AQUEOUS-BASED ASSOCIATED SOLVENTS FOR PROTEIN AND PEPTIDE FRACTIONATION AND PROTEIN PURIFICATION

By

SAJAD TASHAROFI

DISSERTATION

Submitted in partial fulfillment of requirements
For the degree of Doctor of Philosophy at
The University of Texas at Arlington
December 2021

Arlington, TX

Supervising committee:

Dr. Morteza G. Khaledi, Supervising Professor
Dr. Saiful M. Chowdhury
Dr. Kayunta Johnson-Winters
Dr. Daniel W. Armstrong

This thesis is dedicated to my family

for their endless love, support, and encouragement

and

to my son

Maxwell Tasharofi

for bringing joy to our lives

ACKNOWLEDGMENTS

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

I would first like to thank my supervisor, Dr. Morteza G. Khaledi, for his generous support and guidance throughout my PhD studies. His encouragement for learning new things compelled me to further study areas of science that is not trivial for a PhD student tin the field of chemistry.

I would also like to thank my committee members, Dr. Saiful Chowdhury, Dr. Kayunta Johnson-Winters and Dr. Daniel W. Armstrong for their advice and guidance. My special thanks to Dr. Kayunta Johnson-Winters for her generous support and advice.

In addition, I would like to thank my family and parents for their wise counsel and sympathetic ear. You are always there for me. I would like to extend my sincere thanks to my wife Brandi Stephens for her endless support during my hard work in the graduate school by providing encouragement, empowerment, and emotional support.

I very much appreciate the assistance of my friend Mohammadmehdi Azizi in my personal life and with my dissertation research. Also, the support of my friends, Mohsen Talebi, Clyde and Mary Lynne Hudson, and Chris and Lezlie Langford.

Finally, I would like to thank my colleagues, Durga Khanal, Armin Oloumi, Dorra Talili, Shashank Kumbhare and Mohammad Al Amin for helping me with my research.

TABLE OF CONTENTS

TABLE OF ILLUSTRATIONS

LIST OF TABLES

LIST OF ABBREVIATIONS

FOAS	Fluoroalcohol-Organic solvent Associated Solvnets
ypssc	yeast proteome secondary structure calculator
ABC	ammonium bicarbonate
ACN	acetonitrile
BPS	biphasic systems
DMMAPS	dimethylmyristylammoniopropane sulfonate
SDC	sodium deoxycholate
DTT	dithiothreitol
FAiC	fluoroalcohol-Induced Coacervate
FAiC-BPS	fluoroalcohol-induced coacervate biphasic systems
FASP	filter assisted sample preparation
FDR	false discovery rate
GRAVY	grand average of hydropathy
HFIP	hexafluoroisopropanol
IAA	iodoacetamide
ID	inner diameter
IMP	integral membrane proteins
IPA	isopropanol
LCMS	liquid chromatography mass spectrometry
LFQ	label-free quantification
SDS	sodium dodecyl sulfate

TBAB	tetrabutylammonium bromide
TFE	trifluoroethanol

ABSTRACT

AQUEOUS-BASED ASSOCIATED SOLVENTS FOR PROTEIN AND PEPTIDE FRACTIONATION AND PROTEIN PURIFICATION

Sajad Tasharofi, Ph.D.

The University of Texas at Arlington, 2021

Supervising Professor: Morteza G. Khaledi

The focus of this dissertation is studying capabilities of a novel two-phase system in enhancing protein identification in the whole-cell proteomics studies. In this document, two types of biphasic systems have been used for two different purposes.

The first type of biphasic system is formed by the addition of 1,1,1,3,3,3-Hexafluoro-2-propanol (HFIP) to the aqueous solutions of water-miscible organic solvents. A combination of HFIP, organic solvent, and water forms a two-phase system in spite of the fact that every solvent pair in the three-component system is miscible and forms homogeneous solutions in water. This biphasic system is called Fluoroalcohol-Organic Associated Solvent (FOAS) system. This system is used for fractionation of tryptic digest peptide mixture from yeast proteome. This fractionation is shown to be very effective in various aspects of proteomics workflow such as identification of low abundance proteins, improved identification of α-helix structures and identification of proteins with post translational modifications (PTMs).

The second type of fluoroalcohol mediated biphasic system involves the use of surfactants as the amphiphile instead of an organic solvent. These systems are called Fluoroalcohol Induced Coacervate (FAiC) biphasic systems. The FAiC systems are used to fractionate yeast proteome or yeast proteins before digestion. Fractionation of proteins in this system is shown to be effective in improving identification of low abundance proteins and proteins with α-helix structures. The FAiC systems in this study are composed of of a mixture of an anionic and and a cationic amphiphile. The relative mole ratio of the anionic and cationic amphiphile determines the charge of the coacervate phase that influences protein fractionation patterns through electrostatic interaction.

As mentioned earlier, in both biphasic systems, identification of α-helix proteins studied. Determination of secondary structure of proteins is not a trivial task. There are few experiment methods available to determine the secondary structure of proteins, but the demand is exceeding the capacities of those experimental methods to provide an answer. Computational tools have been developed to determine the secondary structure of proteins. One of the most advanced methods in terms of accuracy of prediction, is NetSurP-2.0. This tool predicts secondary structure of proteins with accuracy of 85% but it is not designed to process large number of peptides and provide a coherent picture of secondary structure of proteins corresponding to those peptides. A tool developed as an extension to NetSurP-2.0 that can process large number of peptides form bottom-up proteomics and calculates the percent coverage of secondary structures of the proteins involved in the study; this tool is named yeast proteome secondary structure calculator (ypssc) In collaboration with developers of NetSurP-2.0, we analyzed whole yeast proteome in NetSurP-2.0 and created a database for secondary structures of yeast proteins. The tool we developed, ypssc, uses this database to find secondary structures of peptides and proteins identified in the sample. Without ypssc, the task of secondary structure identification form bottom-up proteomics would be almost impossible, however, ypssc enables the user to process a large bottom-up proteomics data (like the proteomics studies that mentioned earlier) in less than 1 min in a regular desktop computer. ypssc is written in R language, and it is part of Comprehensive R Archive Network (CRAN) as a package which enables the user to use this advanced tool with no knowledge of programming.

In another project, the use of FOAS systems or desalting proteins solutions is demonstrated. The FOAS biphasic systems consist of an organic-rich bottom phase (HFIP- Organic Associate Solvent) and an aqueous-rich top phase. This unique characteristic makes FOAS suitable for desalting proteins solutions especially in the top-down studies that analyzes proteins via direct infusion to mass spectrometers. Results show that proteins, even very hydrophilic ones, are extracted in the organic bottom phase and salts in the sample are almost entirely extracted to the top aqueous phase. Experimental results show successful desalting of 3 proteins form 5 different salts.

CHAPTER 1

INTRODUCTION

Bottom-up proteomics of whole cell proteome is a challenging task because of the complexity of the peptide mixture resulting from the digestion of the proteome and analyzed by mass spectrometry[1]. Incorporation of separation or fractionation steps to simplify the complex mixtures of proteins or peptides would facilitate characterization of proteins. In the bottom-up proteomics workflow, proteins are digested to peptides using enzymes and peptides are usually analyzed by mass spectrometry. Proteins are identified from the analysis of their peptide fragments.

Separation or fractionation could be utilized in two different steps of bottom-up proteomics workflow: fractionation of proteins before digestion or fractionation of peptides after digestion[1, 2]

There are a few methods for separation or fractionation of proteins and peptides in bottom-up proteomics. The most common methods are Reversed-phase Liquid Chromatography (RPLC)[3] and strong cation exchange (SCX) chromatography[4]. Other methods including anion exchange chromatography (AEC) and Hydrophobic Interaction Chromatography (HILIC) are also used for this purpose. These methods can be implemented in tandem in online or offline setup to provide more separation capacity. In a tandem online setup, SCX and RPLC can provide great separation capacity due to the orthogonality of the separation methods[5]. In offline fractionation, any of the separation methods could be used to fractionate sample to simpler fractions[6]. Offline fractionation, is often used as a complementary fractionation step prior to online high-resolution separation[7]. Previously, in our group has developed novel two-phase systems, called Fluoroalcohol Induced Coacervates (FAiC) systems, which could fractionate proteins based on hydrophobicity[8] and charge[9]. The FAiCs are usually composed of HFIP, water and amphiphiles like surfactants. The use of mixed surfactants like DMMAPS as a zwitterionic surfactant and tetra butyl ammonium bromide (TBAB) as a positively charged amphiphile, shows considerable improvement in identification of low abundance proteins and α-helix structures[10]. The use of TBAB alone, resulted in the best improvement in the identification of low abundance proteins among all systems that we reported previously[9].

Although the combination of zwitterionic surfactants and positively charged surfactants has been studied, the effect of combination of a positively and a negatively charged surfactants has not been studied. As a continuation in this line of research, in this study, we have used a mixture of negatively and positively charged surfactant, in different molar ratios, to form biphasic systems with different electrostatic properties to fractionate yeast proteome and to study the effect of fractionation on various aspects of the yeast whole cell proteomics. We used mixtures of sodium deoxycholate (SDC) as a negatively charged surfactant and TBAB, as a positively charged amphiphile.

The protein mixture from the yeast cells was subjected to trypsin digestion and the resulting peptide mixture was fractionated using Fluoroalcohol Organic Associated Solvent biphasic (FOAS) systems. The FOAS systems are similar to FAiC, but different in chemical composition as organic solvent is used, instead of surfactant as the amphiphile. FAiC is composed of water, HFIP and surfactants but FOAS is composed of water, HFIP and organic solvent. FOAS Although FAiC systems can be used to fractionate peptide mixtures, the presence of surfactant in the system would interfere with the LC-MS analysis, thus one would face the challenging task of removing surfactants from the peptide mixture. The FOAS systems are then more suitable for peptide fractionation than FAiC due to lack of surfactants. Adding a simple step of offline using FOAS fractionation prior to RP chromatography can improve identification of proteins with Post Translational Modification (PTM) and proteins containing alpha helix structure.

A unique ability of FOAS for purification of proteins from salts are also studied. Analysis of intact proteins often requires samples that are essentially salt-free. SDS-PAGE and mass spectrometry with electrospray ionization (ESI) are examples that require protein desalting. Salts in protein samples interfere with electrophoresis, and in mass spectrometry suppresses ionization of proteins in the ion source that results in reducing protein signal and in high concentration of salt makes the protein undetectable[11]. There are many methods for desalting the protein samples and the choice of the methods depends on many factors including: volume of the sample, protein concentration in the sample, sample matrix, sensitivity of protein (pH and organic solvents), etc. Common methods that are used for desalting proteins are dialysis[12], ultrafiltration[13], precipitation[14], size exclusion and reversed phase chromatography[15]. In this document, we introduce a new method for desalting proteins using FOAS that can address the drawbacks of the current methods. In this method, protein solutions are desalted in a two-phase system that is composed of water (85% V/V), an organic solvent (butanone, 7.5% V/V) and Hexafluoro-2-propanol (HFIP, 7.5% V/V). The top phase is an aqueous-rich phase because mostly composed of water (96% V/V) and the bottom phase is an enriched in HFIP and organic solvent (HFIP, 35%

V/V; organic solvent, 40% V/V) in associated form that is called as the H-O phase in this document. The FOAS biphasic systems are fundamentally different from the conventional two-phase systems composed of water and an immiscible organic solvent such as chloroform used for protein precipitation. In FOAS, the organic solvent and HFIP are highly soluble or miscible with water. The formation of separate phases in the aqueous solutions of HFIP and polar organic solvents is due to association of HFIP, organic solvent and water molecules. Although the H-O phase is mostly composed of HFIP and organic solvent, there is a considerable amount of water present in this phase (20% V/V) which makes the H-O phase more capable of dissolving proteins and not precipitating them. The FOAS phase is structured due to hydrogen boding between the three constituent solvents and is more condensed and occupy less volume than a simple mixture of three non-interacting solvents. The top aqueous phase solubilizes most of the salt in proteins samples, while the protein is extracted in the FOAS phase. Unlike size exclusion filters and C18 methods, this method does not require expensive materials (filters, C18 columns) and lab equipment (high speed centrifuge).

As part of the proteomics studies, we developed a computational tool for identifying peptide sequences with secondary structures (especially α-helical peptides) in protein structures. Experimental determination of secondary structures is not trivial. There are two methods for determination of protein structures, Nuclear Magnetic Resonance (NMR) and X-ray diffraction. NetSurfP-2.0 is a tool that utilizes deep neural network to predict secondary structures with the accuracy of 85%[16]. In bottom-up proteomics, the proteins are enzymatically digested to peptides; sometimes this process produces hundreds of thousands of peptides. First, NetSurfP-2.0 is not designed to accept very large number of peptides at once, therefore the process of uploading the sequences and waiting for the calculations to be complete is extremely time consuming. Second, even if all sequences uploaded successfully and the results are obtained, it would be almost impossible to combine the results that have been produced for each individual peptide (hundreds of thousands of spread sheets) to get a coherent picture of the secondary structure of the proteins. In this document, an extension for NetSurfP-2.0 is presented which is specifically designed to analyze the results of bottom-up proteomics that has primarily been analyzed with MaxQuant. We call this tool Yeast Proteome Secondary Structure Calculator (ypssc). This tool is written in R language, and it is available in The Comprehensive R Archive Network (CRAN). The ypssc is user friendly and by accessible to users with no programming knowledge.

REFERENCES

(1) Ly, L.; Wasinger, V. C. Protein and peptide fractionation, enrichment and depletion: tools for the complex proteome. *Proteomics* **2011**, *11* (4), 513-534. DOI: 10.1002/pmic.201000394.

(2) Camerini, S.; Mauri, P. The role of protein and peptide separation before mass spectrometry analysis in clinical proteomics. *J Chromatogr A* **2015**, *1381*, 1-12. DOI: 10.1016/j.chroma.2014.12.035.

(3) Henzel, W. J.; Stults, J. T. Reversed-phase isolation of peptides. *Curr Protoc Protein Sci* **2001**, *Chapter 11*, Unit 11.16. DOI: 10.1002/0471140864.ps1106s24.

(4) Josic, D.; Kovac, S. Reversed-phase High Performance Liquid Chromatography of proteins. *Curr Protoc Protein Sci* **2010**, *Chapter 8*, Unit 8.7. DOI: 10.1002/0471140864.ps0807s61.

(5) Tao, D.; Qiao, X.; Sun, L.; Hou, C.; Gao, L.; Zhang, L.; Shan, Y.; Liang, Z.; Zhang, Y. Development of a highly efficient 2-D system with a serially coupled long column and its application in identification of rat brain integral membrane proteins with ionic liquids-assisted solubilization and digestion. *J Proteome Res* **2011**, *10* (2), 732-738. DOI: 10.1021/pr100893j.

(6) Yu, P.; Petzoldt, S.; Wilhelm, M.; Zolg, D. P.; Zheng, R.; Sun, X.; Liu, X.; Schneider, G.; Huhmer, A.; Kuster, B. Trimodal Mixed Mode Chromatography That Enables Efficient Offline Two-Dimensional Peptide Fractionation for Proteome Analysis. *Anal Chem* **2017**, *89* (17), 8884-8891. DOI: 10.1021/acs.analchem.7b01356.

