex proteome mixture (i.e., *S. cerevisiae*) can be effectively isolated from its raw cellular matrix using the sample handling protocols established in Chapter 3. The extracted protein sample is compatible with reversed-phase liquid chromatography (RPLC), a solution-based separation and analysis approach, which is the subject of the current chapter. The goal of this study is to investigate the effectiveness of a solution-based proteome prefractionation strategy ahead of the analysis of the protein digestion products with liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). Prefractionation is implemented to reduce the complexity prior to digestion and LC-MS/MS analysis. While both the 'LC' component of LC-MS/MS and the solution-based prefractionation of the proteome entail RPLC, it is important to distinguish peptide from 'intact' (protein) level prefractionation. Although gel-based platforms (SDS-PAGE) continue to dominate prefractionation strategies, solution-based (gel-free) protein separation approaches provide unique advantages, including higher throughput, ease of automation, and potential for direct (online) as well as offline coupling to mass spectrometry (MS) [109,115,127]. Because of this, solution-based protein separation approaches have generated considerable interest for comprehensive MS-based proteomics [11,109].

The extreme complexity of a cellular proteome, including the vast dynamic range of protein abundance presents a challenge for obtaining a complete profile of the proteome. Perhaps the greatest challenge in MS analysis is avoiding highly abundant proteins that mask those with lower abundance, a task which is often addressed through separation. Even with considerable separation, the current MS platform used to conduct this study will still not be able to detect all protein components in the mixture. Thus, a representative profile of the proteome is used as an assessment in relation to the theoretical distribution of proteins in the yeast proteome.

Strategies for solution-based proteomic separation include solution-phase isoelectric focusing (sIEF) [47], free-flow electrophoresis (FFE) [47], capillary electrophoresis (CE) [128], and liquid chromatography (LC) [64,115]. The last was selected for this study for its versatility, but also in consideration of the limitations of all other separation strategies. For example, when proteins are separated by FFE or sIEF based on their pI, there is a concern for protein loss because of the low solubility of proteins in solution when the pH of the solution is equal to their pI. Moreover, the low resolution is the other limitation by diffusion effects [47]. CE provides a potential for incredible resolution, however, the method is only compatible with minimal quantities of sample (nanograms) and also suffers a large risk of sample loss due to adsorption to the capillary walls [127,128].

As a high-throughput and automated solution-based separation approach, LC is broadly employed for proteome research, given the many available options for selection of stationary and mobile phase. LC strategies for protein separation include RPLC, hydrophobic interaction chromatography (HIC) [115], hydrophilic interaction chromatography (HILIC) [129], ion exchange chromatography (IEX) [115], chromatofocusing (CF) [109], affinity chromatography [115], and size-exclusion chromatography (SEC) [115]. Among the LC strategies, RPLC is the preferred choice for comprehensive proteome prefractionation, based on the potential for high resolution and high yield. HIC and HILIC are mainly for soluble protein separation [115,129]. Both CF and IEX have low protein recovery, and IEX also generates poor resolution due to separation based on surface charges rather than isoelectric point (pI) [130]. Affinity and size-exclusion chromatography are somewhat specialized techniques which have important applications for proteome enrichment or purification, but are considered low resolution techniques, and are therefore not viable options for protein separation. Affinity chromatography essentially generates two fractions (bound *vs* unbound), while SEC is difficult to apply across a broad mass range, and has limited mass resolution for folded proteins. By contrast, RPLC separates proteins according to their hydrophobicity. The selection of stationary and mobile phase provides finite control of the degree of protein binding to the column, which translates into high resolution for proteins. The solvent system is also directly compatible with MS, which offers the potential for top-down (intact protein) MS profiling. Prefractionation of the proteome according to hydrophobicity is therefore selected as the focus of this study.

RPLC separation of proteins is typically accomplished using a water/acetonitrile gradient solvent system [67], providing excellent resolution and recovery for peptides, being the standard approach to bottom-up LC-MS/MS analysis [86]. Using these same conditions, RPLC separation of intact proteins has been conducted by Martosella *et al.* [131] and Neverova *et al.* [132], and was demonstrated to be an effective prefractionation technique ahead of LC-MS/MS analysis. However, a concern of intact protein separation by RPLC is that the integral membrane proteins (IMPs), in particular transmembrane proteins, do not elute efficiently from the column [114,132]. Without a potential for chaotropes or detergents to maintain protein solubility, sample loss can occur before proteins are injected onto the column, or can lead to irreversible protein binding or precipitation onto the column. A variety of mobile and stationary phases are studied in order to optimize RPLC separations of intact proteins. A conventional C_{18} column can be substituted with other stationary phases, such as C_4, Poros R1 (polymer) or non-porous RP column, to increase protein recovery [64,69]. Using four solvents that included water, isopropanol, formic acid and acetonitrile, Martosella *et al.* [133] improved the efficiency of separation for hydrophobic proteins; however, this method was not assessed in terms of comprehensive proteomic separation. An assessment of RPLC for comprehensive proteome prefractionation using various column supports, as well as alternative solvent systems is necessary.

Improved RPLC separation of hydrophobic proteins can be obtained through application of an aqueous formic acid and alcohol solvent system [67,134]. This is attributed to formic acid having unrivaled solubilizing efficiency for hydrophobic proteins. Whitelegge *et al.* [64] found that the integral membrane protein complex, thylakoid-membrane pigment-protein complex, can be dissolved in 60% formic acid/water and was efficiently separated on a polymer RP column using a gradient solvent system comprised of 60% formic acid and isopropanol. Still others have demonstrated effective separation of membrane proteins using formic acid with acetonitrile [66,133] or 1-butanol [67,134]. One of the noted benefits of these solvent systems is that they are still directly compatible with ESI-MS; however, these solvent systems have only been demonstrated for enriched fractions comprising hydrophobic proteins, not for comprehensive proteome separation. It is therefore possible that enhanced separation of hydrophobic proteins comes at the expense of poor retention for the more hydrophilic (and dominating) components of the sample.

In this study, a 60% formic acid/isopropanol solvent system was selected for prefractionation of the extractable proteome from *S. cerevisiae* (see Chapter 3). The solvent system is compared to a conventional acetonitrile/water gradient using multiple stationary phase and supports in RP-HPLC. The merits of each solvent system on the various column supports are discussed in terms of the extent of fractionation and recovery of proteins, as identified through mass spectrometry.

4.2 Experimental

4.2.1 Materials and Reagents

Formic acid (88%, cat. A118P), HPLC-grade isopropanol, HPLC (MS)-grade acetonitrile, HPLC-grade methanol, and acetic acid were purchased from Fisher Scientific (Ottawa, Ontario). Formic acid (≥98%, cat. 06440), trifluoroacetic acid (TFA), ammonium bicarbonate, iodoacetamide, cyanogen bromide (CNBr), bovine trypsin (TPCK treated, cat. T8802), protease inhibitor cocktail (cat. P2714), and yeast extract-peptone-dextrose (YPD) were purchased from Sigma (Oakville, Ontario). SDS, Tris, dithiothreitol (DTT), and other reagents used for casting and running SDS-PAGE were obtained from Bio-Rad (Mississauga, Ontario). All other chemicals were obtained from Sigma and were used without further purification. Milli-Q grade water was purified to 18.2 $M\Omega cm^{-1}$.

4.2.2 Sample Preparation

Yeast cells of *Saccharomyces cerevisiae* (strain BY4741), freshly streaked on an agar plate, were generously supplied by Dr. Melanie Dobson (Department of Biochemistry & Molecular Biology, Dalhousie University) and cultured in the lab as described in Section 3.2.2 of Chapter 3. Cells were suspended in 3.5 mL of lysis buffer (50 mM Tris-HCl pH 7.5 with 4% SDS or 8M urea), to which 250 μL of protease inhibitor cocktail (general use) was added. The cells were lysed using the French press protocol as described in Section 3.2.3.

For adjustment of RPLC gradients of the two solvent systems (Section 4.2.6), the lysate was extracted from yeast strain BY4741 (from Dr. McMaster). The lysate extracted from yeast strain MT302/28B (from Dr. Murray) was used for preparation of subcellular fractions. Preparation of the proteome extract was as described in Section 3.2.2.

Two subcellular fractions, the cytosolic protein (soluble protein) and membrane protein fractions were prepared as described in Section 3.2.4 of Chapter 3.

4.2.3 Proteome Precipitation

The procedures for cold acetone as well as chloroform/methanol/water (CMW) precipitation were used as described in Section 3.2.5 of Chapter 3.

4.2.4 Proteome Digestion

For CNBr/trypsin digestion, dried protein fractions were redissolved in 50 μL of 70% formic acid (Section 4.2.6) and 10 μL of CNBr in 70% formic acid (10.0 μg/μL) was added, followed by bubbling of nitrogen gas through the solution for 30 seconds to prevent methionine oxidation. Protein samples were incubated overnight in the dark (16 hrs), and the reaction was quenched by adding 500 μL of water and dried in a Speedvac.

Dried samples were redissolved in 100 µL of water/acetonitrile (50:50, *v*:*v*) and dried a second time to remove any traces of CNBr.

All trypsin digests involved resuspension of dried protein fractions (Section 4.2.6) in 50 µL of 100 mM NH_4HCO_3 Followed by disulfide bond reduction with 9.5 mM DTT at 56 ºC for 20 min and alkylation with 19.5 mM iodoacetamide at room temperature in the dark for 20 min. Digestion was performed overnight (16 hrs) at 37ºC upon addition of trypsin at a 1: 50 (*w:w*) trypsin:protein ratio. Digestion was quenched through addition of 10% TFA to reduce the pH of the solution to 2-3. Digested samples were dried in the SpeedVac and stored at -20ºC until needed.

Proteins extracted in 50 mM Tris-HCl pH 7.5 with 8 M urea were diluted six fold to a final concentration of 1.33 M urea and 100 mM NH_4HCO_3 prior to digestion according to the protocol described above. The peptide mixture was dried in the SpeedVac and stored at -20 ºC until use.

4.2.5 Strong Cation Exchange

HPLC separation was conducted on an Aligent 1200 HPLC system with UV detection at 214 nm (Palo Alto, CA, USA). A 1.0 × 50 mm strong cation exchange (SCX) column was prepared by packing 5 µm, 1000 Å Polysulfethyl A resin (PolyLC, The Nest Group, Southboro, MA, USA). Peptides were eluted in a gradient using Solvent A (20 % ACN with 0.1% TFA) and Solvent B (1 M NaCl in solvent A) at a flow rate of 100 µL/min. Fractions were dissolved in 100 µL of (Solvent A), and 95 µL of the sample was injected. The gradient is as follows: 0-10 min, 0% B; 12 min, 8% B, 14 min 25% B; 24 min, 50% B; 24.1 min 0% B. Total run time was 40 minutes to re-equilibrate the column prior to loading the next sample. A single fraction was collected over the time window

22.5 to 23.5 min. The SCX run was repeated three times for the 3 cleaned peptide fractions, with each fraction being dried in SpeedVac for storage at -20ºC.

4.2.6 RP-HPLC Separation

The HPLC system is as described in Section 4.2.5. Total extracted yeast protein as shown in Section 4.2.2.1 was separated on four different RP columns: (I) 259VH5110 Polymer column (1×100 mm, 300 Å, 5 µm) from Grace Vydac (Hesperia, CA, USA); (II) 1×100 mm in-house packed 300 Å, 5 µm Magic C_4 column resins (Michrom, Auburn, CA, USA); (III) 1×100 mm in-house packed 2000 Å, 10 µm Poros 50 R_2 particles (Applied Biosystems, Carlsbad, CA, USA); (IV) Polymer X 5u RP-1 column (2×150 mm, 100 Å, 5 µm) from Phenomenex (Torrance, CA, USA).

Both acetonitrile/water and formic acid/isopropanol solvent systems were tested using yeast proteins isolated as described in Section 4.2.2.1 to select the appropriate gradient conditions. A flow rate of 50 µL/min was maintained, and 160 µg of yeast protein dissolved in 60% formic acid/water was injected onto column (I). The final gradient of the 60% formic acid-isopropanol was as follows: 0-10 min, 5% B; 11 min, 10% B; 29 min, 12% B; 44 min, 15% B; 62 min, 22% B; 71 min, 60% B; 71.1 min, 5% B. For the water-acetonitrile solvent system, the final gradient was as follows: 0-10 min, 5% B; 11 min, 20% B; 71 min, 45% B; 74 min, 80% B; 74.1 min, 5% B.

4.2.6.1 Proteome Prefractionation Used for SDS-PAGE Analysis

For SDS-PAGE analysis, the yeast proteins (Dr. McMaster's lab) extracted as described in Sections 4.2.2.1 & 4.2.2.2 were fractionated on reversed phase columns (II), (III), and (IV). Flow rates and injection volumes were adjusted according to the dimensions of the column: 25 µL (135 µg protein) at 50 µL/min was injected on columns

(II) and (III), while 100 μL (540 μg protein) at 200 μL/min was injected on column (IV). Fractions were collected at 3 min intervals, beginning immediately after injection, collecting a total of 28 fractions per run. Fractions were dried in the SpeedVac, redissolved in water dried a second time, and either separated by SDS-PAGE or digested as described above.