(7) Zhang, J.; Xu, X.; Gao, M.; Yang, P.; Zhang, X. Comparison of 2-D LC and 3-D LC with post- and pre-tryptic-digestion SEC fractionation for proteome analysis of normal human liver tissue. *Proteomics* **2007**, *7* (4), 500-512. DOI: 10.1002/pmic.200500880.

(8) Koolivand, A.; Azizi, M.; O'Brien, A.; Khaledi, M. G. Coacervation of Lipid Bilayer in Natural Cell Membranes for Extraction, Fractionation, and Enrichment of Proteins in Proteomics Studies. *J Proteome Res* **2019**, *18* (4), 1595-1606. DOI: 10.1021/acs.jproteome.8b00857.

(9) Azizi, M.; Tasharofi, S.; Koolivand, A.; Oloumi, A.; Rion, H.; Khaledi, M. G. Improving identification of low abundance and hydrophobic proteins using fluoroalcohol mediated

supramolecular biphasic systems with quaternary ammonium salts. *J Chromatogr A* **2021**, *1655*, 462483. DOI: 10.1016/j.chroma.2021.462483.

(10) Khanal, D. D.; Tasharofi, S.; Azizi, M.; Khaledi, M. G. Improved Protein Coverage in Bottom-Up Proteomes Analysis Using Fluoroalcohol-Mediated Supramolecular Biphasic Systems With Mixed Amphiphiles for Sample Extraction, Fractionation, and Enrichment. *Anal Chem* **2021**, *93* (20), 7430-7438. DOI: 10.1021/acs.analchem.1c00030.

(11) Metwally, H.; McAllister, R. G.; Konermann, L. Exploring the mechanism of salt-induced signal suppression in protein electrospray mass spectrometry using experiments and molecular dynamics simulations. *Anal Chem* **2015**, *87* (4), 2434-2442. DOI: 10.1021/ac5044016.

(12) McPhie, P. Methods in Enzymology. 1971.

(13) Pohl, T. Concentration of proteins and removal of solutes. *Methods Enzymol* **1990**, *182*, 68-83. DOI: 10.1016/0076-6879(90)82009-q.

(14) Roger L. Hudgin; William E. Pricer; Gilbert Ashwell; Richard J. Stockert; Anatol G. Morell. The Isolation and Properties of a Rabbit Liver Binding Protein Specific for Asialoglycoproteins. Journal of Biological Chemistry: 1974; Vol. 249, pp 5536-5543.

(15) Pan, Y. C.; Wideman, J.; Blacher, R.; Chang, M.; Stein, S. Use of high-performance liquid chromatography for preparing samples for microsequencing. *J Chromatogr* **1984**, *297*, 13-19. DOI: 10.1016/s0021-9673(01)89024-5.

(16) Klausen, M. S.; Jespersen, M. C.; Nielsen, H.; Jensen, K. K.; Jurtz, V. I.; Sønderby, C. K.; Sommer, M. O. A.; Winther, O.; Nielsen, M.; Petersen, B.; et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins* **2019**, *87* (6), 520-527. DOI: 10.1002/prot.25674.

CHAPTER 2

PEPTIDE FRACTIONATION IN THE BOTTOM-UP PROTEOMICS

Used with permission from Sajad Tasharofi,Duga Khanal, Mohammadmehdi Azizi, Morteza G. Khaledi,

ABSTRACT

A simple and fast offline fractionation methods used to fractionate peptides of Saccharomyces cerevisiae (yeast) after tryptic digestion of its proteins. Fractionation performed in a two-phase system that is composed of 85% water, 7.5% 1,1,1,3,3,3-Hexafluoro-2-propanol and 7.5% of water miscible organic solvents. Such fractionation can reduce complexity of the yeast peptide mixture by fractionating them based on hydrophobicity. The result of such fractionation shown to be beneficial in various aspects of proteomics. As a result of extraction and enrichment of hydrophobic peptides to the bottom phase of the two-phase, the number of identified proteins with α-helix structure increased after fractionation compared to control sample with no phase separation. Coverage of α-helix part of proteins, especially membrane proteins improved 15% in some cases. More importantly, the number of proteins with Post Translational Modification

(PTM) improved; in case of phosphorylated proteins in one of samples there was 43% improvement.

2.1 INTRODUCTION

Improvement in identification of proteins in complex mixtures in bottom-up proteomics can be achieved by fractionation of those mixtures into simpler fractions before identification by tandem mass spectrometry. In bottom-up proteomics, enzymatic digestion of proteins into peptides leads to greater sample complexity. The resolving power of existing mass spectrometers is not adequate to analyze complex mixtures without prior separation [1]. Separation methods fractionate a complex sample into simpler mixtures that can be analyzed more readily by mass spectrometry [1,2]. Reversed-phase Liquid Chromatography (RPLC) and strong cation exchange (SCX) chromatography are two commonly used separation methods for proteins and peptides prior to MS detection. Other methods including anion exchange chromatography (AEC) and Hydrophilic Interaction Chromatography (HILIC) are also used for this purpose. These methods can be implemented in tandem in online or offline setup to add extra separation steps. In a tandem online setup, a combination of SCX and RPLC for 2D separations can provide great peak capacity due to the orthogonality of the two methods[3,4].

In offline fractionation, other separation methods can be used prior to MS or LC-MS analysis[5]. Deng et al. used magnetic beads to perform offline fractionation of peptides prior to online RPLC and they demonstrated that it increases the depth of peptide analyte coverage[6]. Several studies have shown the effect of offline peptide fractionation on improving protein identification using common separation methods. Edward et al. used SCX offline fractionation in combination with RPLC-MS to improve peptide fractionation and protein identification [7]. Ramesh et al. used RPLC to fractionate a peptide mixture into 15 fractions and analyzed each fraction by capillary zone electrophoresis (CZE). They demonstrated that a combination of these two separation methods provides far better sequence coverage of proteins than that by each individual method[8].

Previous work from our group reported the usefulness of Fluoroalcohol Induced Coacervates (FAiC) biphasic systems for extraction, fractionation and enrichment of complex protein samples in proteomics analysis[9]. In bottom-up proteomics workflow, FAiC systems were used to fractionate the yeast proteome into two separate aqueous and coacervate phases prior to enzymatic digestion and analysis of the tryptic maps in each phase by RPLC-MS/MS. The offline fractionation using FAiC improved the protein coverage, especially for low abundance proteins and subproteomic such as proteins located in membranes, mitochondria, and phosphorylated proteins. We further investigated the effects of FAiC composition on protein fractionation patterns according to protein hydrophobicity, charge, and molecular weight[10].

In this study, we investigated fractionation of peptides mixtures in a bottom-up proteomics workflow using a different class of fluoroalcohol induced two-phase system. The tryptic peptide mixture of yeast proteins was fractionated using Fluoroalcohol-Organic Associated Solvents (FOAS) biphasic systems. The FAOS systems are like FAiC but utilize polar organic solvents (instead of surfactants) as the amphiphile that interacts with the fluoroalcohol and leads to phase separation in the aqueous media. FAiC is composed of water, HFIP and surfactants while FAOS is composed of water, HFIP and organic solvent. In both FAiC and FOAS systems, a water soluble fluoroalcohol such as hexafluoroisopropanol (HFIP) interacts with an amphiphile, surfactant in FAiC and polar organic solvent in FAOS, through hydrogen bonding and hydrophobic interactions that results in phase separation in the aqueous media. In these biphasic systems, one phase is enriched in the amphiphile and HFIP and the other phase is composed of mostly water and small amount of HFIP and the amphiphile. In FAiC, the surfactant-amphiphile rich phase is called coacervate, that is a form of organized self-assembly of surfactant molecules and mediated through interaction with the fluoroalcohol[11,12]. In FOAS, the polar organic solvent is miscible or highly soluble in water, like the fluoroalcohol. The formation of a second phase in the aqueous mixtures of HFIP and polar organic solvent in FOAS is primarily driven by hydrogen bonding between the strong H-bond donor hydroxyl group on the fluoroalcohol and the hydrogen bond acceptor group on the polar organic solvent on one hand, and the hydrophobic interaction between the fluoroalcohol's fluorocarbon groups and organic solvent's hydrocarbon groups. Thus, we refer to the HFIP-organic solvent phase as Fluoroalcohol – Organic Associated Solvents (FOAS) to emphasize that solvation and physical properties (such as phase volume) of the FOAS phase are different from a simple mixture of the constituent solvents. A main difference between FAiC and FOAS is that surfactant molecules (the amphiphile in FAiC) self-assemble in aqueous media through hydrophobic interaction, while polar organic solvents do not form self-assemblies. The FAiC and FOAS phases share the common features of being highly concentrated in HFIP and the amphiphile; thus, they have strong solubilizing power for hydrophobic compounds. Due to high concentrations of the amphiphile and fluoroalcohol, these phases are considerably more hydrophobic than the aqueous phase, which makes them attractive for solubilization of more hydrophobic compounds. Hydrophilic compounds have greater affinity toward the aqueous-rich phase in the biphasic systems. As a result, these biphasic systems can be used to separate complex mixtures into simpler hydrophilic and hydrophobic sample fractions for analysis. Their organized molecular structures (coacervates in FAiC and Associate Solvents in FOAS) translate into more condensed phases with much smaller volumes than the initial solution volumes, which results in enrichment of extracted solutes into these phases. The FAiC systems are more suited for extraction, fractionation, and enrichment of proteins than peptides. The surfactant in the FAiC systems would have to be removed from the sample prior to the LC-MS analysis. This is readily accomplished in protein samples with Filter Assisted Sample Preparation (FASP) that separates the surfactant from the proteins through a molecular sieving mechanism. Separation of surfactants from peptides would be a far more challenging task. Thus, the FOAS systems would be better suited for extraction, fractionation, and enrichment of peptides and other small molecules.

2.2 EXPERIMENTAL

2.2.1 Materials

Millipore-DI water was used for sample preparation.1,1,1,3,3,3-Hexafluoro-2-propanol (HFIP) was obtained from Oakwood Chemical, USA. 2-butanone or methyl ethyl ketone (MEK) was purchased from Alpha Aesar with the purity of 99%. Tetrahydrofuran (THF) was purchased from fisher chemicals with the purity of ≥99.9%. Dimethyl sulfoxide (DMSO), for molecular biology was purchased from Sigma with the purity of 99.9%. Formic acid (99%) was purchased from Alfa Aesar. Chemicals for pre-digestion and digestion of proteins like dithiothreitol (DTT), Iodoacetamide (IAA), sequencing grade trypsin, were purchased from Promega Corporation, 2800 Woods Hollow Road, Madison

2.2.2 Cell lysate preparation

Saccharomyces cerevisiae (strains BY4741, Ward's Science®) was grown and lysed according to the our previous publication[13]; however, the lysis buffer was modified by adding 10 µL pepstatin at the concentration of 1 mg/mL to every 10 mL lysis buffer without adding sodium chloride to the lysis buffer. The lysis solution was made by dissolving 1 Mini Tablet of EDTA free Pierce® Protease and Phosphatase Inhibitor in 10 mL of autoclaved deionized water and adding 10µL of pepstatin solution (Roche® Diagnostics GmbH, dissolved in methanol, 1 mg/mL).

2.2.3 Two-phase formation and fractionation

All samples have been prepared in a 1.5 mL Eppendorf low-bind microcentrifuge tubes; each sample has a total volume of 1 mL and is composed of 850 µL water, 75 µL organic solvent, 75 µL HFIP, and 500 µg of yeast tryptic peptide mixture (yeast peptides); procedure for desalting yest peptides mentioned in Appendix 2-1. Three organic solvents that were used for this study included dimethyl sulfoxide (DMSO), butanone, and tetrahydrofuran (THF). In this manuscript we refer to the FAOS systems consisting of water, HFIP, and organic solvent by the name of the organic solvent as DMSO system, butanone system, and THF system. As explained in Appendix 2-1, for desalting of yeast peptides, a C18 cartridge is used to desalt the peptides. The eluent containing acetonitrile for washing peptides from the C18 cartridge interferes with the formation of the two-phase systems thus peptide samples had to be dried before being used for the experiment. After the desalted peptide solution that contains 500 µg of yeast peptides is dried in an Eppendorf vacuum centrifuge, 850 µL of water was added to the dried peptide sample, vortexed, and sonicated for 1 min each. 75 µL of organic solvent and 75 L HFIP added to the mixture, vortexed, and sonicated for 1 min. The mixture is then centrifuged at 4000 g for 4 min and then two -phase is formed. **Error! Reference source not found.** shows the flow chart of the whole process.

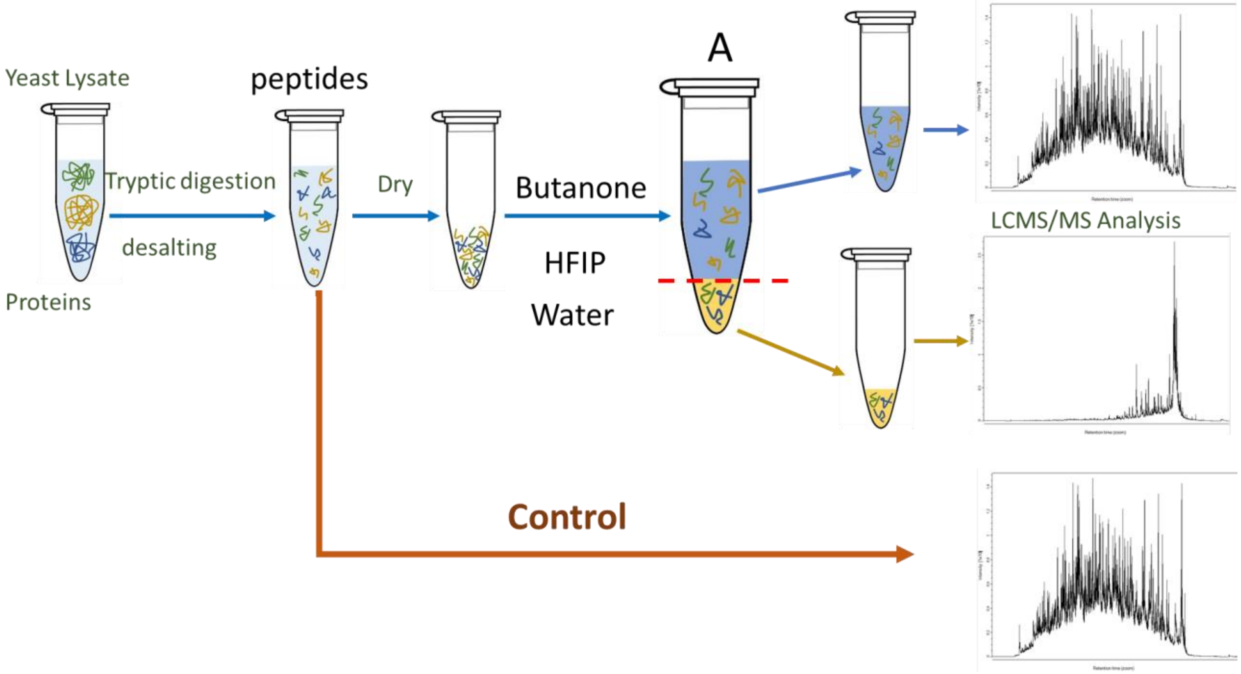


Figure 2-1. Workflow of peptide fractionation prior to LCMS/MS analysis. Organic solvents that could be used to form the two-phase are THF, DMSO and butanone; in this example butanone used to form the two-phase system

In tube A in **Error! Reference source not found.**, the top phase (blue) is the aqueous phase, and t he bottom phase (yellow), is the HFIP-Organic (H-O) phase, which consists of high concentrations of HFIP and the organic solvent, and much smaller levels of water. The top phase is water-rich and has much smaller concentrations of HFIP and the organic solvent and is referred to as the Aqueous (Aq) phase.

2.2.4 Compositional analysis of the two-phase system

The chemical compositions of each phase in the FOAS systems were analyzed using GC-MS analysis. In this experiment, the GCMS-2010SE (Shimadzu) instrument was used with a capillary column (30 m × 0.25 mm × 250 µm) containing 5% Phenyl-Arylene 95% Dimethylpolysiloxane stationary phase. Helium was used as the carrier gas throughout the experiment at the flow rate of 0.9 mL/min. The analysis was done in single ion mode (SIM) to get a better sensitivity. A method was developed for the characterization and separation of solvents and is shown in Table 2-1. Calibration plots were developed and the concentrations of solvents into each phase were determined. Before analysis, the top aqueous phase and the bottom organic phase were diluted several times to avoid overloading the column and the detector.

Table 2-1. GCMS method program and parameters

Samples	Sample injection			Carrier gas flow (mL/min)		Oven profile			Ion source Temp (ºC)
Solvent mixture	Inj. temp.(ºC)	Inj. Vol. (µL	Split ratio	Carrier gas	Column flow (mL/min)	Rate (ºC/min)	Oven temp. (ºC)	Hold time (min)	
	200					-	100	0.5	250
		1	0.3888889	Helium	0.9	25	200	0	

Error! Reference source not found. shows contour plots of the percentages of HFIP and b utanone in each phase at different ratios of HFIP and butanone in the aqueous mixtures.

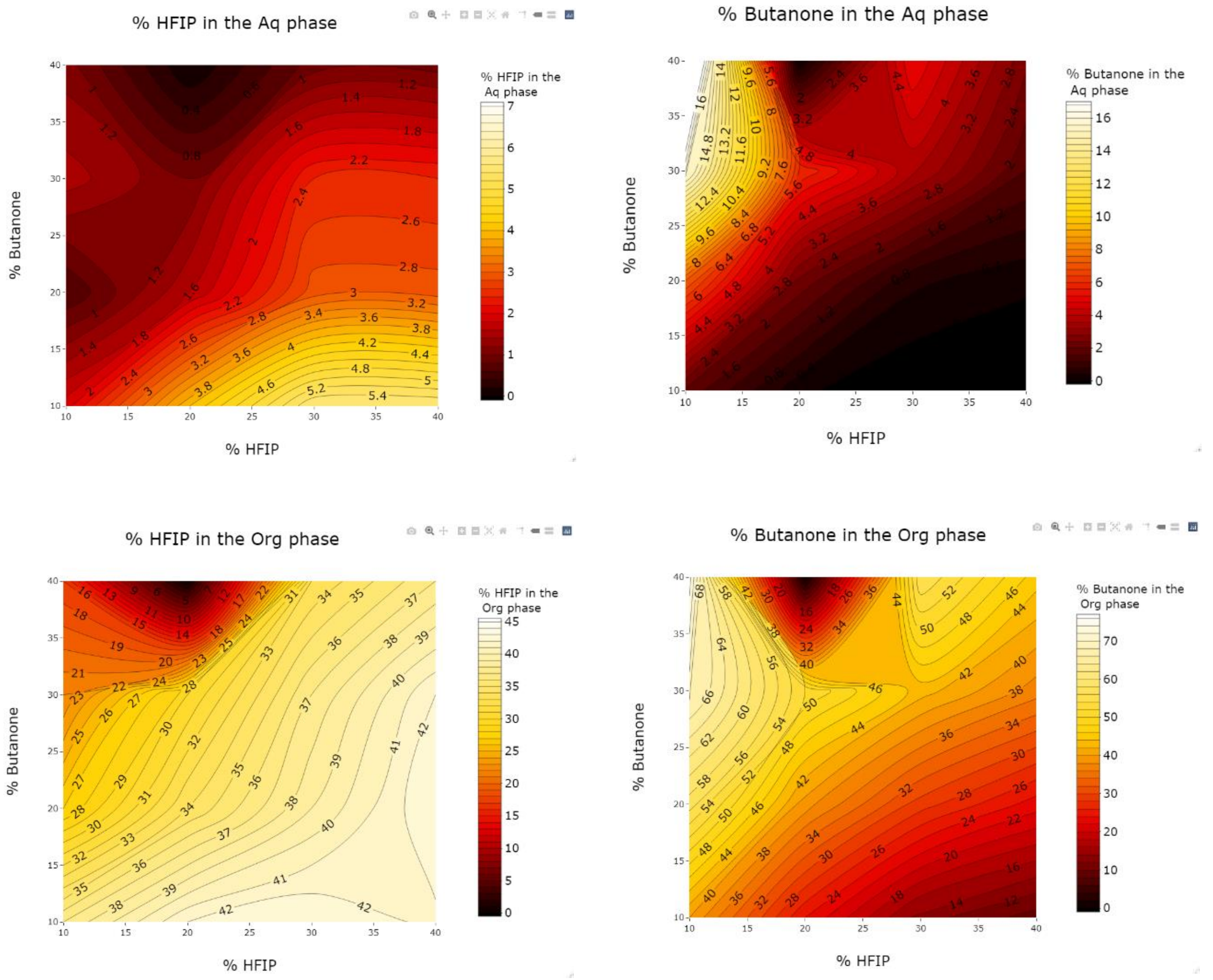


Figure 2-2. Contour plots of the percentages of HFIP and butanone in each phase at different ratios of HFIP and butanone in the aqueous mixtures.

Table 2-2 shows the compositional analysis of the butanone-HFIP two phase system. The concentrations of HFIP and butanone are almost 20 times higher in the organic phase as compared to that in the aqueous phase.

Table 2-2. % HFIP and butanone in aqueous and organic phase of in the butanone system

	Aqueous phase	Organic phase
% HFIP	2	35
% Butanone	2	40

Peptide mixtures form each phase then dried to remove the acetonitrile and then reconstituted using a solvent that contains in 2% acetonitrile with 1% formic acid. During this process concentration of peptides adjusted to 1 mg/mL. concentration of peptides measured using Thermo Scientific^{TM} Nanodrop^{TM} One. After peptide concentration adjustment, samples analyzed in mass spectrometer.

Table 2-3. Concentration of peptides in each phase in all three systems.

	Concentration of peptides aqueous phase (µg/µL)	**Concentration of peptides organic phase (µg/µL)**	**Volume of the aqueous phase (µL)**	**Volume of the organic phase (µL)**
THF system	0.6	0.3	910	90
DMSO system	0.4	3.2	970	30
Butanone system	0.6	0.5	890	110

2.2.5 LCMS analysis

Reconstituted peptides were analyzed by LC-MS/MS (Ultimate 3000 RSLC-Nano liquid chromatography systems, Dionex; coupled with Orbitrap Fusion Lumos MS®, Thermo Electron) analysis. 1 µg of each sample was injected into a C18 column (EasySpray column, i.d.: 75 µm, length: 75 cm, particle size: 3 µm, Thermo®) and eluted with the following gradient, 0-90 min: 0-28% B, flow rate of 350 nL/min. Mobile phase A was 2% (V/V) Acetonitrile (ACN) and 0.1% (V/V) formic acid (FA) in water, and mobile phase B was 80% (V/V) ACN, 10% (V/V) trifluoroethanol (TFE), and 0.1% FA in water. The mass spectrometer operated in positive mode at the following conditions, source voltage: 2.2 kV, ion transfer tube temperature of 275 °C, resolutions: 120,000; number of MS/MS spectra event: up to 10 for each full spectrum, fragmentation: higher-energy collisional dissociation (HCD) for ions with charges of 2-7, and dynamic exclusion: 25 s after an ion was selected for fragmentation.

2.2.6 Data Analysis

MaxQuant (version 1.1.0.1) was used to process the raw data by using the yeast database from http://www.uniprot.org. The following options were used for protein identification, first-search peptide tolerance: 20 ppm, main search peptide tolerance: 4.5 ppm, ITMS MS/MS match tolerance: 0.5 Da, enzyme: trypsin, missed cleavage: 2, fixed modification: carbamidomethyl (C), variable modification: oxidation (M) and acetylation (Protein N-term). False discovery rate (FDR) thresholds were specified at 1% for protein. The minimum unique peptide was set to 1. In Label-free quantification, iBAQ was selected for quantification in LC-MS/MS analysis.

After initial analysis of MS data using MaxQuant, further data analysis and visualization were performed using an in-house written program. We wrote R programs to analyze results of MaxQuant and extract information presented in this manuscript. We have shared this program in the GitHub (https://github.com/Stasharofi/Transmembrane-α-helix-calculator). There is an instruction manual in our GitHub page for using this program as well as the description on how this program works.

2.3 RESULTS AND DISCUSSION

2.3.1 Peptide Fractionation Patterns

Fractionation of yeast peptide digest mixture resulted in significant improvements in coverage of the low abundance proteins and post translationally modified (PTM) proteins.

Fractionation patterns in FOAS biphasic systems is based on solute hydrophobicity where more hydrophobic compounds have greater affinity toward the FOAS phase. This can be seen in Figure 2-3 to Figure 2-5 that illustrate the RPLC tryptic maps of the peptides extracted in the aqueous and HFIP-Organic (H-O) phases for the three FOAS systems. As can be seen, the elution patterns of peptides extracted in the H-O phases are different from those of in the aqueous-rich phase where the peptides in the H-O phases begin to elute after 70 min and are generally later eluting than those in the aqueous-rich phases.

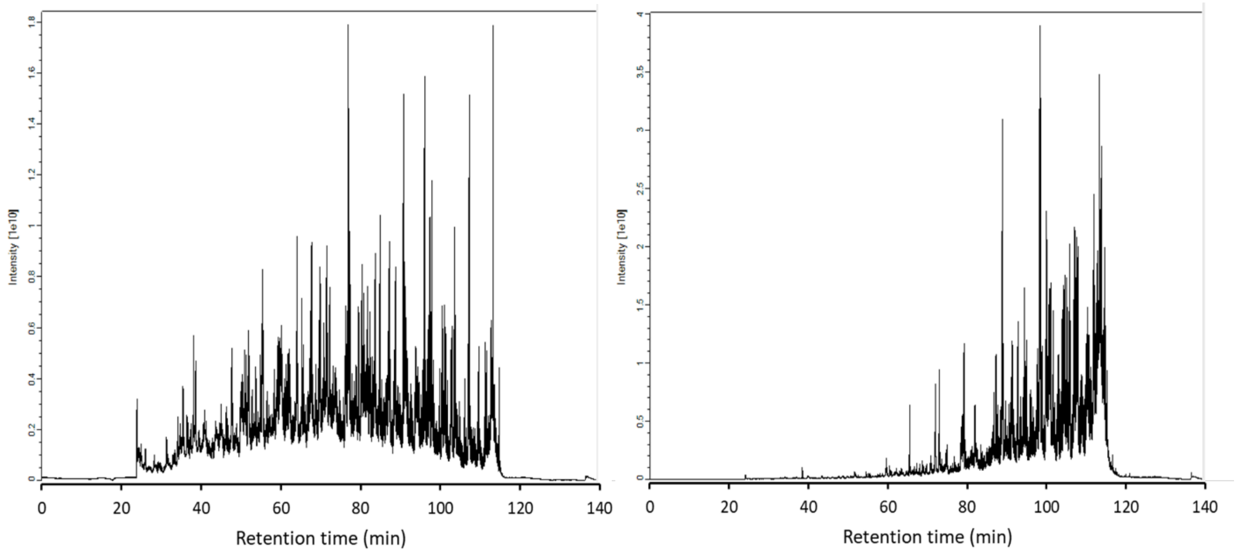


Figure 2-3. Total ion chromatograms of peptides after fractionation in butanone/HFIP system; Right is organic phase and Left is Aq phase

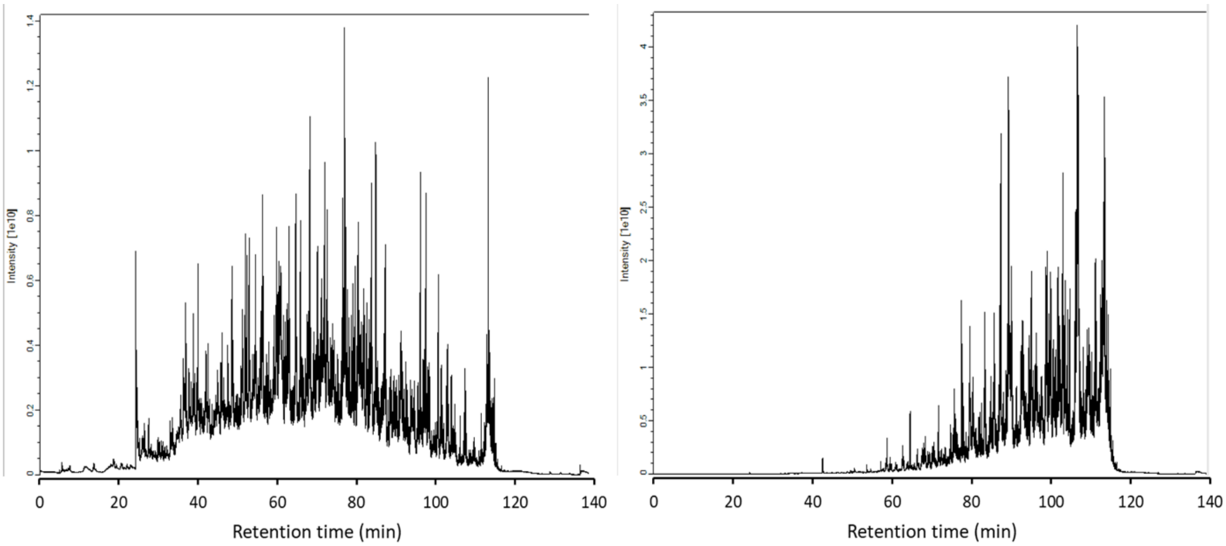


Figure 2-4. Total ion chromatograms of peptides after fractionation in DMSO/HFIP system; Right is organic phase and Left is Aq phase