4.2.6.2 Comparison of Two Solvent Systems

For comparison of the water/acetonitrile and formic acid/isopropanol solvent systems, the extracted proteome from yeast strain BY4741 (see Chapter 3 and Section 4.2.2.1) was prepared at a 6 μg/μL protein concentration in 60% formic acid and a 150 μg aliquot of yeast proteins was injected on column (II) and fractionated using the two solvent systems. 20 fractions per solvent system were collected and dried in a SpeedVac.

4.2.7 SDS-PAGE

Dried fractions from RP separation with the water-acetonitrile or the 60% formic acid/isopropanol solvent system were re-dissolved in 18 μL of a Laemmli gel loading buffer [118], and 16 μL was loaded onto a 15% T SDS-PAGE gel. The fractions collected from the column IV using the formic acid/isopropanol gradient were dissolved in 40 μL of the gel buffer and 10 μL was loaded onto a gel. The performance of SDS-PAGE was according to the procedure described in Section 3.2.6 of Chapter 3

4.2.8 Sample Cleanup and Quantification

Prior to MS, all digested peptides were subjected to RPLC sample cleanup according to the conditions described in Section 3.2.9, dried and stored at -20 ºC prior to LC-MS/MS analysis. RPLC cleaned samples were quantified according to their LC-UV absorbance values using a BSA digest as the standard (from 0.25 to 12.5 μg) to generate a calibration curve of the peak area versus amount of BSA digest.

4.2.9 Nano LC-MS/MS

MS conditions used in the comparison of the two solvent systems are the same as described in Section 3.2.9 of Chapter 3. For fractions 1-8 collected from the water/ acetonitrile gradient (exception), only one injection was analyzed.

The alternative LC gradient used in comparison of the three columns during LC-MS/MS was as follows: 0-5 min, 5% B; 90 min, 28% B; 95 min, 60% B; 95.1 min 5% B to re-equilibrate. Samples were re-dissolved in 20 μL of 5% acetonitrile with 0.1% formic acid (Solvent A), and 15 μL of each sample isolated from the columns (II) and (III), and 3.75 μL from column (IV) was injected onto the column. All LC-MS/MS analyses were performed through duplicate sample analysis.

The simplified yeast peptides were pooled by using the three cleanup fractions as mentioned in Section 4.2.5 and dried in a Speedvac. These peptides were used to generate series of standard peptide solutions prepared as: 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, and 5.0 μg of peptide in 10 μL (injection volume). LC-MS/MS analysis was performed as described in Section 3.2.9 of Chapter 3.

A "tripleplay" dependent scan was employed to record MS spectra. A full MS scan (range 400-1200 *m/z*) used in the comparison of the three columns selects the five most intense doubly or triply charged ions for fragmentation, and comparison between the two solvent systems, a full MS scan (400-1700 *m/z*) was used to select the three most intensive doubly or triply charged ions for fragmentation.

4.2.10 Database Searching

MS/MS spectra were searched against a yeast *S. cerevisiae* database containing 5,802 unique protein entries, using the Bioworks 3.2 software package (ThermoFisher), which uses the SEQUEST search engine. The following search parameters were used to identify proteins: precursor ion tolerance 1.5 amu and fragment ion mass tolerance 1.0 amu; fully tryptic peptides with no more than two missed cleavages; chemical modifications that include fixed carboxyamidomethylation (+ 57 Da) at cystine, and differential oxidation (+16 Da) at methionine. For the formic acid/isopropanol solvent system, an additional modification (differential formylation at serine or threonine) was added [135], and for CNBr and trypsin digestion, partially tryptic cleavage and two additional modifications (differential formylation at serine or threonine and differential modification at methionine set to either -30 Da for Homoserine or -48 Da for Homoserine lactone [45,88]) were used for searching. The search results were filtered with the same settings as described in Section 3.2.10 of Chapter 3.

4.3 Results and Discussion

Proteomic analysis of complex samples can be facilitated by protein fractionation prior to peptide identification by LC-MS/MS. This prefractionation improves detection of low-abundant proteins by separating them from high-abundant proteins, reducing ion suppression in electrospray MS. For optimal MS performance, the appropriate amount of sample must be loaded onto the LC-MS/MS system, therefore the total mass of protein, once determined, can be used to guide the optimization of proteome prefractionation through RPLC.

4.3.1 Sample Loading Effect on LC-MS/MS Analysis

In a typical LC-MS/MS experiment with low concentrations, injecting slightly larger amounts of sample results in an exponentially greater number of detected proteins by MS. Wang *et al.* [136] determined the optimal protein loading for an LC-MS/MS experiment to be 1 μg, because higher sample loadings no longer translated into a proportional increase in the number of identified proteins. This experiment was performed on a whole cell extract from a breast cancer line (MCF-7), using a similar LC-MS platform, though it is noted that the optimal loading will likely depend on the specific instrumental platform. Also, the whole cell extract used in the previous study presents a much more complex mixture than would be obtained following prefractionation. Thus, a similar experiment was conducted using a fractionated yeast sample to determine optimal sample loading of prefractionated protein on the available instrumental platform.

Figure 4.1 demonstrates that in the low concentrations (0.01 to 0.2 μg), injecting slightly larger amounts of peptide material results in an extreme increase in the number of peptide identifications, whereas at larger injected amounts (1.0 to 5.0 μg), there is a plateau effect, wherein the number of the identified peptides increases only slightly as more sample is injected. These results contrast with that reported by Wang *et al.* [136], in which the number of identified proteins dropped as sample loading increased from 1.0 to 2.0 μg. This may be a consequence of the difference in samples analyzed (complete proteome vs fractionated sample), or may be a consequence of instrument. Nonetheless, even though the number of identified peptides increase beyond 1 μg sample loading, the gain is only slight, implying that 1 μg sample injections are sufficient for proteomic analysis.

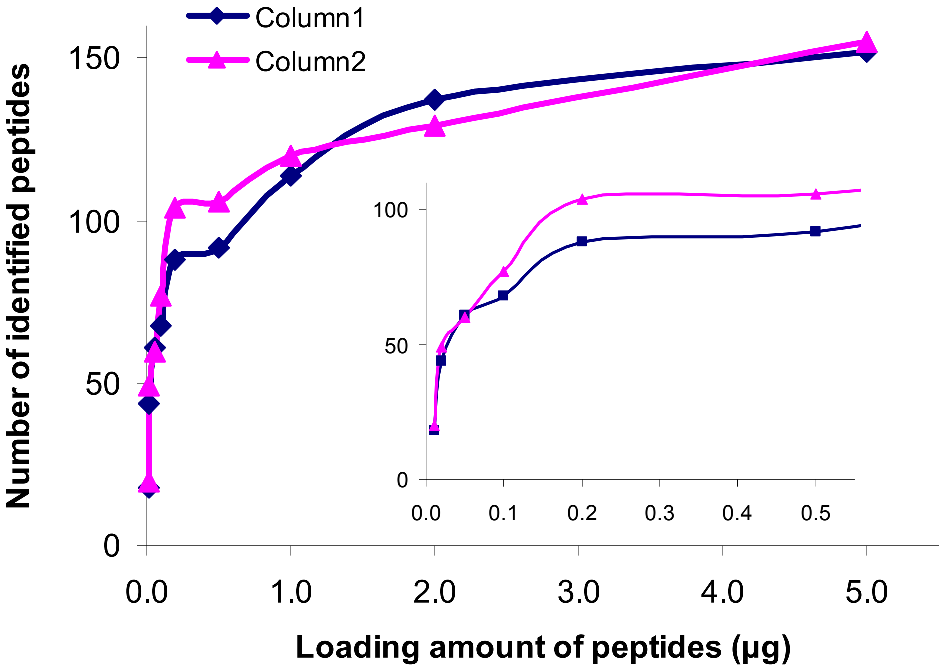


Figure 4.1 Number of identified peptides as a function of the mass of total peptides injected. Insert: the narrow range of mass (0.01 to 0.5 µg). Column 1 & 2 refer to replicate runs on two distinct LC columns.

4.3.2 Suitability of Formic acid/Isopropanol for Whole Proteome Reversed-phase Fractionation

As opposed to the traditional water/acetonitrile solvent system, this study will test a 60% formic acid/isopropanol solvent system for whole proteome fractionation. Here, cytosolic (soluble) and membrane (hydrophobic) protein fractions were separated on a reversed phase polymeric column using a 60% formic acid/isopropanol gradient solvent system as described in Section 4.2.4.1. As shown in Figure 4.2A & B, a clear separation of soluble and hydrophobic proteins is obtained from the each fraction. The efficiency of separation is noted by the unique protein bands present over the many fractions as displayed in the gel. Also noteworthy is the fact that the initial fractions (i.e., the non-retained components) are minimal for both the hydrophobic and hydrophilic proteins. Finally, protein recovery appears to be acceptable when comparing the intensity of bands in each fraction to the control lane (unfractionated yeast), visible in the right-most lane of each gel. A more thorough assessment of recovery is discussed later in this chapter.

4.3.3 Comparison of Solvent Systems for RP separation

Three distinct types of RP columns were used to evaluate the performance of the formic acid/ isopropanol solvent system, using SDS-PAGE to visualize the fractionated proteins. The RP columns include (II) a silica-based RP column packed with C_4 beads, and polystyrene resin supports with a diameter of (III) 10 μm and (IV) 5 μm and pore sizes of (III) 2000 Å and (IV) 100 Å. Prefractionated yeast proteins from these columns are shown in Figure 4.3.

Shown in Figure 4.3, the protein separation patterns are clearly different on the three column types using the 60% formic acid/isopropanol solvent system; however the

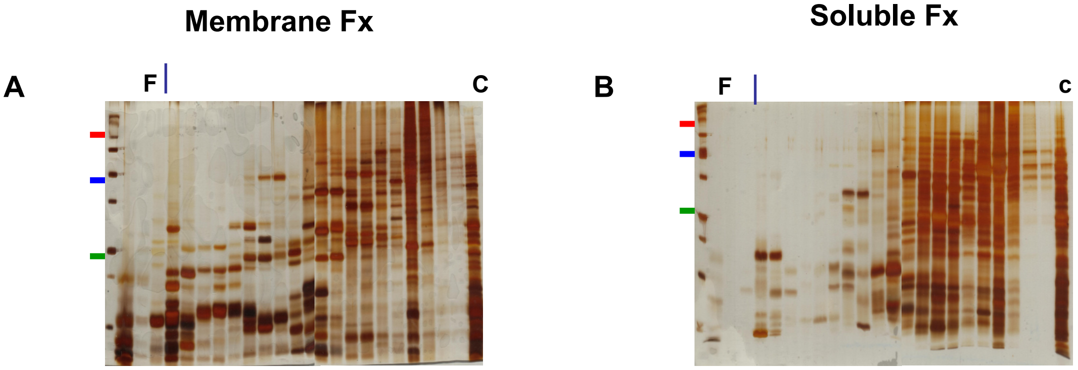


Figure 4.2 SDS-PAGE analysis of fractions collected during RPLC separation. (A) Estimated 190 µg of yeast membrane proteins and (B) 120 µg of yeast soluble proteins were separated on the Grace Vydac Polymer column (Column I) using the 60% formic acid/isopropanol gradient. Protein markers are displayed using 25 kDa (green), 50 kDa (blue) and 100 kDa (red). C stands as a control (loading sample). F represents flow-through.