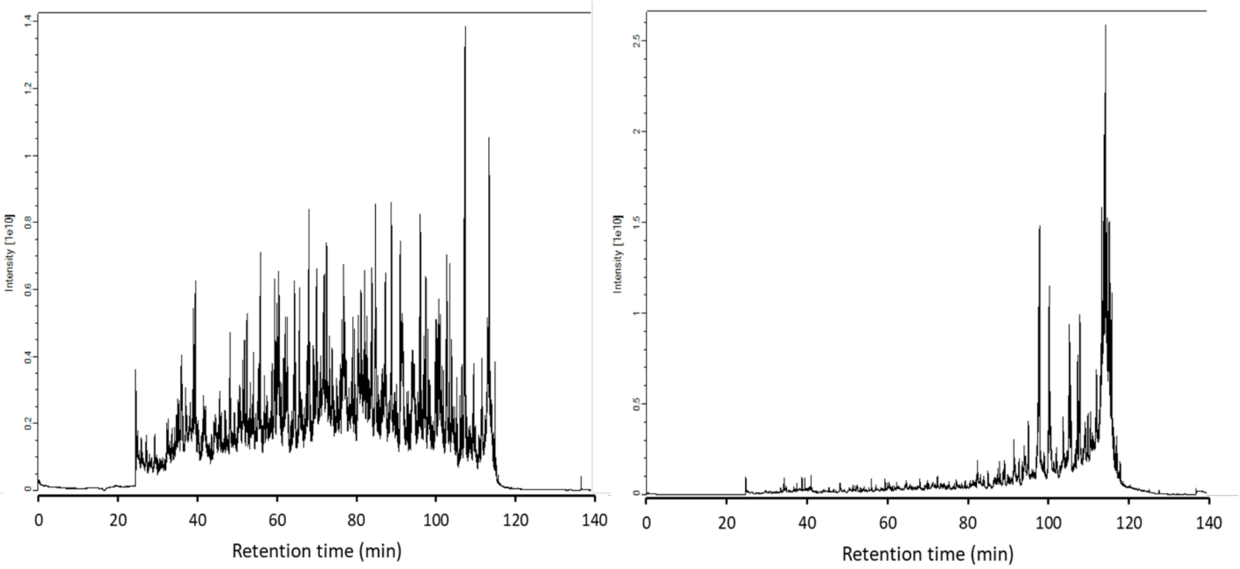


Figure 2-5. Total ion chromatograms of peptides after fractionation in THF/HFIP system; Right is organic phase and Left is Aq phase

Most peptides in the tryptic digest of the yeast proteome are hydrophilic; that means the relative abundance of hydrophobic peptides in the mixture is lower than hydrophilic peptides. Thus, hydrophobic peptides are underrepresented in the MS which results in poor identification of those peptides. Hydrophobic Peptides are enriched upon their extraction into the FOAS phase Figure 2-6 shows there is a significant difference between grand average hydropathy (GRAVY) of peptides that were uniquely identified in the aqueous and the H-O organic phase in all systems GRAVY is a measure of peptide hydrophobicity that corresponds to the amino acid composition[14] Similar trends were observed for the other two FOAS systems

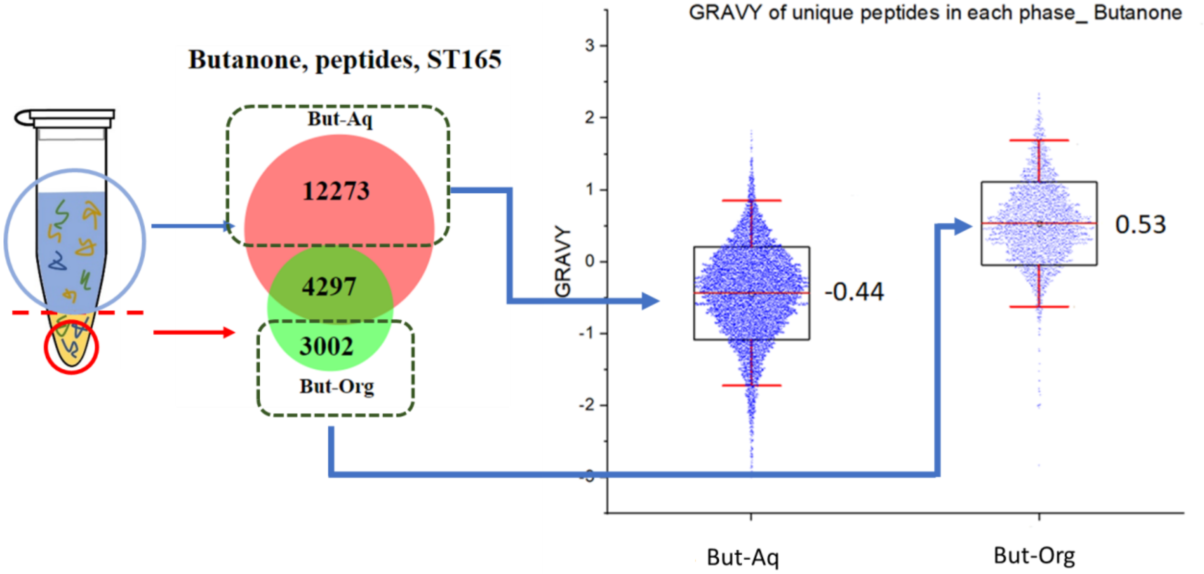


Figure 2-6. For butanone system, plotting the peptides (based on GRAVY) that only identified in the aqueous phase and H-O phase shows great difference between the average values.

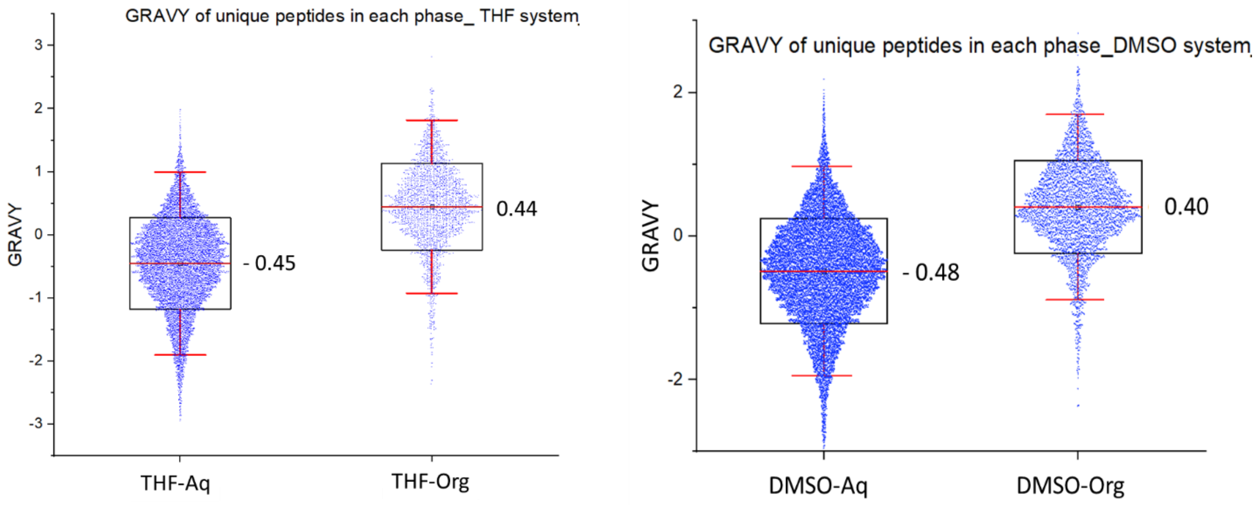


Figure 2-7. For the butanone system, the peptides in the HFIP-Butanone (H-O) phase are more hydrophobic and have larger GRAVY than those in the aqueous phase.

FOAS has a big difference with conventional organic-water two-phase systems in which the organic solvent is immiscible with water. In these systems (octanol/water), because the organic solvent is immiscible with water, there is a big difference between the hydrophobicity of the phases; as a result of this drastic change in hydrophobicity, peptides usually precipitate in the interface of the two-phase. Unlike the conventional two-phase systems, all the components of FOAS are miscible with water and yet the combination of all components form a two-phase system. This special characteristic is the key to successful fractionation without precipitating peptides. We believe that presence of water in the organic phase, although very low, provides better extraction media compared to pure organic solvents that are used in conventional two-phase systems. This characteristic of the phases allows a wide spectrum of peptides (in sense of hydrophobicity) to be used and fractionated.

Partition coefficient of a peptide in the FOAS biphasic system is defined as the ratio of the peptide concentration in the two phases (Eq. 1) and was determined for the peptides that were distributed in both phases from the ratio of their MS intensities.

$$K = \frac{[peptide]_{\ H-O\ phase}}{[peptide]_{\ aqeous\ phase}} \qquad \text{Eq. 1}$$

Figure 2-8 shows the pattern in the log K of peptides distributed between the two phases in the butanone FOAS system as a function of their GRAVY scores. Two populations of peptides are indicated on the graph (by blue rectangles) to differentiate the peptides with Log K less than and greater than zero that differentiates the peptide with the greater affinity toward the aqueous phase ($log\ K < 0$) from those toward the H-O phase. Clearly, the population with log K<0 have lower GRAVY than those with $log\ K > 0$. Figure 2-9 compare the patterns in the three FOAS systems and shows that in the THF system, most peptides have log K < -1 (or K < 0.1), which means the concentration of those peptides are at least 10 time higher in the aqueous phase. This also corresponds to smaller GRAVY values (hydrophilic peptides) and is another indication that these systems are particularly effective in fractionating peptides based on hydrophobicity.

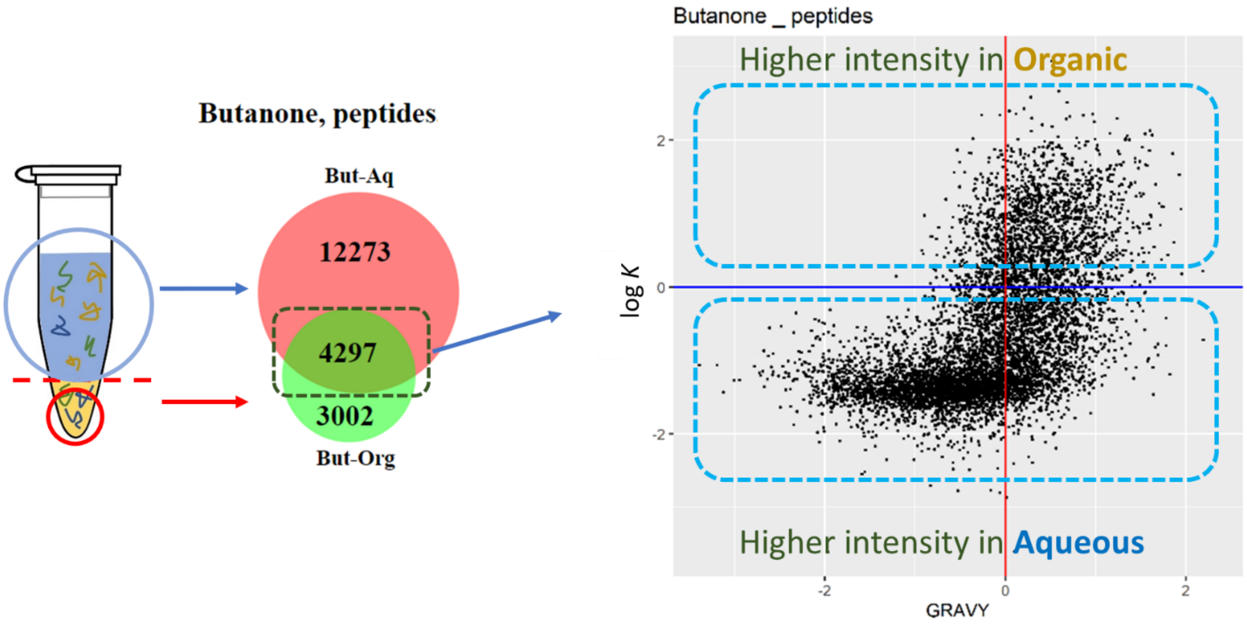


Figure 2-8. Plotting peptides that are shared between phases of butanone system shows that peptides with higher GRAVY have more concentration in the organic phase and peptides with lower GRAVY have higher concentration in the aqueous phase.

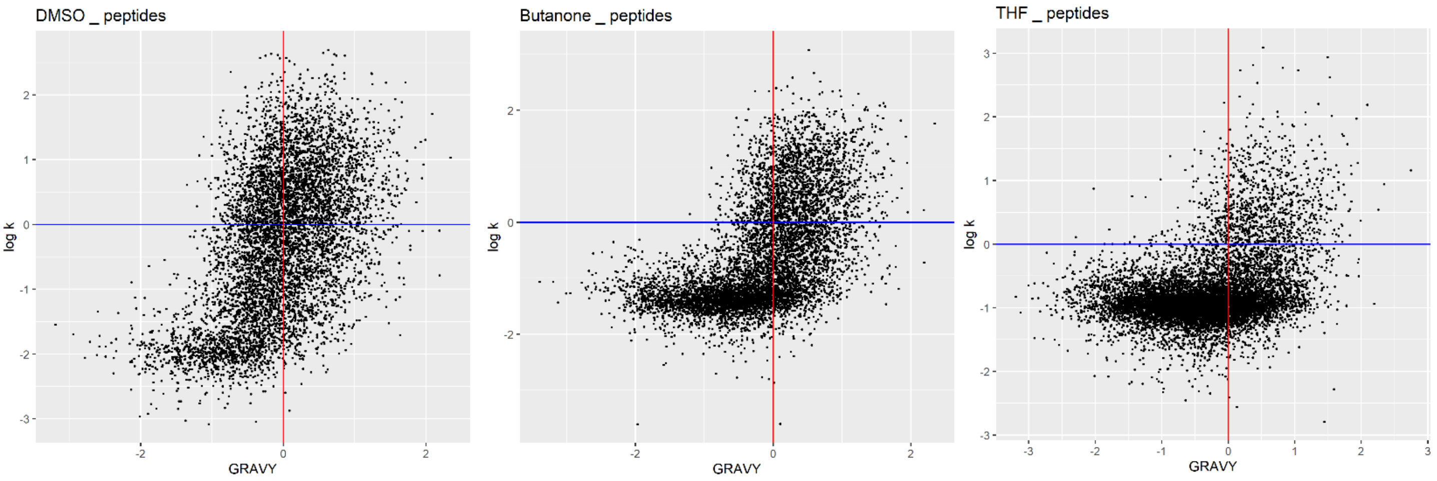


Figure 2-9. Plotting peptides that are shared between phases of all systems show that peptides with higher GRAVY have more concentration in the organic phase and peptides with lower GRAVY have higher concentration in the aqueous phase.

2.3.2 Protein analysis

The FOAS systems were incorporated in the bottom-up proteomics workflow. The results of the proteomics analysis are presented in the following sections that include the number of proteins identified in each phase of all samples and control, comparing and determining the number of unique and shared peptides between each phase and control, identification of proteins with post translational modifications, identification of low abundance proteins and alpha-helix proteins.

Figure 2-10 shows the number of proteins identified in each phase as well as number of proteins identified in both phases.

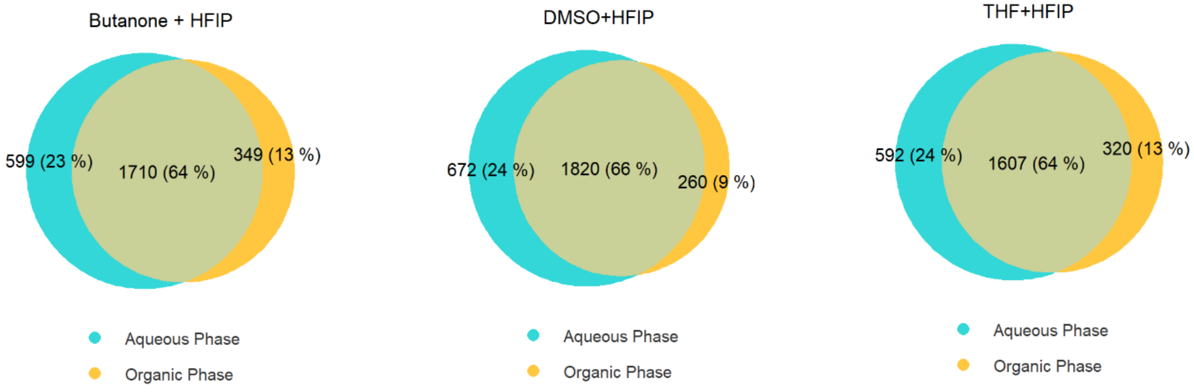


Figure 2-10. Number of proteins identified in each phase of all 3 systems. The Euler diagram differentiates between proteins that identified uniquely in each phase and shared proteins that identified in each phase

Figure 2-11 shows the number of proteins that have been identified in each sample (aqueous + H-O phase) compared to the control.

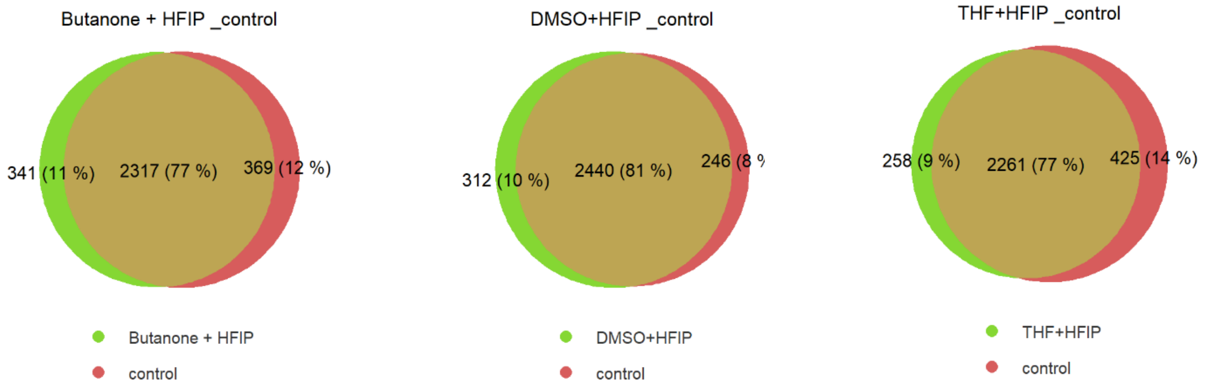


Figure 2-11. Number of proteins that have been identified in each sample (aqueous + H-O phase) compared to control.

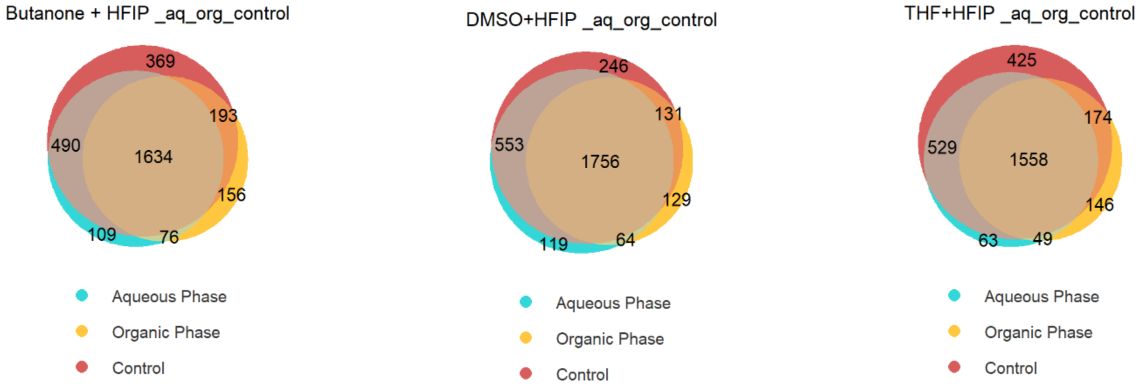


Figure 2-12. shows the population of identified proteins in each phase and control for all 3 samples. The Euler diagram provides information about the number of shared proteins between phases and control

2.3.2.1 Improvement in identification of proteins with PTM

Post translational modifications of proteins can have a profound impact on their bilogical function[15]. Detection of PTM in proteins is necessary for development of therapeutic drugs[16].

There are many types of PTM that can occour in proteins, phosphorylation and Glycolysation are the most common ones. Abnormal phosphorylation of proteins is directly linked to many diseases; for example, Cohen named 19 types of diseses that are linked to abnormal phosphorylation of proteins in human body[17] . Reily et al. related many congenital disorders to glycolysation of different types of proteins[18].

In this study, we also examined the number of peptides with phosphorylation, glycosylation and oxidation. The results have been analyzed to identify the number of proteins with those PTM in the sample. Figure 4 shows the comparison between number of proteins with PTM identified using the three FOAS systems and the control. In the case of phosphorylation, there is a 43% improvement in identification of phosphorylated proteins as a result of offline frcationation in butanone system as compared to the control system . The THF and DMSO systems underperformed as compared to the cotrol. Larger number of glycosylated proteins were identified using the DMSO system. In case of oxidation, all systems are showing improvements as compared to the control system.

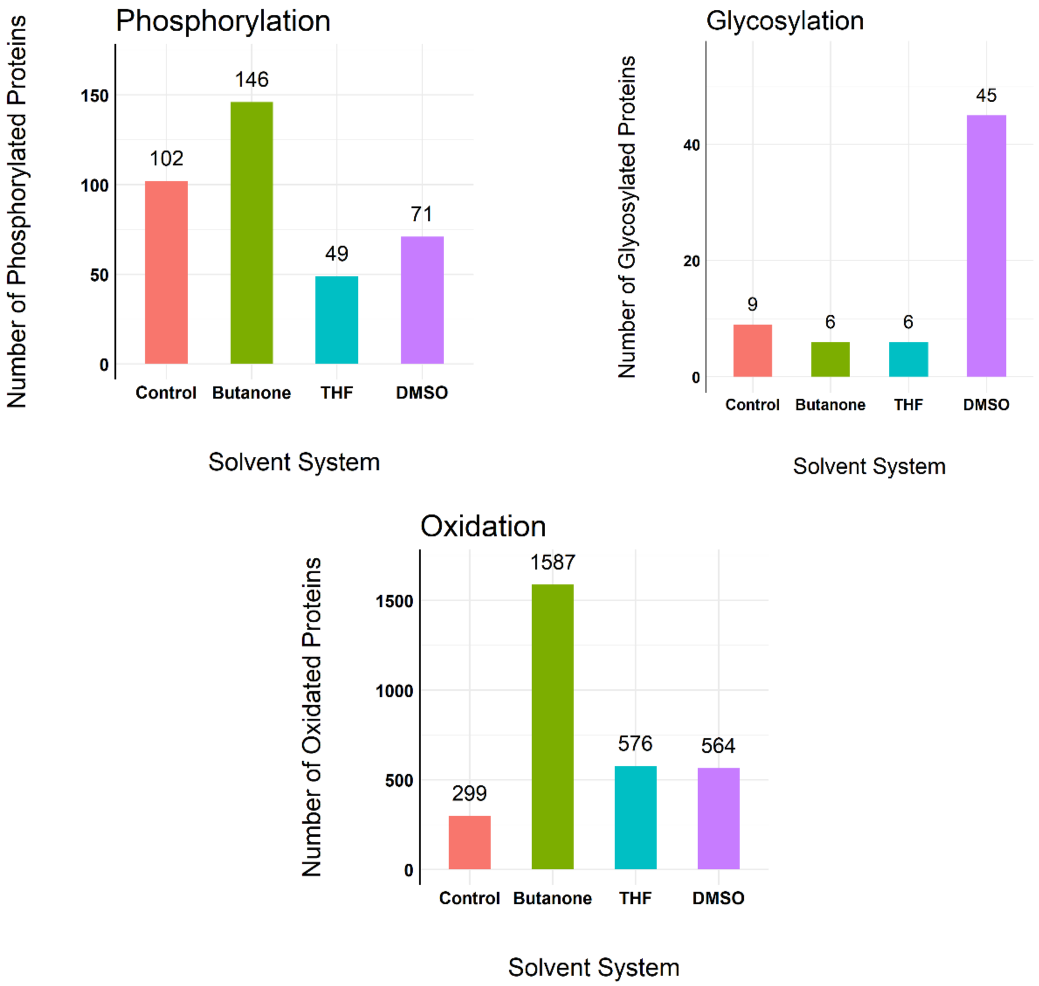


Figure 2-13. comparison between number of proteins with PTM detected in each system and control

2.3.2.2 Improvement in identification of low abundance proteins

Wide dynamic range of protein expression in the cells is one of the reasons that low abundance proteins are underrepresented in proteomics analysis. There are four orders of magnitude difference between protein abundances in the yeast proteome[19]; which leads to poor identification of low abundance proteins.

There are potentially over 6700 proteins in yeast with wide variety of abundances. Yeast proteome divided to groups based on the abundance of proteins and the number of proteins identified from each group calculated. This process done for control and samples done with FOAS and the results compared. Figure 2-14 shows improvement in the number of proteins identified in the samples that treated using FOAS and control, for each group of abundance. Results show that for the group of proteins with abundances < 2000 molecule per cell, there is a significant improvement in identification of proteins using the butanone and DMSO systems as compared to the control. We observed over 55%, 50% and 15% improvement in identification of proteins abundance level below 2000 molecules per cell respectively for the butanone, DMSO and THF systems. This is a significant finding because the results show that a simple step of fractionation can have a considerable impact on identification of proteins that have the lowest abundance in yeast proteome. For the abundance groups higher than 2000 molecules per cell, with few exceptions, we did not see any improvement in identification using the FOAS systems.

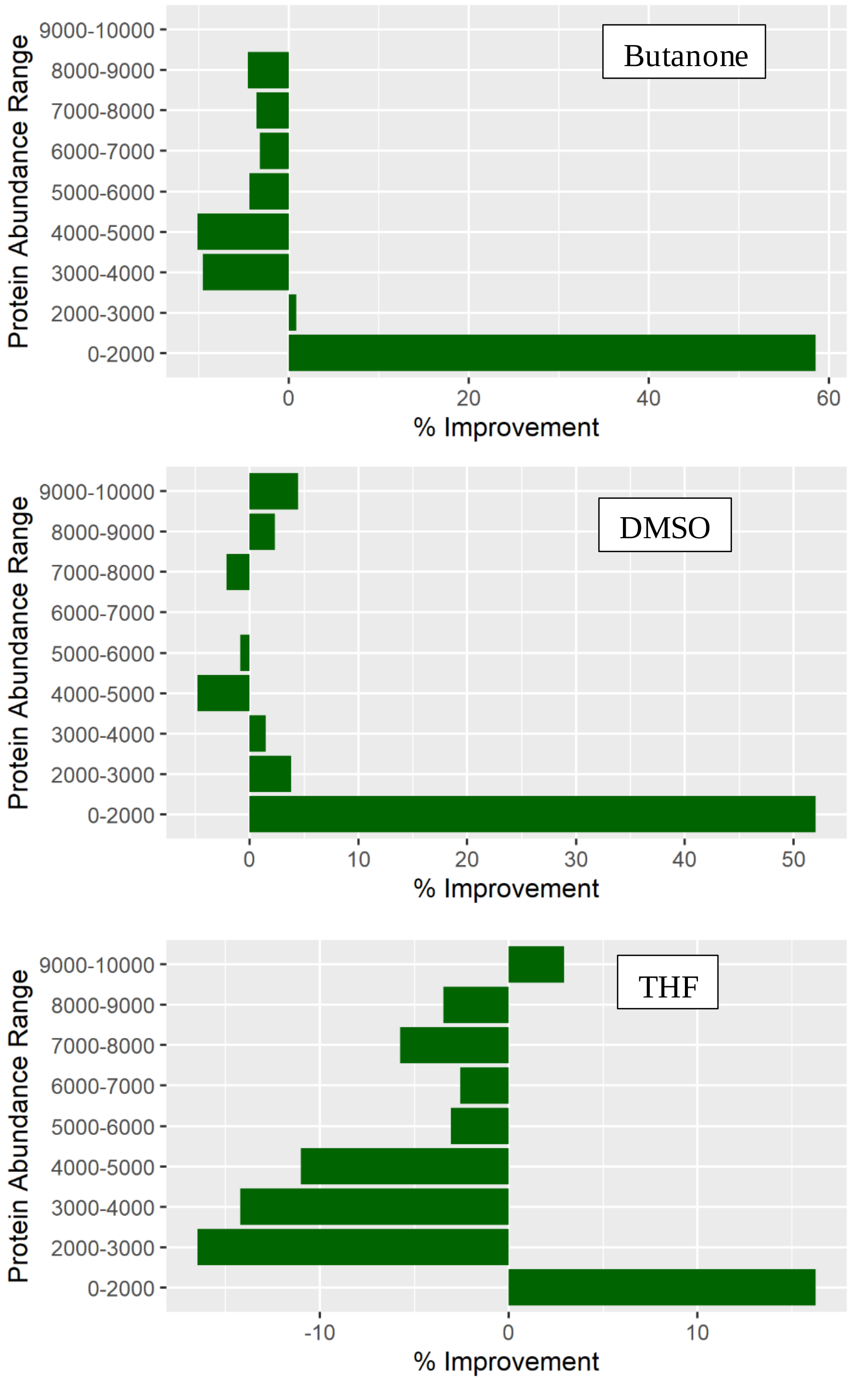


Figure 2- 14. Identification improvements because of fractionation of peptides. Comparison between samples and control, shows that in all systems there is improvement of identification of proteins with abundance <2000 molecules per.

In a different representation of data in Figure 2-15, in Appendix 2-2 the number of proteins identified in each abundance range using the FOAS and the control and samples fractionated using FOAS.