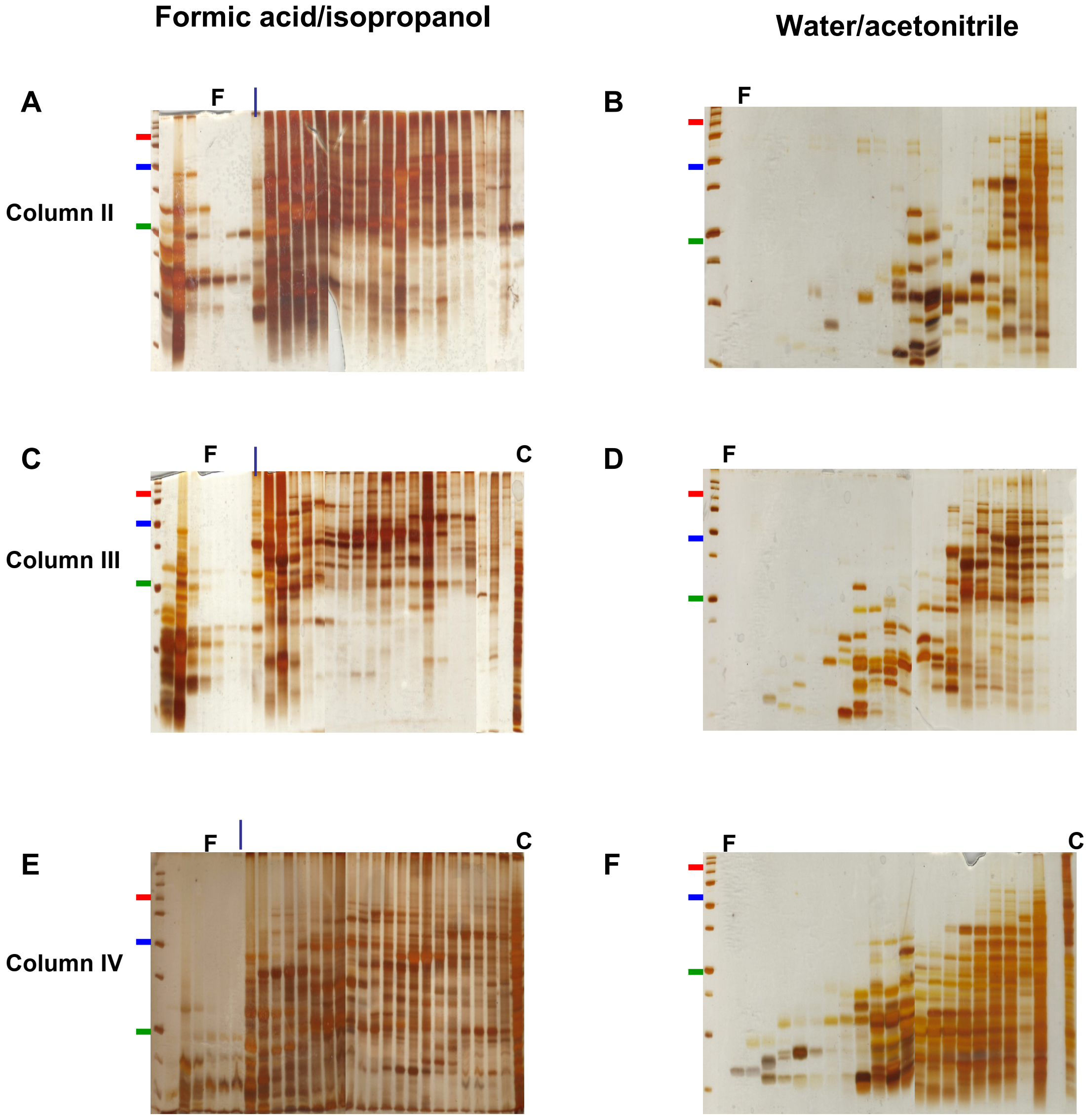


Figure 4.3 SDS-PAGE analysis of fractions collected from the formic acid/ isopropanol, or water/acetonitrile solvent system. Column II (C_4: A, B), column III (R_2: C, D) and column IV (polymer: E, F) were used for comparison of the two solvent systems. A, C and E, show fractions of yeast protein separated using the 60% formic acid/isopropanol gradient, while B, D and F, correspond to fractions collected using the water/acetonitrile gradient. Protein markers are displayed using 25 kDa (green), 50 kDa (blue) and 100 kDa (red). C represents a control (loading sample) and F represents the flow-through, respectively.

separation patterns are in fact quite similar across the three columns when employing a water/acetonitrile solvent system. In particular, a trend towards elution of protein based on molecular weight (MW) is evident in the water/acetonitrile gradient (gels B, D, and F). Badock *et al.* [48] reported a general observation of protein size being correlated with hydrophobicity, which is estimated by using the grand average of hydropathy (GRAVY) score [78]. To test whether this is true, the complete set of 5,802 yeast proteins was used to plot the protein's GRAVY score versus MW. Figure 4.4 demonstrates that there is no correlation of MW to hydrophobicity and as a consequence, the observed trend of size-based protein separation on a reversed phase column is an undesirable result.

Low resolution separation of proteins based on size by RPLC makes development of a 2D separation platform, coupling reversed phase to GeLC or GELFrEE less effective, therefore RPLC separations with water/acetonitrile gradients are determined to be ineffective. By contrast, the 60% formic acid/isopropanol gradient shows no trend of separation based on protein MW (Figure 4.3A, C and E), therefore using the formic acid/ isopropanol solvent system has the potential as a couple for second dimension separations by GeLC or GELFrEE. Protein separation efficiency between the three columns clearly demonstrated in the early fractions (i.e., the flow through or non-retained component of separation) where although the patterns are different, the protein separations between them are acceptable. A high number of unique protein bands are shown only in a single lane or two adjacent lanes for the three gels (Figure 4.3A, C and E). Figure 4.3 shows that using the formic acid/isopropanol solvent system the polystyrene column with smaller diameter beads (column IV) has poor resolution than either column II (C_4 beads) or

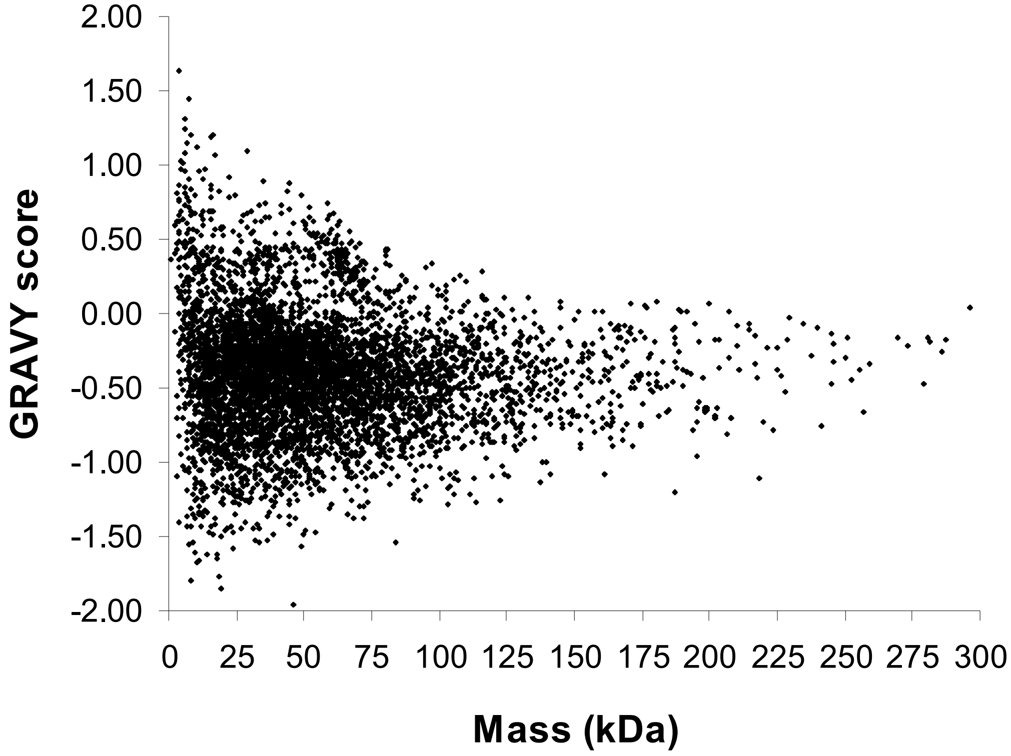


Figure 4.4 Correlation analysis of mass to GRAVY score of protein for the entire yeast proteome. Hydrophobicity of proteins is represented by GRAVY score.

column III (larger pore size polystyrene resin), shown by the number of proteins present in more than one lane of the SDS-PAGE gels. To determine which column performs best for proteome fractionation, a more thorough assessment is provided through LC-MS/MS analysis of the resulting protein fractions.

Protein recovery using the two solvent systems can also be assessed from Figure 4.3. Based on the approximate number of the stained bands, protein recovery from the formic acid/isopropanol solvent system is higher than the water/acetonitrile solvent system, across all three columns. Equivalent sample (150 μg of proteins) loads on columns II and III produced a greater number of protein bands in the gels of Figure 4.3A & C (60% formic acid/isopropanol gradient) than the gels of Figure 4.3B & D (water/ acetonitrile gradient). This difference is most noted between solvent systems using column IV, the biggest column with ID 2.0 mm). For this experiment, approximately four times more volume of sample was required to give the same loading amount for fractions from the water/acetonitrile solvent. These visual comparisons provide clear indication that more proteins are eluted from the column through use of the 60% formic acid/isopropanol gradient.

4.3.4 Characterization of Identified Proteome using LC-MS/MS

To fully estimate proteome separation using the 60% formic acid/isopropanol gradient, the fractionated proteins are identified and assessed by LC-MS/MS. From this assessment, the best column will be selected for comparison between the two solvent systems. To obtain a fair comparison, equivalent sample loading was employed for LC-MS/MS analysis. The identified proteins from the three column separations are shown in Figure 4.5 and Table 4.1.

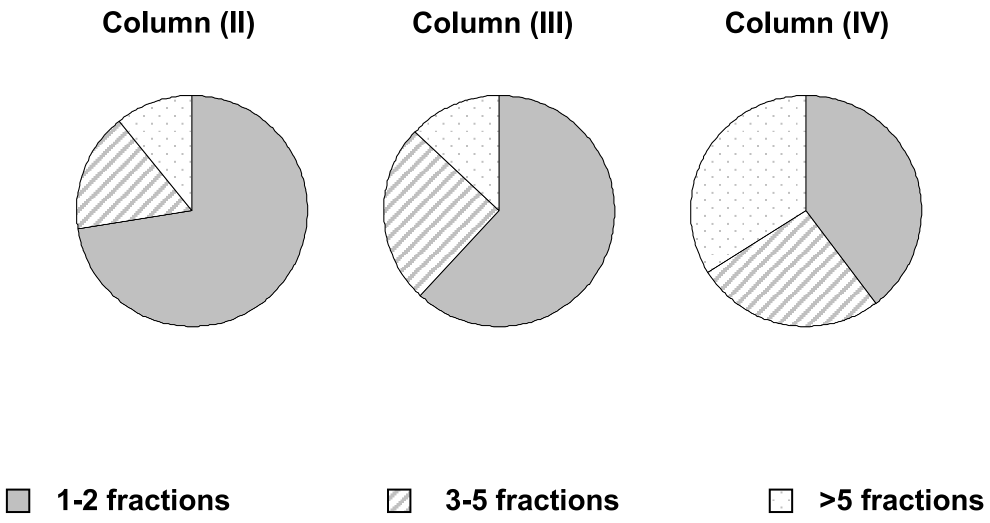


Figure 4.5 Pie charts displaying the degree of protein separation by the three columns. Yeast proteins were separated on the three RP columns by using the 60% formic acid/isopropanol solvent system. The grey indicates that proteins were eluted in one or two fractions. The dashed lines presents that proteins were spread on 3 to 5 fractions and the dots stand for less or no separation (spread in higher than 5 fractions). The C_4 or column (II) provides the best protein separation.

Table 4.1 Comparison of identifying proteins on the three RP columns [A]

Columns	# identified proteins	# unique proteins	# IMPs	# TMPs	% IMPs	% TMPs
II	1,384	410	282	80	20.7	5.8
III	1,020	86	189	48	18.5	4.7
IV	610	38	118	27	19.3	4.4

[A] Yeast proteins were extracted in 25 mM Tris-HCl (pH 7.65) and precipitated by the chloroform/methanol/water protocol.

TMPs represent alpha-helical transmembrane proteins.

As seen in Figure 4.5, column II provides the highest resolution of protein separation. Over two-thirds (72%) of the identified proteins were observed in one or two fractions of column II, as opposed to 62% observed for the column III and significantly less (40%) for the column IV. This is an indication that using the formic acid/isopropanol solvent systems, silica-based columns have higher efficiency than polymer-based columns. One would expect that better protein separation will result in a higher number of identified proteins. The total number of protein identifications, shown in Table 4.1, confirms this. Among the three columns, the silica-based column II gave the greatest number of identified proteins (1,384), compared to 1,020 proteins identified from column III and 610 proteins from column IV. The greatest number of identified IMPs/transmembrane proteins are obtained using column II, therefore, column II (C_4) is selected to be used for full comparison of the two solvent systems.

4.3.4.1 Effect of Digestion

Using the formic acid/isopropanol solvent system, 20.7% IMPs and 5.4% transmembrane proteins were identified from fractions collected from column II. As previously discussed, this is lower than the theoretical number (31.6%, 19.6%), and may be attributed to the MS detection platform, as opposed to being a consequence of ineffective separation of hydrophobic proteins. One such attribute of the detection platform relates to the required digestion step ahead of MS analysis. Trypsin digestion is a common procedure prior to MS; however, because it cleaves proteins at arginine and lysine residues, it may selectively target hydrophilic domains of proteins, and induce a bias against hydrophobic (transmembrane) protein identification. To digest hydrophobic proteins, CNBr digestion is often used prior to trypsin [108,137], CNBr selectively attacks at methionine residues, which is predicted to be more commonly associated with transmembrane alpha-helix (TMH) domains of transmembrane proteins [78] and may allow hydrophobic protein cleavage into shorter segments, allowing trypsin access to its target residues for further digestion. A direct comparison of proteins digested with CNBr/trypsin (1,012) and trypsin alone (1,429) demonstrates no increase in the total number of protein identifications (Figure 4.6). However, the CNBr/trypsin digestion did improve the percentage of transmembrane protein identifications from 6.1% to 5.5%. A total of 95 transmembrane proteins (combination) include 17 transmembrane proteins uniquely identified from the CNBr/trypsin digestion. In spite of this, due to the overall decrease in total proteins identified, it is concluded that the CNBr/trypsin digestion is not suitable for processing the prefractionated proteome. It is speculated that the reason behind the drop in identifications is due to the searching algorithm (SEQUEST) applied to the MS/MS spectra. Peptides that are generated from a combined CNBr/trypsin digestion introduce a lower degree of specificity into the search algorithm, which increases the number of possible (theoretical) peptide matches, thus lowering the confidence of a true identification.