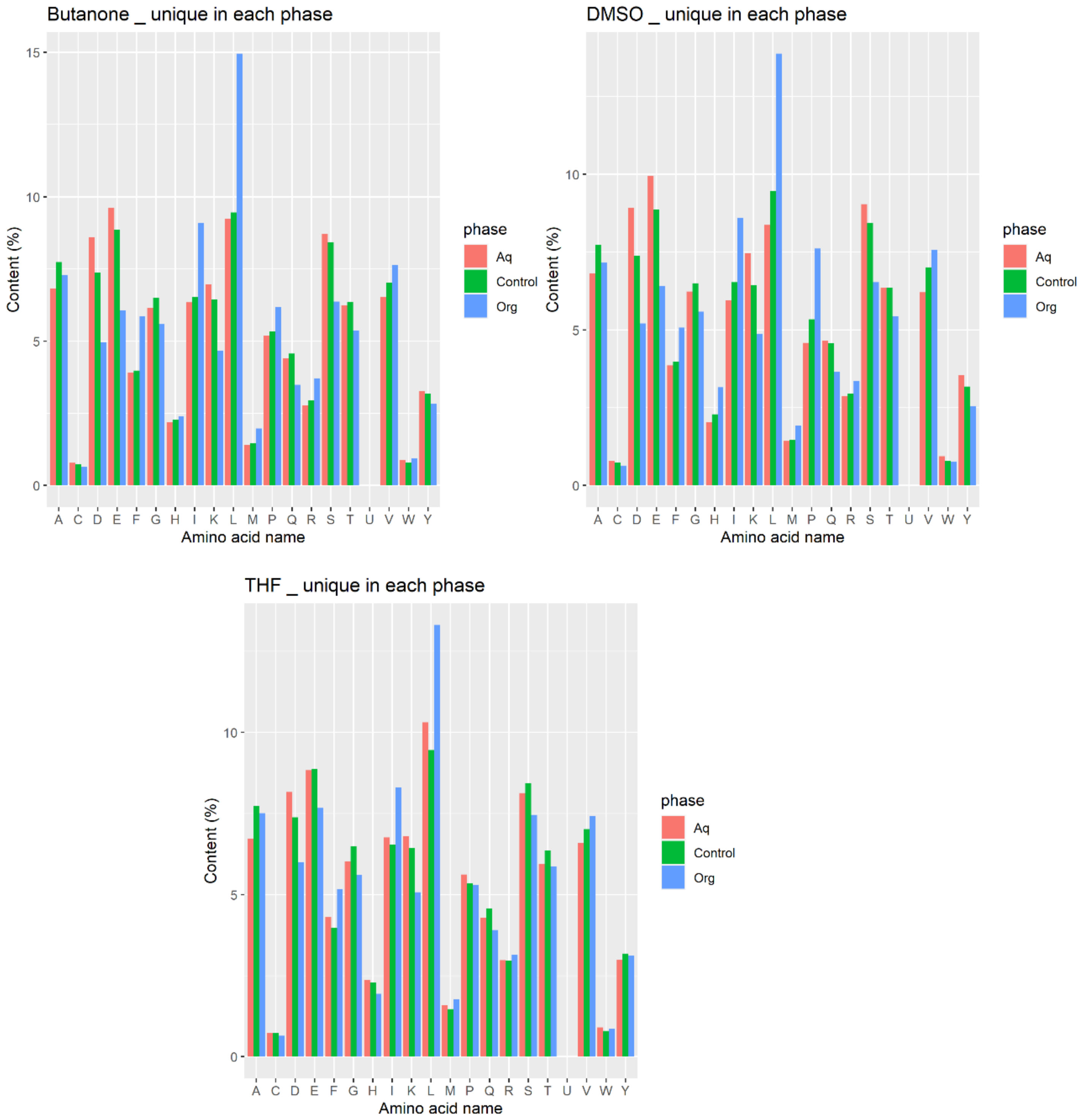
2.3.2.3 Enrichment of peptides with α-helix structure into H-O phase

To predict the sequences in the yeast proteome that are forming alpha-helix structure, we used NetSurfP-2.0 which is a structural prediction tool that can identify alpha-helix structure in any amino acid sequence[20].

Calculations based on the database from NetSurfP-2.0 show that alpha-helix structures make up to 38% of total yeast proteome sequences and 42 % of the yeast membrane proteins. The experimental data from the control sample in our study shows that 36% of yeast proteome is alpha-helix, which is very close to NetSurfP-2.0 prediction. However, for membrane proteins only 35% of sequences that have been identified are alpha helices. This shows that the alpha-helix structures in the membrane proteins of yeast are more susceptible to be lost in the proteomics analysis. This could be because of the hydrophobicity of the alpha-helix structures in the membrane proteins that with poor solubility in the aqueous media.

In transmembrane peptides, the polar functional groups form hydrogen bonding in the inner part alpha helical structure. The amino acids side chains are pointing out of the helical structure[21]. The hydrophobicity of the side chains of amino acids will determine the hydrophobicity of the outer shell of the helical structure which is in contact with its lipid bilayer environment in the cell membrane. In other words, the more hydrophobic side chains on the amino acids of the helical structure, the more hydrophobic the structure becomes. Since the probability of finding amino acids with hydrophobic side chains like Ala, Leu, and Met (which is noted as A, L and M in Figure 2-15) is higher than other amino acids[22], usually, the alpha helix structures have more hydrophobicity on the outer shell compared to other secondary structures of proteins.

Like lipid bilayer in the cells, the H-O phase in the FAOS is more hydrophobic than the aqueous phase due to the high organic content. It is expected that alpha-helix structures be enriched in the H-O phase due to this property. To confirm this, an amino acid analysis of the sequences of peptides in each phase has been performed to evaluate the abundance of the amino acids in each phase. Amino acid analysis of the peptides shows that the abundance of Ala, Leu, and Met (A, L and M in Figure 2-15) are higher in the H-O phase which is indicative of the fact that α-helix peptides are enriched in the H-O phase. Figure 2-15 shows that for all systems under study the abundance of Ala, Leu, and Met (A, L and M in Figure 2-15) is higher in the sequences found in the H-O phase.


Butanone _ unique in each phase

DMSO _ unique in each phase

THF _ unique in each phase

Figure 2-15. In all systems under study the abundance of Ala, Leu, and Met (A, L and M) is higher in the sequences found in H-O phase.

In addition to amino acid analysis of the peptides identified in each phase, a structural analysis of the peptides has been performed as well. A sophisticated program written in R language has been used to predict secondary structure of the peptides in the corresponding protein.

A database that contains the secondary structure of all proteins in the yeast was provided by NetSurfP-2.0. To find out the secondary structure of the peptide that has been found in the sample, the ypssc first reconstruct the protein by finding and arranging all the peptides related to the protein. Reconstructed protein may have missing sequences which does not affect further calculation. Then ypssc performs a side-by-side comparison between the reconstructed protein and the and the database and by doing so can calculate the secondary structure coverage in the protein found in the sample. ypssc released in The Comprehensive R Archive Network (CRAN), (https://cran.r-project.org/web/packages/ypssc/index.html).

Calculation for the peptide sequences that have been found in each phase of all three systems shows that the percentage of alpha-helix structures is significantly higher in the H-O phases of all systems as compared to aqueous phase and control. This is a robust evidence that the H-O system has the capacity for enriching the hydrophobic structures like alpha-helices.

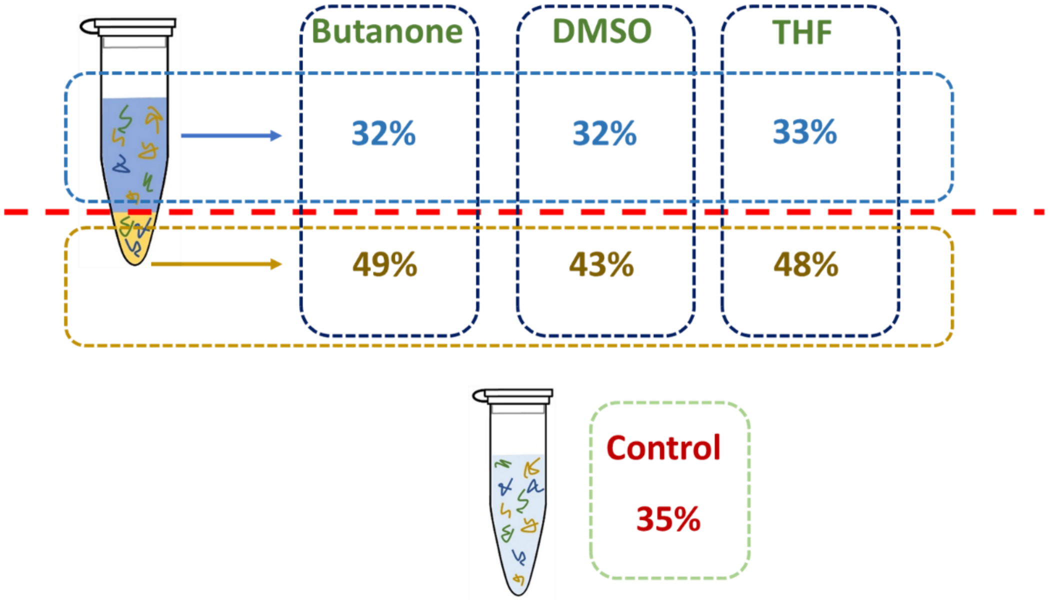


Figure 2-16. Structural analysis of the peptides found in each phase of all systems shows that the percent content of α-helix structure is much higher in the H-O phase compared to aqueous and control

Results also show that, due to the enrichment of alpha-helical peptides into the H-O phase, the total number of identified proteins with alpha-helix structure in all systems are higher than that in the control. This difference is greater in butanone and DMSO systems where 267 and 262 more proteins (compared to control) with alpha-helix structure have been identified, respectfully.

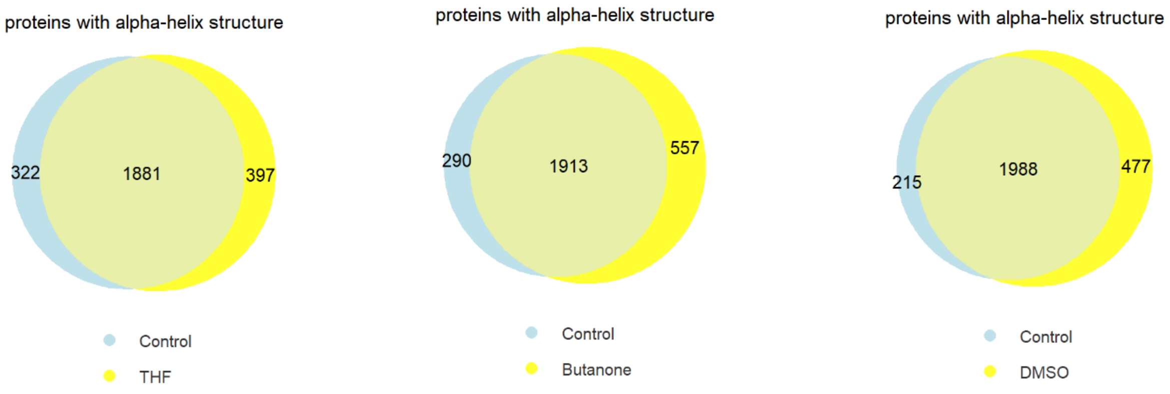
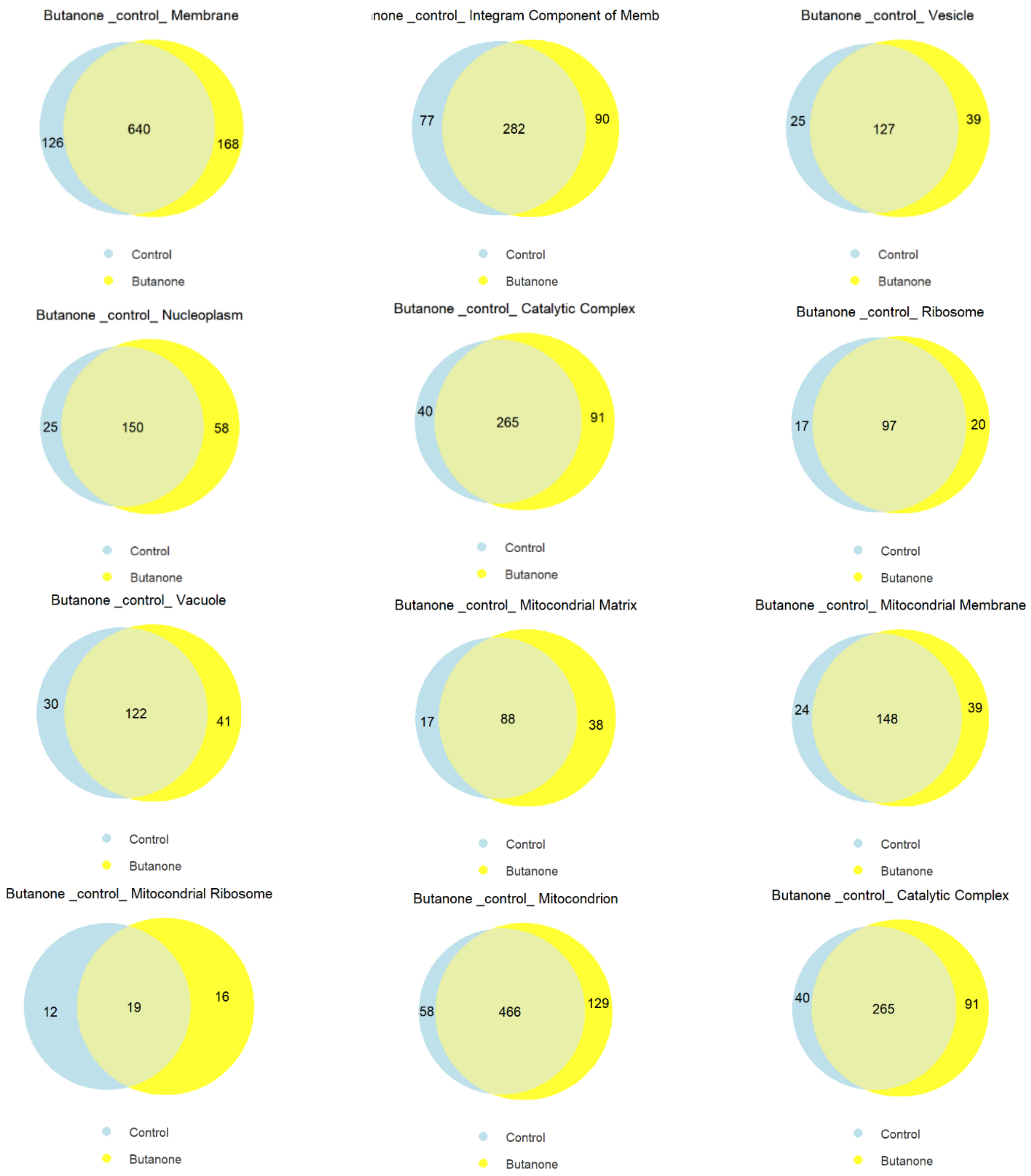


Figure 2-17. Total number of proteins with α-helix structure that have been identified in each system compared to control

In addition to the total number of identified alpha-helical proteins in the whole sample, the number of identified alpha-helix proteins in different cellular locations derived from the gene ontology in each system has been examined and compared with control. There is an improvement in the number of identified alpha-helix proteins in different cellular locations. Figure 2-18 and Figure2-19 shows the number of identified α-helix proteins in major yeast cells.

It is noteworthy that for the membrane and integral component of membrane groups, there is a considerable improvement in the number of identified α-helix proteins. This improvement is a result of the enrichment of alpha helix structures in the H-O phase.



Butanone _control_ Membrane

ınone _control_ Integram Component of Memb

Butanone _control_ Vesicle

Butanone _control_ Nucleoplasm

Butanone _control_ Catalytic Complex

Butanone _control_ Ribosome

Butanone _control_ Vacuole

Butanone _control_ Mitocondrial Matrix

Butanone _control_ Mitocondrial Membrane

Butanone _control_ Mitocondrial Ribosome

Butanone _control_ Mitocondrion

Butanone _control_ Catalytic Complex

Figure 2-18. Number of identified α-helix proteins in major yeast gene ontologies, a comparison between control and butanone sample.

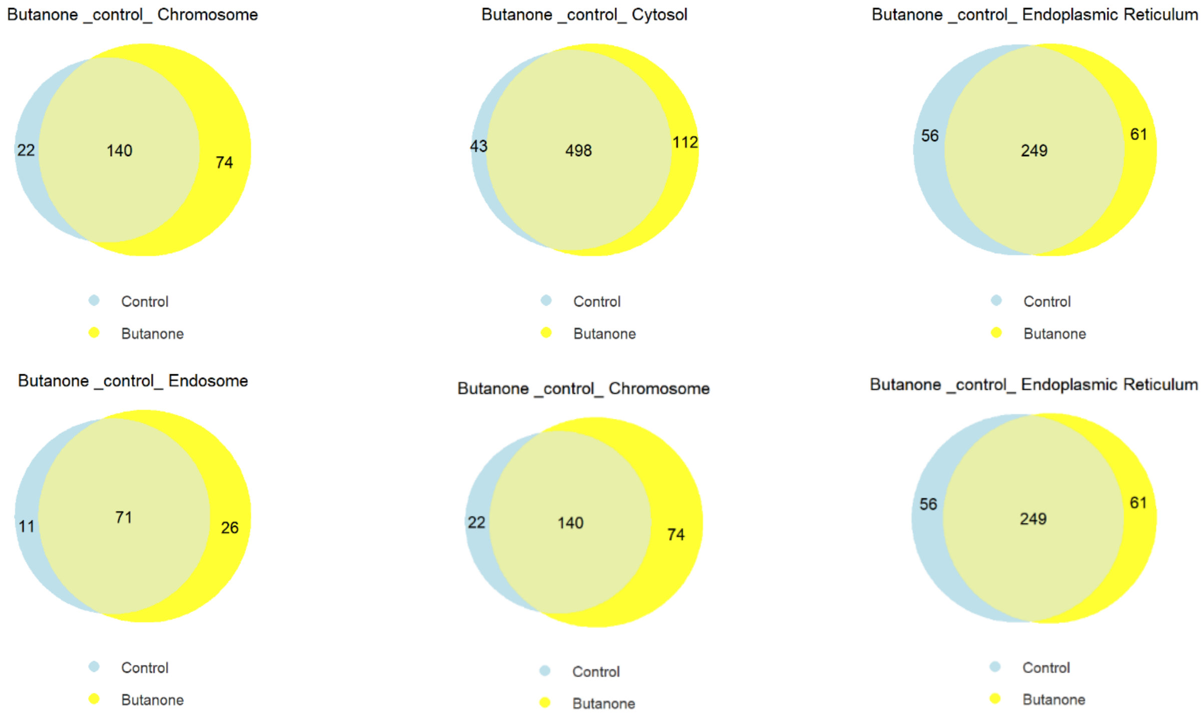
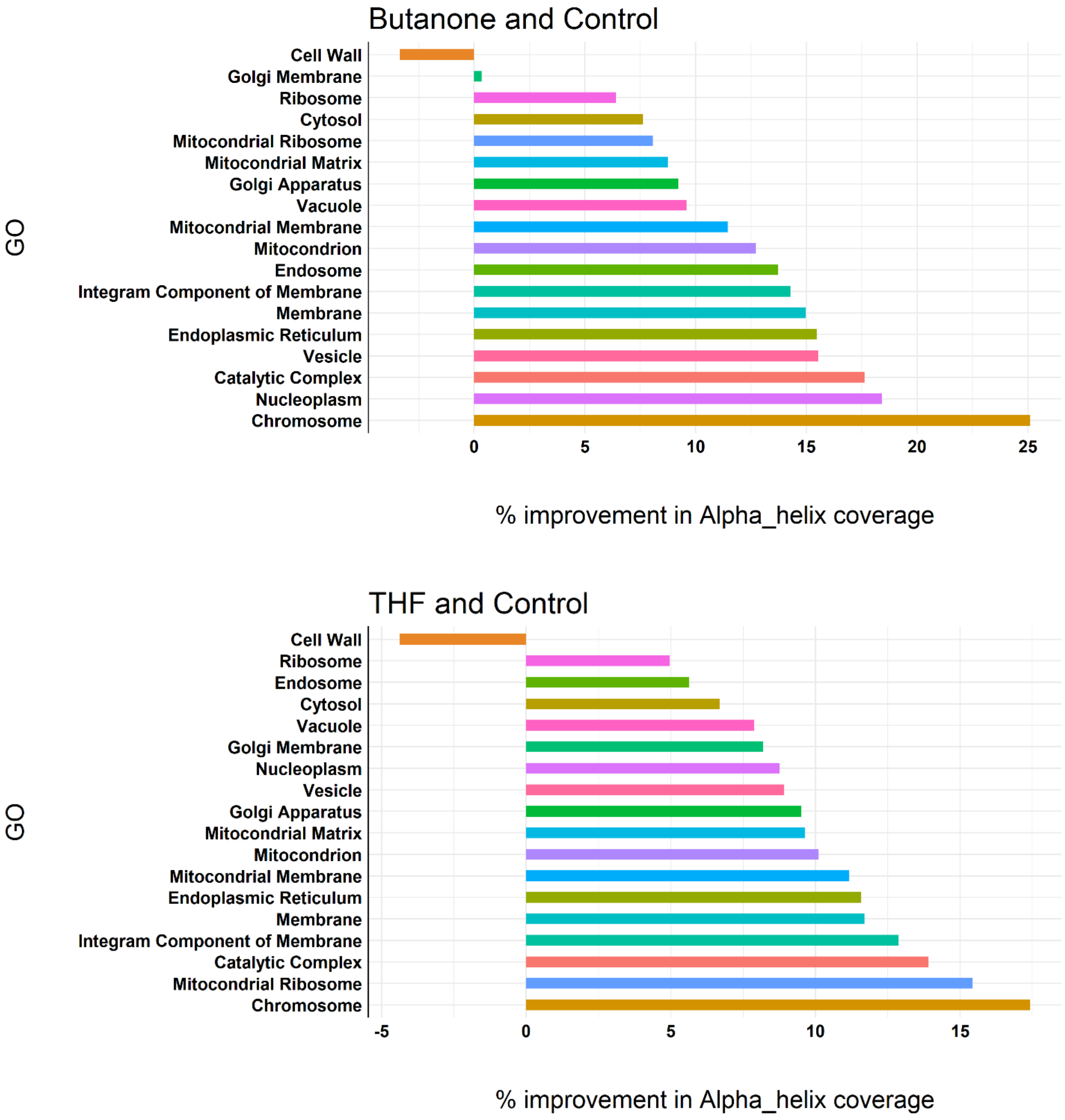


Figure 2-19. Number of identified α-helix proteins in major yeast gene ontologies, a comparison between control and butanone sample.

The analysis that mentioned earlier, shows the number of proteins that have α-helix structure. This abnalysis does not provide information about the length of the α-helix part of the protein that has been identified. To understand how much of α-helix part of the protein identified compared to other parts of proteins, the coverage in the α-helix part of the identified α-helix proteins of each gene ontology calculated as well. This analysis helps to understand the depth of improvement in identification of α-helix proteins.

The results show that for butanone and THF samples, in all gene ontologies, except the cell wall, there is considerable improvement in the coverage of alpha-helix part of the identified proteins with alpha-helix structure compared to control. It is noticable that, as compared to the control, the coverage of alpha-helix in the membrane proteins improved by 15% and 12% respectively in the butanone and THF systems. Figure 2-20 and Figure 2-21 show the % improvement in α-helix coverage of proteins of major gene ontologies compared to control.



Butanone and Control

THF and Control

Figure 2-20. Percent improvement in α-helix coverage of proteins of major gene ontologies compared to control.

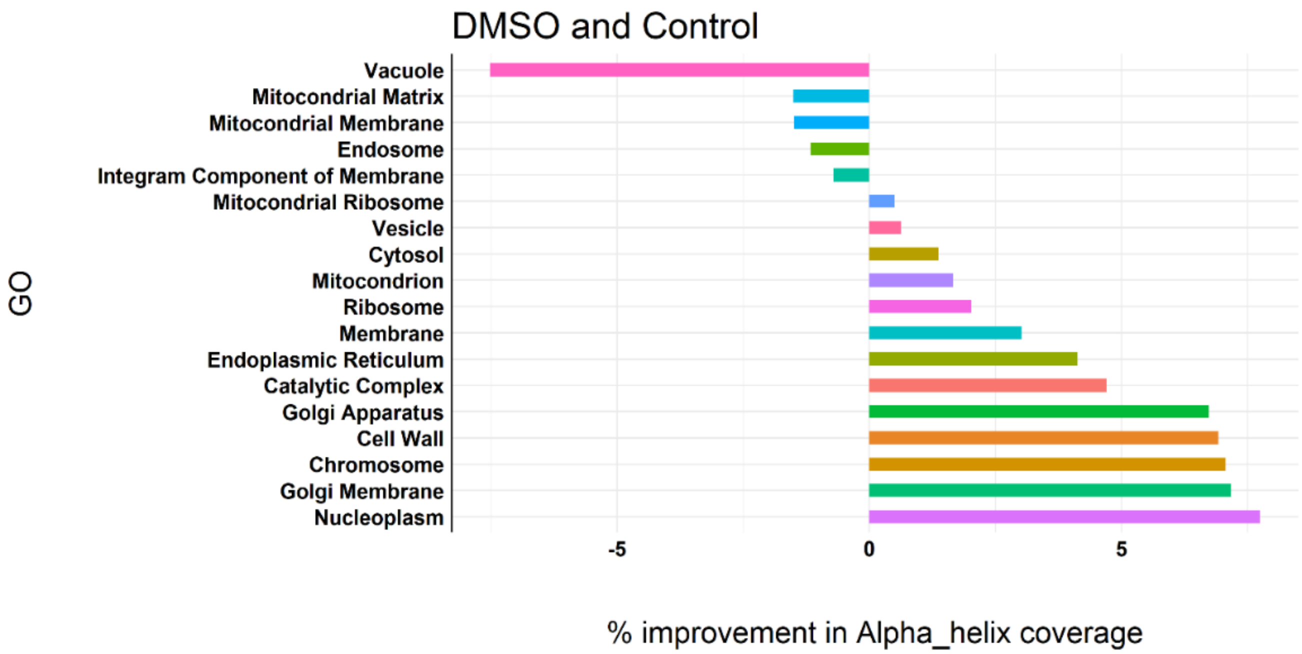


Figure 2-21. Percent improvement in α-helix coverage of proteins of major gene ontologies compared to control.

For butanone sample, in all gene ontologies that was examined there was considerable improvement in α-helix coverage of proteins compared to control sample. The highest improvement belongs to chromosome category which is 25% more than control. This is a significant improvement for this category of gene ontology and opens an oppotunity for the researchers that are focusing on the proteomics of this gene ontology. For the nucleoplasm, catalytic complex, vesicle, endoplasmic reticulum and most importantly membrane, there is improvement more than 15%. in α-helix coverage compared to control. For integral component of membrane there is almost 14% improvement. Improvements in α-helix coverage of category of membrane proteins are every important since α-helix structures are very dominant and improvement will directly improve the total coverage of proteins.

For THF sample, similar to butanone sample, chromosome has the largest value of improvement among the gene ontologies that are investigated with over15% improvement. For all gene ontologies that are examined, the minimum improvement is almost 5% except cell wall category. Especial attention is always toward membrane proteins because of their importance in the cell function and therapeutic treatment. In the THF sample both for integral component of membrane and membrane there are over 10% improvement in coverage of α-helix structures compared to control.

For the DMSO sample, however, the α-helix coverages are not as big as the other two systems mentioned earlier. Among 16 gene ontologies, for 5 of them there was not any improvement compared to control. For nucleoplasm, golgi membrane, chromosome, cell wall and golgi apparatus there is over 5% improvement. Interesting fact is this is the only system that is showing improvement in the cell wall category.

2.4 CONCLUSION

In bottom-up proteomics, especially in case of whole cell proteomics, the number of peptides is so high that it may overwhelm the mass spectrometer and leads to poor detection of peptides. in this document we proposed a method for frcationation of peptides using FOAS. In this method peptides frcationated into two-phase in which the bottom phase extrcats and enriches the hydrophobic peptides more than hydrophilic peptides. this simple fractionation step proven to be effctive in enrichement of α-helical peptides which are more hydrophobic than other peptides due to the spetial shape and types of amino acids in that structure. Enrichement of hydrophobic peptides lead to better identification of proteins with α-helical structures. For membrane proteins that are very likely to have α-helical structure in them we saw 15% improvement in identification of α-helical part of the proteins. The chamical used in FOAS does not have interference with ESI-MS and eliminates need for sample clenup prior to mass spectrometry. Three organic solvents used to form FOAS and butanone shows superior results compared to other solvents in sense of %improvement in identification of α-helical structures and identification of phosphorylated proteins.

2.5 REFERENCES

(1) Ly, L.; Wasinger, V. C. Protein and peptide fractionation, enrichment and depletion: tools for the complex proteome. *Proteomics* **2011**, *11* (4), 513-534. DOI: 10.1002/pmic.201000394.

(2) Camerini, S.; Mauri, P. The role of protein and peptide separation before mass spectrometry analysis in clinical proteomics. *J Chromatogr A* **2015**, *1381*, 1-12. DOI: 10.1016/j.chroma.2014.12.035.

(3) Tao, D.; Qiao, X.; Sun, L.; Hou, C.; Gao, L.; Zhang, L.; Shan, Y.; Liang, Z.; Zhang, Y. Development of a highly efficient 2-D system with a serially coupled long column and its application in identification of rat brain integral membrane proteins with ionic liquids-assisted solubilization and digestion. *J Proteome Res* **2011**, *10* (2), 732-738. DOI: 10.1021/pr100893j.

(4) Zhu, M. Z.; Li, N.; Wang, Y. T.; Liu, N.; Guo, M. Q.; Sun, B. Q.; Zhou, H.; Liu, L.; Wu, J. L. Acid/Salt/pH Gradient Improved Resolution and Sensitivity in Proteomics Study Using 2D SCX-RP LC-MS. *J Proteome Res* **2017**, *16* (9), 3470-3475. DOI: 10.1021/acs.jproteome.7b00443.

(5) Zhang, J.; Xu, X.; Gao, M.; Yang, P.; Zhang, X. Comparison of 2-D LC and 3-D LC with post- and pre-tryptic-digestion SEC fractionation for proteome analysis of normal human liver tissue. *Proteomics* **2007**, *7* (4), 500-512. DOI: 10.1002/pmic.200500880.