4.3.5 Comparison of the Two Solvent Systems

From the gel images of Figure 4.3, the two solvent systems generated two different separation patterns on a silica-based C_4 column (II). A further assessment is provided through MS analysis of the resulting fractions. The numbers of identified proteins from the two solvent systems are summarized in Table 4.2.

As shown in Table 4.2, the number of identified proteins from the formic acid/isopropanol solvent system (910) is higher than those from the water/acetonitrile

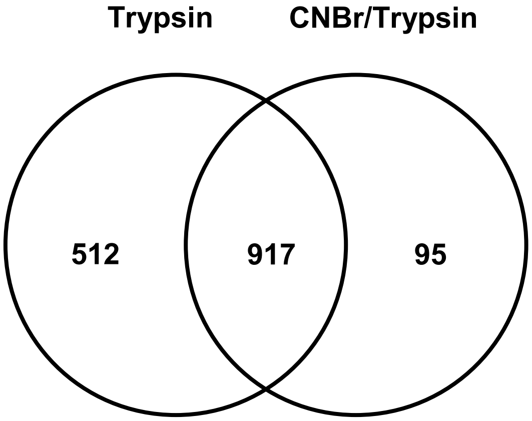


Figure 4.6 Venn diagram showing a comparison between trypsin and CNBr/trypsin digestion. The yeast proteins were fractionated on each of the three columns (II, III, IV) using the 60% formic acid/isopropanol gradient and 20 fractions for each column were collected. Each of the 20 fractions was divided by half. One half was digested by the trypsin method and the other half was used for the CNBr/trypsin method. All identified proteins by the trypsin digestion from the three columns are compared with all identified proteins from the CNBr/trypsin digestion.

Table 4.2 Comparison of protein separation on the two solvent systems.

Solvent systems	# identified proteins (peptides)	# IMPs	# TMPs	% IMPs	% TMPs
FA-IPA	910 (3,356)	193	54	21.2	5.9
Water-ACN	795 (2,602)	145	32	18.2	4.0

Notes: TMPs represent alpha-helical transmembrane proteins. FA-IPA represents 60% formic acid/isopropanol solvent system and water-ACN represents water/acetonitrile one.

solvent system (795). The difference in identified peptides from each solvent system favors the formic acid/ isopropanol (3.356 vs 2,602). Nonetheless, it is expected that these differences would have been greater still, had it not been for the "optimized" preparation of the water/acetonitrile sample (see Section 4.2.6). Specifically, a standard protocol would involve preparing the sample in the solvent starting conditions (i.e. 5% ACN with 0.1% TFA, see Chapter 3). However, such a solvent system has very poor protein recovery as described in Chapter 3. In this study, a 60% formic acid/water solvent system was used to prepare samples for loading on the RPLC column, allowing more proteins to be injected on the column and more proteins to be identified. Even using this sample preparation, the numbers of total identified proteins as well as identified IMPs/transmembrane proteins from the formic acid/isopropanol solvent system are still higher than the water/acetonitrile solvent system. Therefore, the formic acid/isopropanol solvent system appears to be a better choice for proteome fractionation.

Further evidence for formic acid/isopropanol solvent system superiority is that protein recovery of the formic/acid solvent system is higher than for the water/acetonitrile. The gel images of Figure 4.2 show a greater number of protein bands visualized following formic acid/isopropanol separation compared to acetonitrile/water separation. These results are confirmed by peptide quantification. Using each of the two solvent systems, the recovered peptides in each fraction were quantified by a RPLC-UV peptide assay as described in Section 4.2.8. A 67% total recovery of peptides was obtained from the formic acid/isopropanol solvent system versus 41% recovery using the water/acetonitrile solvent system. Higher protein recovery translates into improved MS identification, lending further evidence to formic acid/isopropanol solvents being a better solvent system for intact protein separation by RPLC than water/acetonitrile.

4.3.5.1 Protein Distribution on Fractions

One of the main goals of proteome prefractionation is to obtain a nearly even distribution of proteins into each fraction. Figure 4.7A shows that fractionated yeast proteins are evenly distributed across the 20 fractions using the formic acid/isopropanol solvent system, whereas the majority of proteins are eluted in the later fractions (9 to 20) with the water/acetonitrile solvent system. A more even distribution of proteins translates into a more even number of MS-identified proteins among all 20 fractions (Figure 4.7B).

For intact proteome prefractionation, it would be an added benefit that the proteins separate in a predictable fashion (i.e., according to mass, isoelectric point (pI) or hydrophobicity). Reversed phase separation predicts a trend according to hydrophobicity, which is indicated by the GRAVY score of the proteins. Figure 4.8A to C indicate that protein separation by both of the solvent systems is indeed based on hydrophobicity, with the formic acid/isopropanol solvent system resulting in the recovery of more hydrophobic proteins than the water/acetonitrile solvent system.

It is noted that RPLC protein separation is also correlated to the MW of the protein. As shown in Figure 4.8D-F, higher mass proteins are generally recovered as the separation progresses. Also, larger proteins are eluted by the formic acid/isopropanol than the water/acetonitrile solvent system making the formic acid/isopropanol solvent system better for eluting a wider protein mass range; however this relationship to size-based separation could have a negative impact on the second dimension of separation (SDS-PAGE).

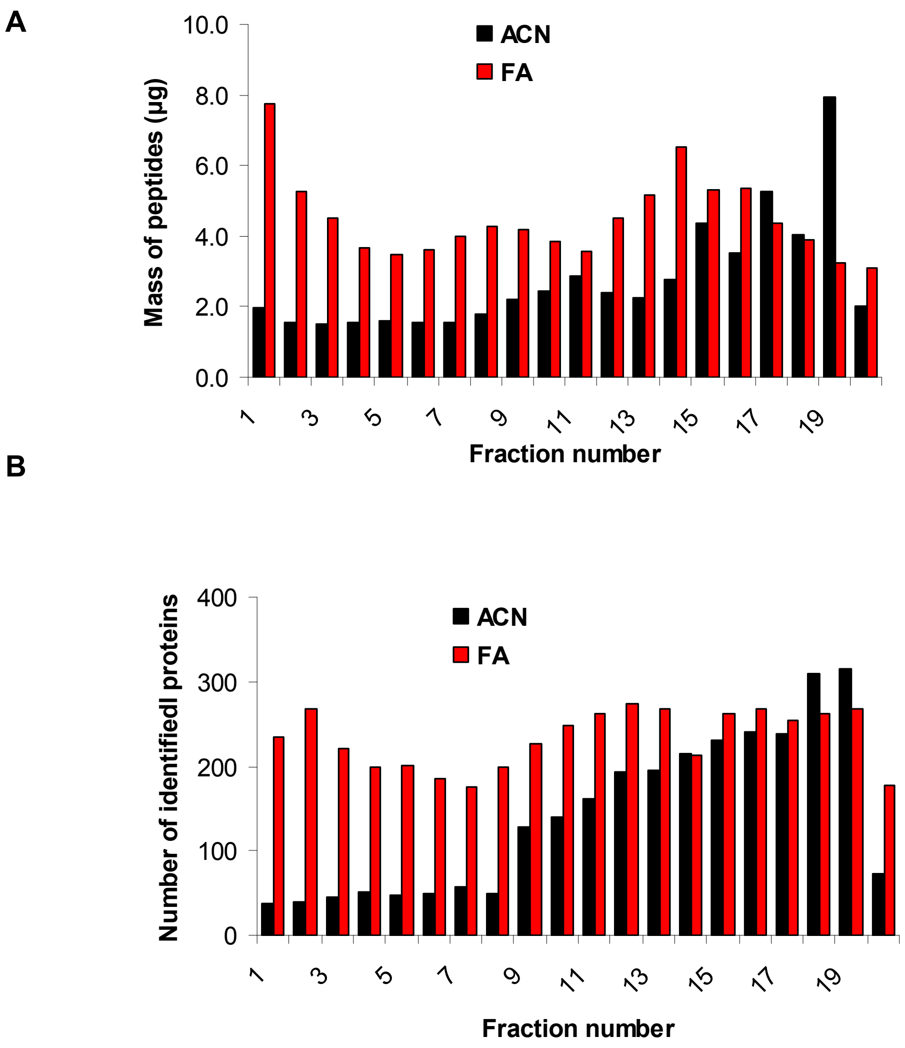


Figure 4.7 Assessing the two solvent systems on distribution of fractionated proteins across 20 fractions. (A) A histogram shows mass of peptides distribution over 20 fractions for the 60% formic acid/isopropanol gradient (FA, red) and the water/acetonitrile gradient (ACN, black). The column II was selected for the comparison. (B) A histogram shows the distribution of number of identified proteins in each fraction. Both figures indicate that the formic acid/isopropanol solvent system has more even distribution of number of identified proteins and mass of proteins (implying by mass of peptides) across 20 fractions than water/acetonitrile.

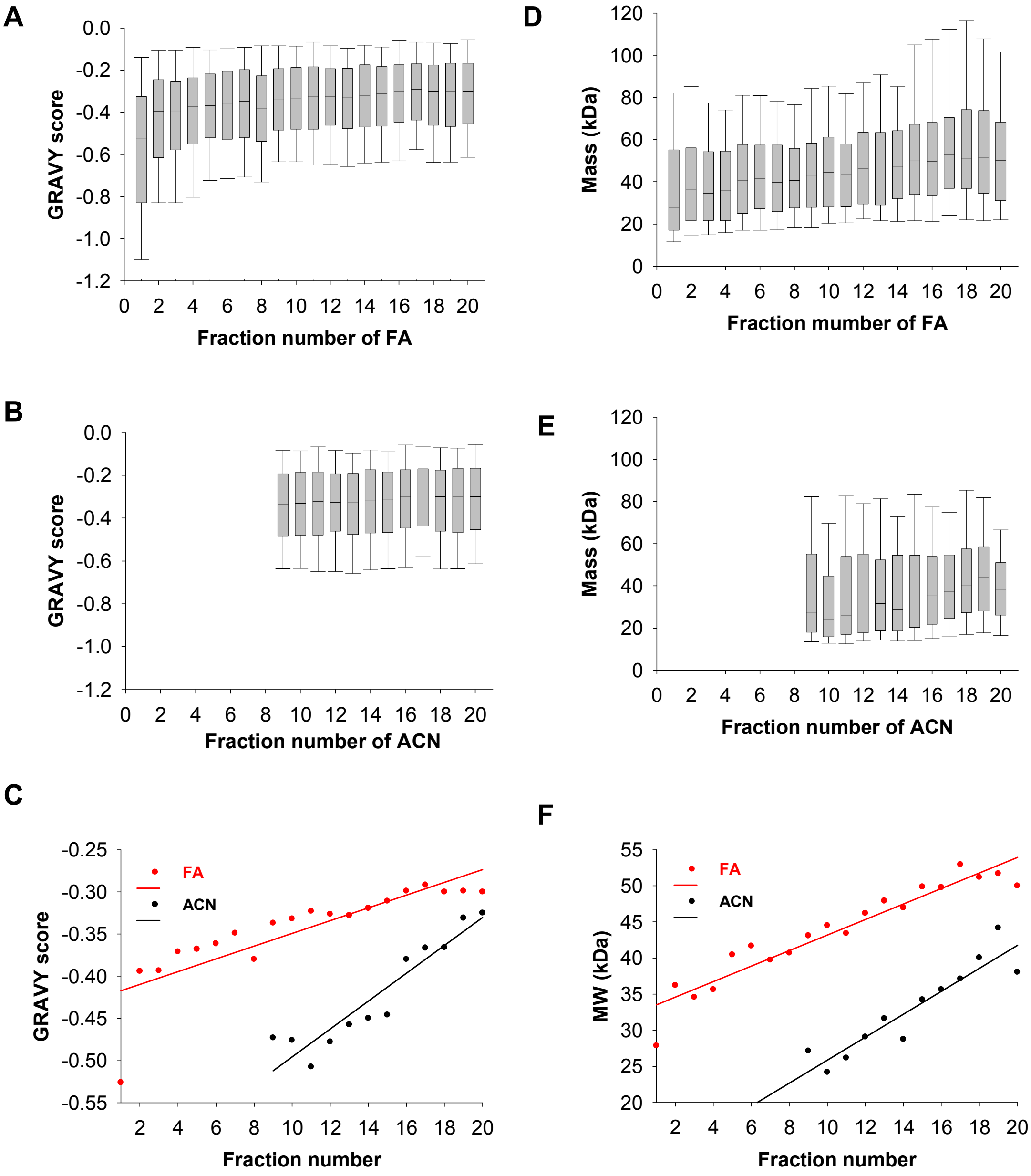


Figure 4.8 Comparison of the two solvent systems by elution profile. A box and whisker plot of the GRAVY and mass distributions of proteins identified in each fraction from (A, D) the formic acid/isopropanol solvent system (FA) and (B, E) the water-ACN solvent system (ACN). The plot was generated to represent the 5th/95th percentile (lower/upper whisker), and the 25th/75th percentile (box), with the median molecular weight per fraction displayed by the line in the box. A clear trend of the GRAVY (C) and mass (F) elution profiles between the two solvent systems is demonstrated, with the median GRAVY from A & B or median mass from D & E of identified proteins in each fraction plotted against the fraction number.

4.3.5.2 Mass Distribution

The formic acid/isopropanol solvent system preferentially separates moderate MW proteins, which is shown by the mass distribution of identified proteins (Figure 4.9A), while the water/acetonitrile solvent system shows a preference toward low-mass proteins (Figure 4.9A). According to chi-square (χ^2) tests, the mass distribution of proteins identified from the formic acid/isopropanol solvent system is more similar to that distribution from that of the entire yeast proteome.

4.3.5.3 GRAVY and pI Distributions

According to χ^2 tests, the GRAVY and pI distributions of the identified proteins from the formic acid/isopropanol solvent system is more similar to the theoretical yeast proteome than water/acetonitrile one. It is expected that protein solubility improves significantly using 60% formic acid (see Fig. 3.8). More proteins dissolved in the solution leads to more identified proteins which preferable to represent the classes of the proteome. In particular, hydrophobic proteins with GRAVY higher than 0.25 are still not identified using formic acid/isopropanol solvent system (see Fig. 4.9B). The identified proteins from this solvent system are perrable toward acidic proteins with pI less than 6.5 (see Fig. 4.9C).

4.3.5.4 Percentage of Identified Integral Membrane/Transmembrane Proteins

Although the percentages of IMPs/transmembrane proteins identified following separation with formic acid/isopropanol (21.2% and 5.9%) are lower than the predicted percentages from the yeast protein database (31.6% and 19.6%), separation with formic acid/isopropanol yields a higher percentages (or number) of identified proteins than with water/acetonitrile (18.2% and 4.0%). These numbers are consistent with the GRAVY

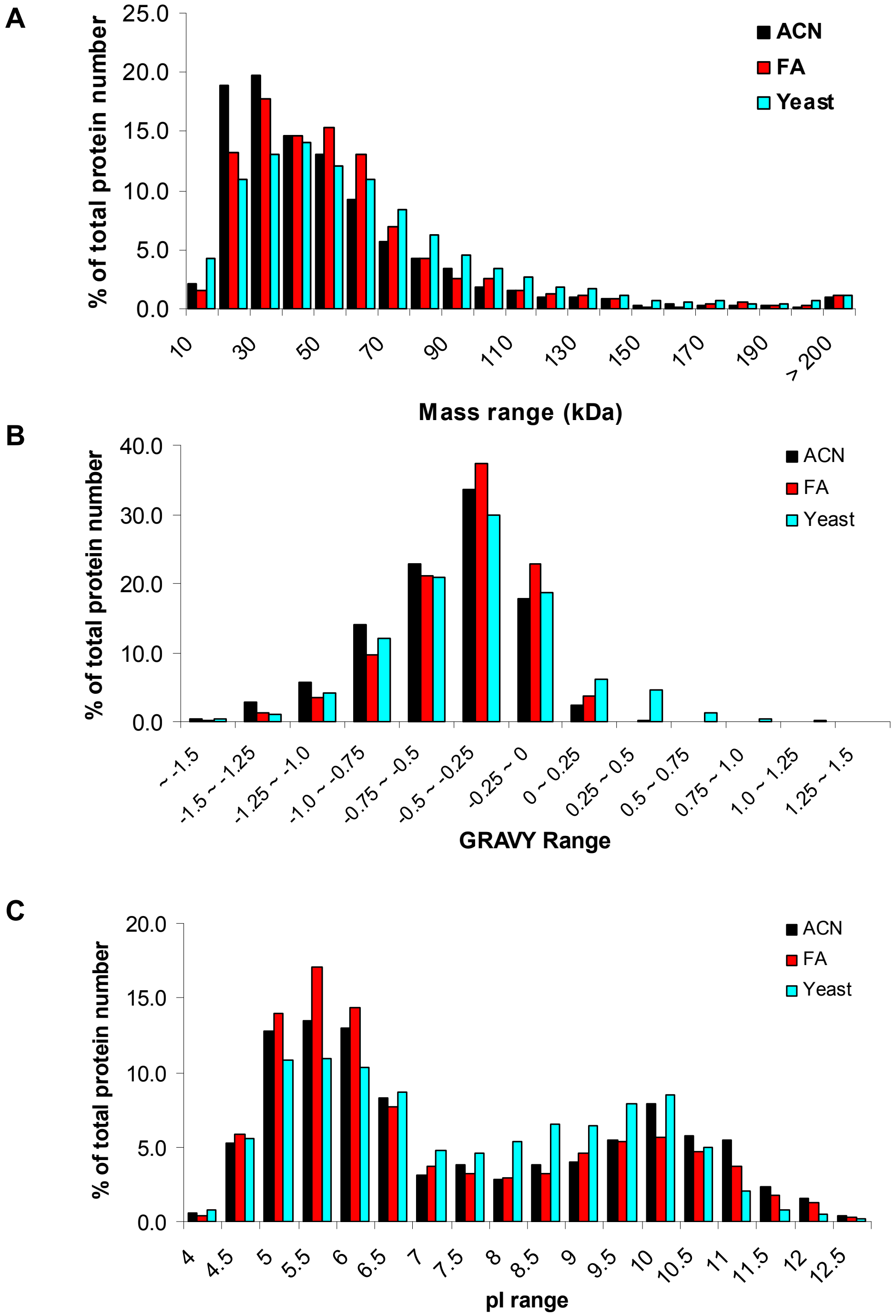


Figure 4.9 Comparison of the identified proteins from the two solvent systems based on (A) mass, (B) GRAVY score, and (C) and pI. Distributions of the identified proteins from the 60% formic acid/Isopropanol (FA) and the water/acetonitrile (ACN) gradients were compared with the theoretical yeast proteome (Yeast). Each distribution of identified proteins between the two solvent systems shows a similar trend.

distribution of identified proteins following separation with either solvent system wherein the identified transmembrane proteins only contain one or two transmembrane alpha helices (data not shown). The result is likely caused by hydrophobic protein retention on the column, which could account for 33% loss of protein masses. These results imply that even with the improved efficiency of RPLC separation with the formic acid/isopropanol solvent system, solution-based RPLC protein separation is still not an effective means for complete proteome fractionation.

4.3.5.5 Considerations of the RPLC Platform

Although RPLC efficacy is greatly improved with the formic acid/isopropanol solvent system, several considerations need to be mentioned. First, water/acetonitrile solvent system has greater resolution than formic acid/isopropanol, meaning that the percentage of proteins identified in one or two adjacent fractions in the water/acetonitrile solvent system (62%) is higher than one in the formic acid/isopropanol (48%). A potential explanation of this result is that the more simplified mixture causes the better separation, meaning that less interference among proteins results in improved separation [138]. Second, the efficiency of the C_4 column (column II) decreases quickly when using the formic acid/isopropanol solvent system. In the early experiments (see Table 4.1), 1,384 proteins were identified (72% in one or two adjacent fractions), but further testing showed only 910 proteins identified, with only 48% of proteins identified in one or two fractions. The decreased efficiency of separation is caused by accumulation of non-eluted proteins, which was reported by Whitelegge *et al.* [4]. Third, proteins in 60% formic acid may lead to chemical modifications such as formylation at serine or threonine residues [135]. The extent of formylation can be minimized by shortening the time spent in formic acid. For peptide identification, a smaller percentage of peptides (less than 5%) were modified at serine or threonine residues. This could have a negative impact on intact protein identification.

4.3.6 Solution vs Gel-based Platforms

Although the establishment of a protein separation protocol that separates proteins by RPLC using the formic acid/isopropanol solvent system improves performance over the conventional water/acetonirtile solvent system, the challenge of separation and identification of hydrophobic proteins remains.

As described in Chapter 3, the procedure of protein identification using solution-based platform (RPLC) includes three steps: (1) sample extraction, (2) protein precipitation, and (3) protein fractionation and identification. Therefore, the final identified proteins are dependant on the first two steps. As indicated in chapter 3, the extracted and precipitated proteins identified using GeLC MS/MS platform (a gel-based method shown in Figure 1.9) can be a representative sampling of the yeast proteome. The number of totally identified proteins using the GeLC MS/MS platform is 1,204 for the extract or 1,451 for protein precipitation. Both numbers are greater than the 910 proteins identified from RPLC using the formic acid/isopropanol solvent system.

In an early experiment, 1,384 proteins were identified by RPLC using the formic acid/isopropanol solvent system. This number was close to ones for the extract (1,204) or precipitation (1,451), indicating that the number of proteins identified using RPLC can be comparable with the GeLC method. Nonetheless, significant difference of hydrophobic protein (transmembrane protein) identification between RPLC and GeLC platforms is clearly shown. The percentages of identified IMPs/transmembrane proteins in the early

RPLC experiment (total 1,384 identified proteins) were 20.7% and 5.8%, similar to the 21.2% and 5.9% of the later experiment which identified 910 identified proteins. Both numbers obtained from RPLC separation are lower than the 27.0% and 13.0% for the sample extraction or 27.7% and 13.6% for the protein precipitation followed by GeLC. These numbers, although less than the theoretical numbers of IMPs and transmembrane proteins (31.6% and 19.6%) are still better than those obtained from RPLC.

The distribution of identified proteins from different protein separation methods, relative to that of a theoretical distribution of yeast proteins, can be evaluated in terms of their size, hydrophobicity, and pI.

4.3.6.1 Size

As shown in Chapter 3, the mass distributions of the extract and the precipitated proteins identified by GeLC MS/MS are similar with each other, and also similar to the RPLC platform (formic acid/isopropanol). A wide mass range of proteins not only separated by SDS gel-based platforms, but also prefractionated using the RPLC platform.

4.3.6.2 GRAVY & pI

Identified proteins from GeLC and RPLC platforms are biased towards hydrophilic proteins (see Figure 4.9B and Figure 3.7C). However, the degree of bias for RPLC platform is worse than GeLC platform according to χ^2 test of GRAVY distribution. It is clearly noted that none of proteins with GRAVY higher than 0.25 are identified from RPLC platform. This is agreement with higher percentages of identified hydrophobic proteins from GeLC than RPLC platform. The discrepancy of hydrophobic protein identification between GeLC and RPLC platforms are probably related to sample loss. All proteins are separated in a SDS-PAGE gel, but around one-third of a proteome is lost during RPLC separation (see Section 4.3.5). Transmembrane proteins prefractionated by RPLC platform are also lost when they are redissolved in 100 mM NH_4HCO_3 solution (for digestion) after evaporation of formic acid.

Identified proteins from GeLC and RPLC are also biased towards acidic proteins (see Figure 4.9A and Figure 3.7C). According to χ^2 test of pI distribution, the degree of bias for RPLC platform is worse than GeLC platform. Therefore, for a comprehensive protein analysis, GeLC is better than RPLC platform

4.4 Conclusion

The ultimate purpose of the present study was to establish the utility of RP separation as a method for fractionation of a complex protein mixture. In this study, the 60% formic acid/isopropanol solvent system was used for proteome separation and compared with the conventional solvent system (water/acetonitrile). The formic acid/isopropanol solvent system proved better than water/acetonitrile in that a greater number of proteins were eluted and identified, as well as more IMPs and transmembrane proteins. The most important difference between the two solvent systems is that proteins can be evenly distributed across 20 fractions which can facilitate subsequent protein separations by SDS-PAGE or a second dimension of LC. However, when compared to a GeLC MS/MS platform, the RPLC platform was less effective for total protein identification.

Chapter 5

Conclusions

5.1 Thesis Summary

The work presented in this thesis focuses on techniques for proteome prefractionation. Proteome separation and identification platforms were assessed, using a whole protein extract from yeast *S. cerevisiae* as the model complex protein sample. Considering that approximately 30% of proteins in yeast are integral membrane proteins (IMPs) and that these proteins are essential for fundamental cellular processes, identification of these proteins is of utmost importance in proteomics experiments. Most IMPs, especially alpha-helical transmembrane proteins (herein referred to as transmembrane proteins), are hydrophobic, and poor protein solubility causes problems in proteomic analysis. Thus, a comprehensive method for proteome fractionation and identification should consider identification of IMPs, especially transmembrane proteins.

For evaluation of any proteome prefractionation technique, a rigorous dataset of yeast IMPs/transmembrane proteins must be established. Current identification approaches employ computational prediction algorithms to determine if a protein is either a transmembrane or cytosolic protein. However, there is considerable disagreement among the prediction programs. Thus, one of the objectives of Chapter 2 was to develop a more reliable and rapid approach to classify transmembrane proteins from the yeast proteome. The established datasets of yeast IMPs/transmembrane proteins were used to assess the universal approaches of proteome fractionation and identification.

Because SDS is an efficient solubilizing agent of both hydrophilic and hydrophobic proteins, SDS-PAGE separation platforms are suggested to be compatible techniques for whole proteome characterization. GeLC MS/MS takes advantage of the separating efficiency of SDS-PAGE for identification of a wide range of proteins, both hydrophilic and hydrophobic [57]. GELFrEE MS/MS is a relatively new SDS-based intact protein separation method similar to SDS-PAGE that is able to separate proteins based on MW in solution. Treatment of samples in each platform differs greatly post separation, thus it was important to compare the two platforms for comprehensive proteome characterization. Chapter 2 directly compares these two platforms in a comprehensive proteomic analysis according to several parameters such as percentage of identified IMPs/transmembrane proteins. Identified proteins from the two platforms were assessed relative to a theoretical yeast protein database in terms of size, hydrophobicity (GRAVY), and isoelectric point (pI). A representative sampling profile of the yeast proteome obtained from GeLC or GELFrEE platforms was used to determine whether each platform was biased towards hydrophilic proteins. Based on the comparison shown in Chapter 2, the GELFrEE MS/MS matched the performance of GeLC MS/MS for proteome fractionation and identification.

Although the SDS-based platforms provide a useful tool for whole proteome characterization, these platforms are still difficult to automate and are limited in certain applications. An easily automated alternative to this separation platform involves solution-based separation by reversed phase liquid chromatography (RPLC). However, solution platforms present entirely new demands on sample preparation. In particular, SDS can no longer be used to solubilize the sample, as it is not compatible with the

RPLC. Sample preparation strategies for the solution-based separation platform generally involve the following: (1) cell disruption and proteome extraction; (2) protein cleanup; and (3) final resolubilization in a LC-MS compatible solvent. GeLC MS/MS was used to characterize the relative efficiency of each step (Chapter 3). The three characteristics of identified proteins (mentioned above) were chosen for assessing each sample preparation step using the datasets generated in Chapter 2. Once the optimum sample preparation protocol used for RPLC proteome prefractionation was selected, a prepared yeast proteome solution was ready for yeast proteome fractionation by RPLC (Chapter 4).

Chloroform/methanol/water and acetone protein precipitation methods were investigated in chapter 3 and found to be a suitable source of protein for fractionation testing due to the representative portion of the yeast proteome isolated.

Chapter 4 elucidated an alternative solvent system for RPLC for proteome separation. This solvent system (60% formic acid/isopropanol) was found to improve protein recovery and provide approximately even number of proteins across all the 20 collected fractions through comparing with the conventional solvent system. Also, the numbers of identified total/integral membrane/transmembrane proteins were increased using the formic acid/isopropanol solvent system. This solvent system was found to be suitable for proteome separation based on mass; however it was found to be ineffective for separation according to either hydrophobicity or isoelectric point. The inefficiency of the RPLC separation process makes gel-based methods still better than the solution-based method RPLC platform for a comprehensive protein analysis.

5.2 Future Work

Much of the future work surrounding hydrophobic protein identification needs to be investigated in order to obtain a representative profile of a proteome. As GeLC MS/MS and GELFrEE MS/MS platforms were assessed in Chapter 2, percentages of IMPs/transmembrane proteins were slightly lower than the theoretical numbers. Also, the current conditions of LC-MS/MS were proved to slightly prefer hydrophilic peptides. Therefore, comparing the GRAVY distribution of the entire yeast proteome to the identified proteins from the two SDS-based platforms, the SDS platforms showed a lower percentage of proteins in the GRAVY range 0.25 to 1.5. While the two SDS-based platforms were assessed to be comprehensive proteome analysis methods, an increased efficiency of hydrophobic protein identification (i.e., high number of identified transmembrane proteins) is needed in these methods.

Further experiments for this improvement would involve optimization of LC-MS/MS for hydrophobic protein identification. Several conditions of LC (RPLC for peptide separation) need to be investigated as follows: (1) different RP columns, e.g., monolithtic column (polymeric large-pore RP column) [139], (2) the gradients to elute very hydrophobic peptides, (3) elevated temperature (e.g., 30-80ºC) of a column to assist in hydrophobic peptide elution [108]. Settings of mass range in the linear ion trap (LIT) mass spectrometer may be adjusted to collect spectra of hydrophobic peptides corresponding to LC gradients. Thus, more hydrophobic proteins can be identified through more hydrophobic peptides identified. The other way is that an exclusion list obtained from the first run is used in the second run to find low abundant hydrophobic peptides. These optimal settings of LC-MS/MS for hydrophobic peptide identification might improve the SDS-based platform for proteome analysis.

Also, these optimal settings of LC-MS/MS could be used as post-column identification from protein separation by RPLC using 60% formic acid/isopropanol solvent system. In spite of this better performance of protein separation by this RPLC platform, this platform is still not as good as the SDS-based platforms for a comprehensive method. One of main reasons for biased proteome fractionation by this RPLC platform is non-eluted hydrophobic proteins (see Chapter 4). However, RPLC protein separation has advantages such as high-throughput, automation, high loading and coupling to MS. Thus, this RPLC platform is worthy of further investigation to improve the efficiency of proteome identification.

Additional work for improving RPLC protein separation will lie as following. First, an additional solvent system is used to remove very hydrophobic proteins from the column, e.g., 60% formic acid and 33% n-butanol in 60% formic acid solvent system [134] or chloroform (for generation of RP columns). Second, RP columns with macroporous pore size can be tested, e.g., monolithtic columns. These two are for very hydrophobic protein elution. The digestion step also needs to be considered. As described in Chapter 4, CNBr/trypsin digestion slightly improves transmembrane protein identification, but the number of total identified proteins decreases. An alternative digestion method (e.g., using Lys-C digestion or trypsin digestion with surfactants) increases percentage of identified IMPs/transmembrane proteins [59,140]. Finally, fractionated proteins by this RPLC platform are coupled to a GeLC MS/MS platform. Higher percentages of identified IMPs/transmembrane proteins might be achieved because of high resolution of 2D protein separation and hydrophobic protein identification by the latter platform. With all the efforts toward the improvement of hydrophobic protein identification, RPLC might be applied as an unbiased proteome fractionation method.

Appendix

1. The Theoretical Yeast Proteome

A complete yeast *Saccharomyces cerevisiae* proteome, 5,802 yeast proteins, can be downloaded from UniProt website (http://www.uniprot.org/). Clicking on ‘complete proteome sets’ icon in the UniProt web page and searching for yeast *S. cerevisiae* species, all yeast protein sequences with their accession numbers are downloaded and saved as a file with FASTA format.

2. Integral Membrane Proteins Obtained from a GO term

The dataset of integral membrane proteins of yeast is predicted according to a Gene Ontology annotation term through the UniProt website (http://www.uniprot.org/). In this website, ‘retrieve’ icon was firstly clicked so the yeast database (FASTA format) can be input under the ‘file’ icon of the new web page through browse function. Clicking on ‘retrieve’ icon again, all the proteins were filed to have a job name (i.e., job: xxxx), which is shown in the following web page. Under the ‘Query’ icon shown in the web page, ‘and GO: 0016020’ command (for proteins that directly interact with membrane) should be typed behind the job name in that window. Clicking ‘search’ icon, the ‘UniProtKB’ icon is shown in a new web page. By clicking this icon (UniProtKB), a list of the predicted integral membrane proteins of the yeast were displayed. Clicking on the ‘download’ icon, a dataset of predicted integral membrane proteins of the yeast is generated and can be saved in a proper format (e.g., Eexcel). The dataset of 1,832 integral membrane proteins for the yeast is created.

3. Transmembrane Proteins Predicted from the Four Selected Methods

Alpha-helical transmembrane proteins can be predicted through freely accessible prediction algorithms such as TMHMM Sever v 2.0 (http://www.cbs.dtu.dk/services/TMHMM/), Phobius (http://www.phobius.cbr.su.se/), HMMTOP (http://www.enzim.hu/hmmtop/) and SOSUI (http://bp.nuap.nagoya-u.ac.jp/sosui/sosuiG/). Prediction algorithms such as TMHMM or PHOBIUS are easey to use. The protein sequence file (FASTA format) can be input under a ‘sequence’ window. Clicking the ‘submit’ icon, the number of TMHs for a protein is quickly predicted. A list of all input proteins with zero (cytosolic proteins) or numbers of transmembrane helix or helices is provided. This list can be copied and pasted into a text file. Using the Microsoft Excel program, the number of TMHs and accession number of predicted transmembrane proteins are obtained.

Using HMMTOP and SOSUI algorithms, a maximum of protein sequences is set. Thus, each time around 500 protein sequences can be submitted in a ‘sequence’ window by pasting. The number of TMHs of a pasted protein is quickly predicted. These proteins submitted for prediction are discriminated by the number of predicted TMHs (zero or integer). This result was copied and pasted to a text file. While all yeast proteins are predicted, the number of TMHs and accession number of predicted transmembrane proteins were acquired using Microsoft Excel.

References

[1] Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; Oliver, S. G. Life with 6000 genes. *Science* **1996***, 274*, 546-567.

[2] Wu, C. C.; Yates, J. R. The application of mass spectrometry to membrane proteomics. *Nat. Biotechnol.* **2003***, 21*, 262-267.

[3] Gorg, A.; Weiss, W.; Dunn, M. J. Current two dimensional electrophoresis technology for proteomics [vol 4, pg 3665, 2004]. *Proteomics* **2005***, 5*, 826-827.

[4] Whitelegge, J. P.; Gomez, S. M.; Faull, K. F. Proteomics of membrane proteins. *Adv. Protein Chem.* **2003***, 65*, 271-307.

[5] Santoni, V.; Molloy, M.; Rabilloud, T. Membrane proteins and proteomics: Un amour impossible? *Electrophoresis* **2000***, 21*, 1054-1070.

[6] Tyers, M.; Mann, M. From genomics to proteomics. *Nature* **2003***, 422*, 193-197.

[7] Ullrich, B.; Ushkaryov, Y. A.; Sudhof, T. C. Cartography of neurexins - more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. *Neuron* **1995***, 14*, 497-507.

[8] Petsko, G. A.; Ringe, D. In *Protein Structure and Function;* New Science Press: London, 2004.

[9] Anderson, N. L.; Anderson, N. G. The human plasma proteome - History, character, and diagnostic prospects. *Mol. Cell. Proteomics* **2002***, 1*, 845-867.

[10] Ghaemmaghami, S.; Huh, W.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. Global analysis of protein expression in yeast. *Nature* **2003***, 425*, 737-741.

[11] Yates, J. R.; Ruse, C. I.; Nakorchevsky, A. Proteomics by mass spectrometry: approaches, advances, and applications. *Annu. Rev. Biomed. Eng.* **2009***, 11*, 49-79.

[12] Yamashita, M.; Fenn, J. B. Electrospray ion-source - another variation on the free-jet theme. *J. Phys. Chem.* **1984***, 88*, 4451-4459.

[13] Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. Protein and polymer analyses up to *m/z* 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **1988***, 2*, 151-153.

[14] Makarov, A. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Anal. Chem.* **2000***, 72*, 1156-1162.

[15] Makarov, A.; Denisov, E.; Kholomeev, A.; Baischun, W.; Lange, O.; Strupat, K.; Horning, S. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **2006***, 78*, 2113-2120.

[16] Cox, J.; Mann, M. Is proteomics the new genomics? *Cell* **2007***, 130*, 395-398.

[17] Domon, B.; Aebersold, R. Review - Mass spectrometry and protein analysis. *Science* **2006***, 312*, 212-217.

[18] Lin, D.; Tabb, D. L.; Yates, J. R. Large-scale protein identification using mass spectrometry. *Biochim. Biophys. Acta.* **2003***, 1646*, 1-10.

[19] Wilm, M.; Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **1996***, 68*, 1-8.

[20] Schmidt, A.; Karas, M.; Dulcks, T. Effect of different solution flow rates on analyte ion signals in nano-ESI MS, or: When does ESI turn into nano-ESI? *J. Am. Soc. Mass Spectrom.* **2003***, 14*, 492-500.

[21] Vanlear, G. E.; Mclaffer.Fw Biochemical Aspects of high-resolution mass spectrometry. *Annu. Rev. Biochem.* **1969***, 38*, 289-322.

[22] Biemann, K. Mass spectrometry. *Annu. Rev. Biochem.* **1963***, 32*, 755-780.

[23] James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **1993***, 195*, 58-64.

[24] Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **1993***, 3*, 327-332.

[25] Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007***, 4*, 787-797.

[26] Nilsson, T.; Mann, M.; Aebersold, R.; Yates,John R.,,III; Bairoch, A.; Bergeron, J. J. M. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **2010***, 7*, 681-685.

[27] Hunt, D. F.; Yates, J. R.; Shabanowitz, J.; Winston, S.; Hauer, C. R. Protein sequencing by tandem mass-spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **1986***, 83*, 6233-6237.

[28] Roepstorff, P.; Fohlman, J. Proposal for a common nomenclature for sequence ions in mass-spectra of peptides. *Biomed. Mass Spectrom.* **1984***, 11*, 601-601.

[29] Dongre, A. R.; Jones, J. L.; Somogyi, A.; Wysocki, V. H. Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model. *J. Am. Chem. Soc.* **1996***, 118*, 8365-8374.

[30] Cox, K. A.; Gaskell, S. J.; Morris, M.; Whiting, A. Role of the site of protonation in the low-energy decompositions of gas-phase peptide ions. *J. Am. Soc. Mass Spectrom.* **1996***, 7*, 522-531.

[31] Yalcin, T.; Khouw, C.; Csizmadia, I. G.; Peterson, M. R.; Harrison, A. G. Why are B ions stable species in peptide spectra? *J. Am. Soc. Mass Spectrom.* **1995***, 6*, 1165-1174.

[32] Yalcin, T.; Csizmadia, I. G.; Peterson, M. R.; Harrison, A. G. The structure and fragmentation of Bn [n≥3] ions in peptide spectra. *J. Am. Soc. Mass Spectrom.* **1996***, 7*, 233-242.

[33] Kinter, M.; Sherman, N. E. In *Protein Sequencing and Identification Using Tandem Mass Sepectrometry;* John Wiley & Sons: New York, 2000.

[34] Dookeran, N. N.; Yalcin, T.; Harrison, A. G. Fragmentation reactions of protonated alpha-amino acids. *J. Mass Spectrom.* **1996***, 31*, 500-508.

[35] Vachet, R. W.; Ray, K. L.; Glish, G. L. Origin of product ions in the MS/MS spectra of peptides in a quadrupole ion trap. *J. Am. Soc. Mass Spectrom.* **1998***, 9*, 341-344.

[36] Johnson, R. S.; Martin, S. A.; Biemann, K.; Stults, J. T.; Watson, J. T. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass-spectrometer - differentiation of leucine and isoleucine. *Anal. Chem.* **1987***, 59*, 2621-2625.

[37] Bairoch, A.; Boeckmann, B. The Swiss-Prot protein-sequence data-bank. *Nucleic Acids Res.* **1991***, 19*, 2247-2248.

[38] Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003***, 31*, 365-370.

[39] Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H. Z.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. L. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **2004***, 32*, D115-D119.

[40] Bairoch, A.; Consortium, U.; Bougueleret, L.; Altairac, S.; Amendolia, V.; Auchincloss, A.; Argoud-Puy, G.; Axelsen, K.; Baratin, D.; Blatter, M.; Boeckmann, B.; Bolleman, J.; Bollondi, L.; Boutet, E.; Quintaje, S. B.; Breuza, L.; Bridge, A.; deCastro, E.; Ciapina, L.; Coral, D.; Coudert, E.; Cusin, I.; Delbard, G.; Dornevil, D.; Roggli, P. D.; Duvaud, S.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gehant, S.; Farriol-Mathis, N.; Ferro, S.; Gasteiger, E.; Gateau, A.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hulo, N.; James, J.; Jimenez, S.; Jungo, F.; Junker, V.; Kappler, T.; Keller, G.; Lachaize, C.; Lane-Guermonprez, L.; Langendijk-Genevaux, P.; Lara, V.; Lemercier, P.; Le Saux, V.; Lieberherr, D.; Lima, T. d. O.; Mangold, V.; Martin, X.; Masson, P.; Michoud, K.; Moinat, M.; Morgat, A.; Mottaz, A.; Paesano, S.; Pedruzzi, I.; Phan, I.; Pilbout, S.; Pillet, V.; Poux, S.; Pozzato, M.; Redaschi, N.; Reynaud, S.; Rivoire, C.; Roechert, B.; Schneider, M.; Sigrist, C.; Sonesson, K.; Staehli, S.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Veuthey, A.; Yip, L.; Zuletta, L.; Apweiler, R.; Alam-Faruque, Y.; Antunes, R.; Barrell, D.; Binns, D.; Bower, L.; Browne, P.; Chan, W. M.; Dimmer, E.; Eberhardt, R.; Fedotov, A.; Foulger, R.; Garavelli, J.; Golin, R.; Horne, A.; Huntley, R.; Jacobsen, J.; Kleen, M.; Kersey, P.; Laiho, K.; Leinonen, R.; Legge, D.; Lin, Q.; Magrane, M.; Martin, M. J.; O'Donovan, C.; Orchard, S.; O'Rourke, J.; Patient, S.; Pruess, M.; Sitnov, A.; Stanley, E.; Corbett, M.; di Martino, G.; Donnelly, M.; Luo, J.; van Rensburg, P.; Wu, C.; Arighi, C.; Arminski, L.; Barker, W.; Chen, Y.; Hu, Z.; Hua, H.; Huang, H.; Mazumder, R.; McGarvey, P.; Natale, D. A.; Nikolskaya, A.; Petrova, N.; Suzek, B. E.; Vasudevan, S.; Vinayaka, C. R.; Yeh, L. S.; Zhang, J. The Universal Protein Resource [UniProt] 2009. *Nucleic Acids Res.* **2009***, 37*, D169-D174.

[41] MacCoss, M. J.; Wu, C. C.; Yates, J. R. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **2002***, 74*, 5593-5599.

[42] Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999***, 20*, 3551-3567.

[43] Eng, J. K.; Mccormack, A. L.; Yates, J. R. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994***, 5*, 976-989.

[44] Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2005***, 2*, 667-675.

[45] Washburn, M. P.; Wolters, D.; Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001***, 19*, 242-247.

[46] Zolotarjova, N.; Mrozinski, P.; Chen, H.; Martosella, J. Combination of affinity depletion of abundant proteins and reversed-phase fractionation in proteomic analysis of human plasma/serum. *J. Chromatogr. A* **2008***, 1189*, 332-338.

[47] Righetti, P. G.; Castagna, A.; Herbert, B.; Reymond, F.; Rossier, J. S. Prefractionation techniques in proteome analysis. *Proteomics* **2003***, 3*, 1397-1407.

[48] Badock, V.; Steinhusen, U.; Bommert, K.; Otto, A. Prefractionation of protein samples for proteome analysis using reversed-phase high-performance liquid chromatography. *Electrophoresis* **2001***, 22*, 2856-2864.

[49] Ofarrell, P. H. High-resolution 2-dimensional electrophoresis of proteins. *J. Biol. Chem.* **1975***, 250*, 4007-4021.

[50] Gorg, A.; Weiss, W.; Dunn, M. J. Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **2004***, 4*, 3665-3685.

[51] Klose, J.; Kobalz, U. 2-Dimensional electrophoresis of proteins - an updated protocol and implications for a functional-analysis of the genome. *Electrophoresis* **1995***, 16*, 1034-1059.

[52] Sedmak, J. J.; Grossberg, S. E. Rapid, sensitive, and versatile assay for protein using coomassie brilliant blue G250. *Anal. Biochem.* **1977***, 79*, 544-552.

[53] Blum, H.; Beier, H.; Gross, H. J. Improved silver staining of plant-proteins, RNA and DNA in polyacrylamide gels. *Electrophoresis* **1987***, 8*, 93-99.

[54] Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels. *Anal. Chem.* **1996***, 68*, 850-858.

[55] Gygi, S. P.; Corthals, G. L.; Zhang, Y.; Rochon, Y.; Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. U. S. A.* **2000***, 97*, 9390-9395.

[56] Schirle, M.; Heurtier, M. A.; Kuster, B. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2003***, 2*, 1297-1305.

[57] de Godoy, L. M. F.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Froehlich, F.; Walther, T. C.; Mann, M. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **2008***, 455*, 1251-1254.

[58] Reynolds, J. A.; Tanford, C. The gross conformation of protein-sodium dodecyl sulfate complexes. *J. Biol. Chem.* **1970***, 245*, 5161-5165.

[59] Speers, A. E.; Wu, C. C. Proteomics of integral membrane proteins-theory and application. *Chem. Rev.* **2007***, 107*, 3687-3714.

[60] Havlis, J.; Thomas, H.; Sebela, M.; Shevchenko, A. Fast-response proteomics by accelerated in-gel digestion of proteins. *Anal. Chem.* **2003***, 75*, 1300-1306.

[61] Shevchenko, A.; Tomas, H.; Havlis, J.; Olsen, J. V.; Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **2006***, 1*, 2856-2860.

[62] Tran, J. C.; Doucette, A. A. Gel-eluted liquid fraction entrapment electrophoresis: An electrophoretic method for broad molecular weight range proteome separation. *Anal. Chem.* **2008***, 80*, 1568-1573.

[63] Wall, D. B.; Kachman, M. T.; Gong, S. Y. S.; Parus, S. J.; Long, M. W.; Lubman, D. M. Isoelectric focusing nonporous silica reversed-phase high-performance liquid chromatography/electrospray ionization time-of-flight mass spectrometry: a three-dimensional liquid-phase protein separation method as applied to the human erythroleukemia cell-line. *Rapid Commun. Mass Spectrom.* **2001***, 15*, 1649-1661.

[64] Nilsson, C. L.; Davidsson, P. New separation tools for comprehensive studies of protein expression by mass spectrometry. *Mass Spectrom. Rev.* **2000***, 19*, 390-397.

[65] Whitelegge, J. P.; Gundersen, C. B.; Faull, K. F. Electrospray-ionization mass spectrometry of intact intrinsic membrane proteins. *Protein Sci.* **1998***, 7*, 1423-1430.

[66] Lee, R. P.; Doughty, S. W.; Ashman, K.; Walker, J. Purification of hydrophobic integral membrane proteins from Mycoplasma hyopneumoniae by reversed-phase high-performance liquid chromatography. *J. Chromatogr. A* **1996***, 737*, 273-279.

[67] Heukeshoven, J.; Dernick, R. Reversed-phase high-performance liquid-chromatography of virus proteins and other large hydrophobic proteins in formic-acid containing solvents. *J. Chromatogr.* **1982***, 252*, 241-254.

[68] Wildner, G. F.; Fiebig, C.; Dedner, N.; Meyer, H. E. The use of HPLC for the purification of the Qb-protein. *Z. Naturforsch. C: J. Biosci.* **1987***, 42*, 739-741.

[69] Gohshi, T.; Shimada, M.; Kawahire, S.; Imai, N.; Ichimura, T.; Omata, S.; Horigome, T. Molecular cloning of mouse p47, a second group mammalian RuvB DNA helicase-like protein: Homology with those from human and Saccharomyces cerevisiae. *J. Biochem.* **1999***, 125*, 939-946.

[70] Garrels, J. I.; Futcher, B.; Kobayashi, R.; Latter, G. I.; Schwender, B.; Volpe, T.; Warner, J. R.; Mclaughlin, C. S. Protein identifications for a saccharomyces-cerevisiae protein database. *Electrophoresis* **1994***, 15*, 1466-1486.

[71] Skou, J. C. Enzymatic basis for active transport of Na+ and K+ across cell membrane. *Physiol. Rev.* **1965***, 45*, 596-617.

[72] Ott, C. M.; Lingappa, V. R. Integral membrane protein biosynthesis: why topology is hard to predict. *J. Cell. Sci.* **2002***, 115*, 2003-2009.

[73] Torres, J.; Stevens, T. J.; Samso, M. Membrane proteins: the 'Wild West' of structural biology. *Trends Biochem. Sci.* **2003***, 28*, 137-144.

[74] Chou, K. C.; Elrod, D. W. Prediction of membrane protein types and subcellular locations. *Proteins* **1999***, 34*, 137-153.

[75] Elofsson, A.; von Heijne, G. Membrane protein structure: Prediction versus reality. *Annu. Rev. Biochem.* **2007***, 76*, 125-140.

[76] Pedersen, S. K.; Harry, J. L.; Sebastian, L.; Baker, J.; Traini, M. D.; McCarthy, J. T.; Manoharan, A.; Wilkins, M. R.; Gooley, A. A.; Righetti, P. G.; Packer, N. H.; Williams, K. L.; Herbert, B. R. Unseen proteome: Mining below the tip of the iceberg to find low abundance and membrane proteins. *J. Proteome Res.* **2003***, 2*, 303-311.

[77] Wallin, E.; von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **1998***, 7*, 1029-1038.

[78] Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982***, 157*, 105-132.

[79] Zinser, E.; Daum, G. Isolation and biochemical-characterization of organelles from the yeast, Saccharomyces-cerevisiae. *Yeast* **1995***, 11*, 493-536.

[80] Bodzon-Kulakowska, A.; Bierczynska-Krzysik, A.; Dylag, T.; Drabik, A.; Suder, P.; Noga, M.; Jarzebinska, J.; Silberring, J. Methods for samples preparation in proteomic research. *J. Chromatogr. B-Analyt. Technol. Biomed. Life Sci.* **2007***, 849*, 1-31.

[81] Simpson, K. L.; Wilson, A. W.; Nakayama, T. O.; Chichester, C. O.; Burton, E. Modified French press for disruption of microorganisms. *J. Bacteriol.* **1963***, 86*, 1126-1127.

[82] Barritault, D.; Expertbezancon, A.; Guerin, M. F.; Hayes, D. Use of acetone precipitation in isolation of ribosomal-proteins. *Eur. J. Biochem.* **1976***, 63*, 131-135.

[83] Lemaire, M.; Deschamps, S.; Moller, J. V.; Lecaer, J. P.; Rossier, J. Electrospray-ionization mass-spectroscopy on hydrophobic peptides electroeluted from sodium dodecyl-sulfate polyacrylamide-gel electrophoresis application to the topology of the sarcoplasmic-reticulum Ca2+ ATPase. *Anal. Biochem.* **1993***, 214*, 50-57.

[84] Wessel, D.; Flugge, U. I. A method for the quantitative recovery of protein in dilute-solution in the presence of detergents and lipids. *Anal. Biochem.* **1984***, 138*, 141-143.

[85] Puchades, M.; Westman, A.; Blennow, K.; Davidsson, P. Removal of sodium dodecyl sulfate from protein samples prior to matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **1999***, 13*, 344-349.

[86] Ishihama, Y. Proteomic LC-MS systems using nanoscale liquid chromatography with tandem mass spectrometry. *J. Chromatogr. A* **2005***, 1067*, 73-83.

[87] Kang, A. H. Studies on location of intermolecular crosslinks in collagen - isolation of a CNBr peptide containing delta-hydroxylysinonorleucine. *Biochemistry* **1972***, 11*, 1828-1835.

[88] Zhang, X. Y.; Dillen, L.; Vanhoutte, K.; VanDongen, W.; Esmans, E.; Claeys, M. Characterization of unstable intermediates and oxidized products formed during cyanogen bromide cleavage of peptides and proteins by electrospray mass spectrometry. *Anal. Chem.* **1996***, 68*, 3422-3430.

[89] Graham, R. L. J.; O'Loughlin, S. N.; Pollock, C. E.; Ternan, N. G.; Weatherly, D. B.; Jackson, P. J.; Tarleton, R. L.; McMullan, G. A combined shotgun and multidimensional proteomic analysis of the insoluble subproteome of the obligate thermophile, Geobacillus thermoleovorans T80. *J. Proteome Res.* **2006***, 5*, 2465-2473.

[90] Rahbar, A. M.; Fenselau, C. Unbiased examination of changes in plasma membrane proteins in drug resistant cancer cells. *J. Proteome Res.* **2005***, 4*, 2148-2153.

[91] Weinglass, A. B.; Whitelegge, J. P.; Kaback, H. R. Integrating mass spectrometry into membrane protein drug discovery. *Curr. Opin. Drug Discov. Devel.* **2004***, 7*, 589-599.

[92] Jones, D. T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **2007***, 23*, 538-544.

[93] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000***, 28*, 235-242.

[94] Bowie, J. U. Solving the membrane protein folding problem. *Nature* **2005***, 438*, 581-589.

[95] Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G.; Gene Ontology Consortium Gene Ontology: tool for the unification of biology. *Nat. Genet.* **2000***, 25*, 25-29.

[96] Moller, S.; Croning, M. D. R.; Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **2001***, 17*, 646-653.

[97] Nilsson, J.; Persson, B.; von Heijne, G. Prediction of partial membrane protein topologies using a consensus approach. *Protein Sci.* **2002***, 11*, 2974-2980.

[98] Huang, X. D.; Jack, M. A. hidden Markov modeling of speech based on a semicontinuous model. *Electron. Lett.* **1988***, 24*, 6-7.

[99] Rabiner, L. R. A tutorial on hidden Markov-models and selected applications in speech recognition. *Proc IEEE* **1989***, 77*, 257-286.

[100] Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **2001***, 305*, 567-580.

[101] Tusnady, G. E.; Simon, I. Principles governing amino acid composition of integral membrane proteins: Application to topology prediction. *J. Mol. Biol.* **1998***, 283*, 489-506.

[102] Tusnady, G. E.; Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **2001***, 17*, 849-850.

[103] Kall, L.; Krogh, A.; Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **2004***, 338*, 1027-1036.

[104] Hirokawa, T.; Boon-Chieng, S.; Mitaku, S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **1998***, 14*, 378-379.

[105] Botelho, D.; Wall, M. J.; Vieira, D. B.; Fitzsimmons, S.; Liu, F.; Doucette, A. top-down and bottom-up proteomics of SDS-containing solutions following mass-based separation. *J. Proteome Res.* **2010***, 9*, 2863-2870.

[106] Amico, M.; Finelli, M.; Rossi, I.; Zauli, A.; Elofsson, A.; Viklund, H.; von Heijne, G.; Jones, D.; Krogh, A.; Fariselli, P.; Martelli, P. L.; Casadio, R. PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. *Nucleic Acids Res.* **2006***, 34*, W169-W172.

[107] Sonnhammer, E. L.; von Heijne, G.; Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1998***, 6*, 175-82.

[108] Speers, A. E.; Blackler, A. R.; Wu, C. C. Shotgun analysis of integral membrane proteins facilitated by elevated temperature. *Anal. Chem.* **2007***, 79*, 4613-4620.

[109] Wang, H.; Hanash, S. Intact-protein based sample preparation strategies for proteome analysis in combination with mass spectrometry. *Mass Spectrom. Rev.* **2005***, 24*, 413-426.

[110] Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Na. Methods* **2009***, 6*, 359-362.

[111] Zhang, N.; Chen, R.; Young, N.; Wishart, D.; Winter, P.; Weiner, J. H.; Li, L. Comparison of SDS- and methanol-assisted protein solubilization and digestion methods for Escherichia coli membrane proteome analysis by 2-D LC-MS/MS. *Proteomics* **2007***, 7*, 484-493.

[112] Rundlett, K. L.; Armstrong, D. W. Mechanism of signal suppression by an ionic surfactants in capillary electrophoresis electrospray ionization mass spectrometry. *Anal. Chem.* **1996***, 68*, 3493-3497.

[113] Shi, Y.; Xiang, R.; Horvath, C.; Wilkins, J. A. The role of liquid chromatography in proteomics. *J. Chromatogr. A* **2004***, 1053*, 27-36.

[114] Opiteck, G. J.; Ramirez, S. M.; Jorgenson, J. W.; Moseley, M. A. Comprehensive two-dimensional high-performance liquid chromatography for the isolation of overexpressed proteins and proteome mapping. *Anal. Biochem.* **1998**, *258*, 349-361.

[115] Lescuyer, P.; Hochstrasser, D. F.; Sanchez, J. C. Comprehensive proteome analysis by chromatographic protein prefractionation. *Electrophoresis* **2004**, *25*, 1125-1135.

[116] Whitelegge, J.; Halgand, F.; Souda, P.; Zabrouskov, V. Top-down mass spectrometry of integral membrane proteins. *Expert Rev. Proteomics* **2006**, *3*, 585-596.

[117] Lubman, D. M.; Kachman, M. T.; Wang, H. X.; Gong, S. Y.; Yan, F.; Hamler, R. L.; O'Neil, K. A.; Zhu, K.; Buchanan, N. S.; Barder, T. J. Two-dimensional liquid separations-mass mapping of proteins from human cancer cell lysates. *J. Chromatogr. B-Analyt. Technol. Biomed. Life Sci.* **2002**, *782*, 183-196.

[118] Laemmli, U. K. Cleavage of structural proteins during assembly of head of bacteriophage-T4. *Nature* **1970**, *227*, 680-685.

[119] Horvath, A.; Riezman, H. Rapid protein extraction from Saccharomyces-cerevisiae. *Yeast* **1994**, *10*, 1305-1310.

[120] Taylor, M. T.; Belgrader, P.; Furman, B. J.; Pourahmadi, F.; Kovacs, G. T. A.; Northrup, M. A. Lysing bacterial spores by sonication through a flexible interface in a microfluidic system. *Anal. Chem.* **2001**, *73*, 492-496.

[121] Kushnirov, V. V. Rapid and reliable protein extraction from yeast. *Yeast* **2000**, *16*, 857-860.

[122] Kadiyala, C. S. R.; Tomechko, S. E.; Miyagi, M. Perfluorooctanoic acid for shotgun proteomics. *Plos One* **2010**, *5*, e15332.

[123] Soulie, S.; Denoroy, L.; Le Caer, J. P.; Hamasaki, N.; Groves, J. D.; le Maire, M. Treatment with crystalline ultra-pure urea reduces the aggregation of integral membrane proteins without inhibiting N-terminal sequencing. *J. Biochem.* **1998**, *124*, 417-420.

[124] Wang, N.; MacKenzie, L.; De Souza, A. G.; Zhong, H.; Goss, G.; Li, L. Proteome profile of cytosolic component of zebrafish liver generated by LC-ESI MS/MS combined with trypsin digestion and microwave-assisted acid hydrolysis. *J. Proteome Res.* **2007**, *6*, 263-272.

[125] Visser, N.; Lingeman, H.; Irth, H. Sample preparation for peptides and proteins in biological matrices prior to liquid chromatography and capillary zone electrophoresis. *Anal. Bioanal. Chem.* **2005***, 382*, 535-558.

[126] Jiang, L.; He, L.; Fountoulakis, M. Comparison of protein precipitation methods for sample preparation prior to proteomic analysis. *J. Chromatogr. A* **2004***, 1023*, 317-320.

[127] Yates, J. R. Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct.* **2004***, 33*, 297-316.

[128] Fonslow, B. R.; Yates,John R.,,III Capillary electrophoresis applied to proteomic analysis. *J. Sep. Sci.* **2009***, 32*, 1175-1188.

[129] Zhang, X.; Fang, A.; Riley, C. P.; Wang, M.; Regnier, F. E.; Buck, C. Multi-dimensional liquid chromatography in proteomics-A review. *Anal. Chim. Acta* **2010***, 664*, 101-113.

[130] Tran, J. C.; Wall, M. J.; Doucette, A. A. Evaluation of a solution isoelectric focusing protocol as an alternative to ion exchange chromatography for charge-based proteome prefractionation. *J. Chromatogr. B-Analyt. Technol. Biomed. Life Sci.* **2009***, 877*, 807-813.

[131] Martosella, J.; Zolotarjova, N.; Liu, H. B.; Nicol, G.; Boyes, B. E. Reversed-phase high-performance liquid chromatographic prefractionation of immunodepleted human serum proteins to enhance mass spectrometry identification of lower-abundant proteins. *J. Proteome Res.* **2005***, 4*, 1522-1537.

[132] Neverova, I.; Van Eyk, J. E. Application of reversed phase high performance liquid chromatography for subproteomic analysis of cardiac muscle. *Proteomics* **2002***, 2*, 22-31.

[133] Martosella, J.; Zolotarjova, N.; Liu, H. B.; Moyer, S. C.; Perkins, P. D.; Boyes, B. E. High recovery HPLC separation of lipid rafts for membrane proteome analysis. *J. Proteome Res.* **2006***, 5*, 1301-1312.

[134] Segawa, M.; Niino, K.; Mineki, R.; Kaga, N.; Murayama, K.; Sugimoto, K.; Watanabe, Y.; Furukawa, K.; Horigome, T. Proteome analysis of a rat liver nuclear insoluble protein fraction and localization of a novel protein, ISP36, to compartments in the interchromatin space. *FEBS J.* **2005***, 272*, 4327-4338.

[135] Loo, R. R. O.; Loo, J. A. Matrix-assisted laser desorption/ionization-mass spectrometry of hydrophobic proteins in mixtures using formic acid, perfluorooctanoic acid, and sorbitol. *Anal. Chem.* **2007***, 79*, 1115-1125.

[136] Wang, N.; Xie, C.; Young, J. B.; Li, L. Off-line two-dimensional liquid chromatography with maximized sample loading to reversed-phase liquid Chromatography-Electrospray Ionization Tandem Mass Spectrometry for shotgun proteome analysis. *Anal. Chem.* **2009***, 81*, 1049-1060.

[137] Kraft, P.; Mills, J.; Dratz, E. Mass spectrometric analysis of cyanogen bromide fragments of integral membrane proteins at the picomole level: Application to rhodopsin. *Anal. Biochem.* **2001***, 292*, 76-86.

[138] Winnik, W. M.; Ortiz, P. A. Proteomic analysis optimization: Selective protein sample on-column retention in reverse-phase liquid chromatography. *J. Chromatogr. B-Analyt. Technol. Biomed. Life Sci.* **2008***, 875*, 478-486.

[139] Josic, D.; Clifton, J. G. Use of monolithic supports in proteomics technology. *J. Chromatogr. A* **2007***, 1144*, 2-13.

[140] Chen, E. I.; Cociorva, D.; Norris, J. L.; Yates, J. R. Optimization of mass spectrometry-compatible surfactants for shotgun proteomics. *J. Proteome Res.* **2007***, 6*, 2529-2538.