(6) Deng, W.; Sha, J.; Plath, K.; Wohlschlegel, J. A. Carboxylate-Modified Magnetic Bead (CMMB)-Based Isopropanol Gradient Peptide Fractionation (CIF) Enables Rapid and Robust Off-Line Peptide Mixture Fractionation in Bottom-Up Proteomics. *Mol Cell Proteomics* **2021**, *20*, 100039. DOI: 10.1074/mcp.RA120.002411.

(7) Lau, E.; Lam, M. P.; Siu, S. O.; Kong, R. P.; Chan, W. L.; Zhou, Z.; Huang, J.; Lo, C.; Chu, I. K. Combinatorial use of offline SCX and online RP-RP liquid chromatography for iTRAQ-based quantitative proteomics applications. *Mol Biosyst* **2011**, *7* (5), 1399-1408. DOI: 10.1039/c1mb05010a.

(8) Kumar, R.; Shah, R. L.; Rathore, A. S. Harnessing the power of electrophoresis and chromatography: Offline coupling of reverse phase liquid chromatography-capillary zone electrophoresis-tandem mass spectrometry for peptide mapping for monoclonal antibodies. *J Chromatogr A* **2020**, *1620*, 460954. DOI: 10.1016/j.chroma.2020.460954.

(9) Gundry, R. L.; White, M. Y.; Murray, C. I.; Kane, L. A.; Fu, Q.; Stanley, B. A.; Van Eyk, J. E. Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow. *Curr Protoc Mol Biol* **2009**, *Chapter 10*, Unit10.25. DOI: 10.1002/0471142727.mb1025s88.

(10) Koolivand, A.; Clayton, S.; Rion, H.; Oloumi, A.; O'Brien, A.; Khaledi, M. G. Fluoroalcohol - Induced coacervates for selective enrichment and extraction of hydrophobic proteins. *J Chromatogr B Analyt Technol Biomed Life Sci* **2018**, *1083*, 180-188. DOI: 10.1016/j.jchromb.2018.03.004.

(11) Khaledi, M. G.; Jenkins, S. I.; Liang, S. Perfluorinated alcohols and acids induce coacervation in aqueous solutions of amphiphiles. *Langmuir* **2013**, *29* (8), 2458-2464. DOI: 10.1021/la303035h.

(12) Nejati, M. M.; Khaledi, M. G. Perfluoro-alcohol-induced complex coacervates of polyelectrolyte-surfactant mixtures: phase behavior and analysis. *Langmuir* **2015**, *31* (20), 5580-5589. DOI: 10.1021/acs.langmuir.5b00444.

(13) Koolivand, A.; Azizi, M.; O'Brien, A.; Khaledi, M. G. Coacervation of Lipid Bilayer in Natural Cell Membranes for Extraction, Fractionation, and Enrichment of Proteins in Proteomics Studies. *J Proteome Res* **2019**, *18* (4), 1595-1606. DOI: 10.1021/acs.jproteome.8b00857.

(14) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **1982**, *157* (1), 105-132. DOI: 10.1016/0022-2836(82)90515-0.

(15) Ramazi, S.; Zahiri, J. Posttranslational modifications in proteins: resources, tools and prediction methods. *Database (Oxford)* **2021**, *2021*. DOI: 10.1093/database/baab012.

(16) Walsh, G.; Jefferis, R. Post-translational modifications in the context of therapeutic proteins. *Nat Biotechnol* **2006**, *24* (10), 1241-1252. DOI: 10.1038/nbt1252.

(17) Cohen, P. The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur J Biochem* **2001**, *268* (19), 5001-5010. DOI: 10.1046/j.0014-2956.2001.02473.x.

(18) Reily, C.; Stewart, T. J.; Renfrow, M. B.; Novak, J. Glycosylation in health and disease. *Nat Rev Nephrol* **2019**, *15* (6), 346-366. DOI: 10.1038/s41581-019-0129-4.

(19) Futcher, B.; Latter, G. I.; Monardo, P.; McLaughlin, C. S.; Garrels, J. I. A sampling of the yeast proteome. *Mol Cell Biol* **1999**, *19* (11), 7357-7368. DOI: 10.1128/MCB.19.11.7357.

(20) Klausen, M. S.; Jespersen, M. C.; Nielsen, H.; Jensen, K. K.; Jurtz, V. I.; Sønderby, C. K.; Sommer, M. O. A.; Winther, O.; Nielsen, M.; Petersen, B.; et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins* **2019**, *87* (6), 520-527. DOI: 10.1002/prot.25674.

(21) Robinson, S. W.; Afzal, A. M.; Leader, D. P. Handbook of Pharmacogenetics and Stratified Medicine. Academic Press, 2014; pp 259-287. Boyle, A. L. Applications of de novo designed peptides. In *Peptide Applications in Biomedicine, Biotechnology and Bioengineering*, Koutsopoulos, S. Ed.; Woodhead Publishing, 2018; pp Pages 51-86.

(22) Clark, D. P.; Pazdernik, N. J.; McGehee, M. R. *Molecular biology*; Academic Cell, 2018.

CHAPTER 3

IMPROVING IDENTIFICATION OF LOW ABUNDANCE PROTEINS BY FRACTIONATION IN BIPHASIC SYSTEMS OF MIXED SURFACTANTS

Used with permission from Sajad Tasharofi, Mohammadmehdi Azizi, Morteza G. Khaledi

3.1 ABSTRACT

whole cell proteomics of yeast using bottom-up method is very challenging due to the complexity of the system. offline fractionation of proteins prior to digestion will result in plainer fractions that would decrease the complexity of the system and improve protein identification. We used a biphasic system to perform yeast proteome fractionation prior to digestion; the biphasic system is formed by addition of HFIP to aqueous solution of SDC and TBAB. Fractionation of proteins in such biphasic system resulted in 40% improvement in identification of proteins with abundance less than 2000 molecule per cell. Since SDC and TBAB have opposite charges, changes in the ratio between them can determine the fractionation pattern of proteins based on the pI value. Overall, this fractionation method can improve total number of identified proteins compared to control.

3.2 INTRODUCTION

In whole cell proteomics, preferred method of protein analysis is bottom-up proteomics[1-3]. In bottom-up proteomics, especially if it performed for whole cell proteomics, the mixture of peptides from digestion of proteins is very complex; in some cases fractionation of proteins or peptides prior to mass spectrometry analysis seems to be inavitable[4]. Reverse phase chromatography and ion exchange chromatography are usually performed to separate the complex mixture of proteins or peptides before mass spectrometry analysis[5-7]. In addition to chromatographic methods, the value of fractionation and enrichment of proteins and peptides in bottom-up proteomics is very important.

Wide dynamic range of protein expression in the cells is one of the reasons that low abundance proteins are lost the workflow of bottom-up proteomics. There is four orders of magnitude difference between protein abundances in yeast proteome[8]; this will lead to poor identification of low abundance proteins specifically in the data dependent acquisition mode in mass spectrometry. In addition to mass spectrometry method used to analyze peptides, there are some other sources that suppress low abundance protein identification; this includes: co-elution of peptides in the chromatography that is performed prior to mass spectormetry[9]; suppression in ionization of peptides of low abundance proteins by high abundance peptides[10] and sensitivity of the instrument.

To avoid suppression of low abundance proteins due to complexity of the sample, the mixture of protein should be fractionated to simpler mixtures to improve identification of low abundance proteins. It has been proven that fractionation of proteins, as a supplement for reverse phase chromatography, can improve protein identification significantly[11-13]. The common methods of fractionation used for this purpose are: electrophoretic/ charge, Strong cation exchange (SCX)[14-

[16], C18[17] and fractionation based on the mass of protein or size exclusion chromatography (SEC)[18, 19].

Previously, we have developed a biphasic system composed of fluorinated alcohols, water and amphiphiles that would fractionate proteins based on different physicochemical properties. We have reported different compositions and types of amphiphiles and surfactants could be used to form such biphasic systems[20]; a thorough analysis of phase behavior has been performed on different phases to characterize them to the greatest extend[21]. The potential of such system as a good extraction media for membrane proteins has been shown before[22] and the potential of such system for extraction and enrichment of hydrophobic proteins has been demonstrated in previous reports[23]. As an alternative for use of surfactants, in our previous study, we used the natural lipids in yeast cell to form a biphasic system that helps with identification improvement in whole proteome studies of yeast[24]. Use of mixed surfactants like DMMAPS as a zwitterionic surfactant and tetra butyl ammonium bromide (TBAB) as a positively charged surfactant, shows considerable improvement in identification of low abundance proteins and α-helix structures[25]. The use of TBAB alone, resulted in the best improvement in the identification of low abundance proteins among all systems that we reported previously[26].

Although the combination of zwitterionic surfactants and positively charged surfactants has been studied, the effect of combination of a positively and a negatively charged surfactant has not been studied. As a continuation in this line of research, in this study, we have used a mixture of negatively and positively charged surfactant, in different ratios, to form a biphasic system to fractionate yest proteome and study the effect of fractionation on various aspects of the yeast whole cell proteomics. We used sodium deoxycholate (SDC) as a negatively charged surfactant and TBAB, as a positively charged surfactant.

3.3 EXPERIMENTAL

3.3.1 Materials, Chemicals and Reagents

Tetrabutylammonium bromide (TBAB) were purchased from ACROS Organics. 1,1,1,3,3,3-Hexafluoro-2-propanol (HFIP) was obtained from Oakwood Chemical, USA. Trifluoroacetic acid (99%) and formic acid (99%) were purchased from Alfa Aesar. Tris HCl and Tris base were purchased from Sigma-Aldrich. Dithiothreitol (DTT), Iodoacetamide (IAA) and Sequencing grade trypsin were purchased from Fisher BioReagents, Alfa Aesar™, and Promega Corporation, respectively. LC-MS acetonitrile (ACN), isopropanol (IPA) and deionized water were provided from Fisher Chemical, USA.

3.3.2 Biphasic system containing mixed surfactants

The two-phase system is formed by addition of HFIP to aqueous solution of mixed surfactants, SDC and TBAB; each sample has total volume of 1 mL, 8% HFIP, SDC and TBAB. Concentration of surfactants (SDC+TBAB) in all samples is 100 mM but the ratio between SDC and TBAB changes in different samples. First system has 30 mM SDC and 70 mM TBAB and we refer to this sample as SDC/TBAB-30/70, second system has 50 mM SDC and 50 mM TBAB and we refer to this sample as SDC/TBAB-50/50, and third system has 70 mM SDC and 30 mM TBAB and we refer to this sample as SDC/TBAB-70/30. These three systems are used to fractionate yeast whole proteome. Each sample contains 400 µg of yeast proteins which is equivalent of 50 µL of yeast cell lysate.

After addition of yeast cell lysate to aqueous solution of mixed surfactants, HFIP added to the mixture to induce the two-phase system. Then, the two phases have been separated and dried to remove HFIP that interferes with the workflow that follows in the protocol. Each phase has been reconstituted in 8 M urea and added to size exclusion filters for the surfactant removal. The method used for surfactant removal using size exclusion filters is called Filter Assisted Sample Preparation (FASP) which is explained in detail in the Appendix 3-2. For the control sample, which is not subject to phase separation, the yeast cell lysate is dissolved in the 8M urea and FASP protocol has been applied on it to keep the conditions of sample and control as identical as possible. Figure 3-1 demonstrates the workflow of sample preparation for the biphasic system and control.

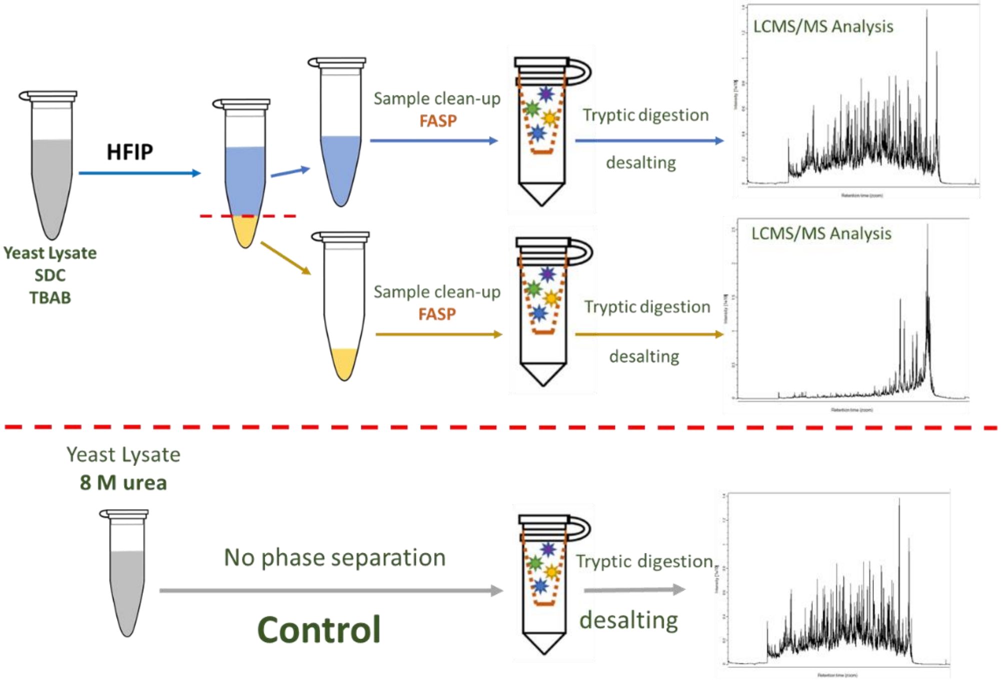


Figure 3-1. the workflow of sample preparation for mixed surfactants and control

3.3.3 TBAB and SDC concentration measurement

To provide a better understanding of the biphasic system, using mass spectrometry, concentration of SDC and TBAB have been measured in aqueous and coacervate in all 3 sample with different ratios of SDC and TBAB using mass spectrometry. Table 3-1 shows all 4 measurements for TBAB solutions of different concentrations.

Table 3-1. measurements for different solutions of TBAB

TBAB						
TBAB concentration in (µM)	**Measuremnt 1**	**Measuremnt 2**	**Measuremnt 3**	**Measuremnt 4**	**Average**	**Standard deviation**
0.1	5557968	5941405	5746256	5748543	5748543	156546
0.25	13946113	14165737	13998850	14036900	14036900	93611
0.5	29824808	29685132	28951406	29487115	29487115	383072
1	59547472	62368599	60985598	60967223	60967223	1151794
1.5	105679428	106337380	105104731	105707180	105707180	503609

Table 3-2. measurements for different solutions of SDC

SDC						
SDC concentration in (µM)	**Measuremnt 1**	**Measuremnt 2**	**Measuremnt 3**	**Measuremnt 4**	**Average**	**Standard deviation**
10	975268	1164213	1249328	1129603	1129603	114530
30	3668805	3554912	3627066	3616928	3616928	47046
50	3624062	4500697	3843033	3989264	3989264	372523
70	5896326	6797228	6541938	6411831	6411831	379124
90	7408321	7726560	8105021	7746634	7746634	284781

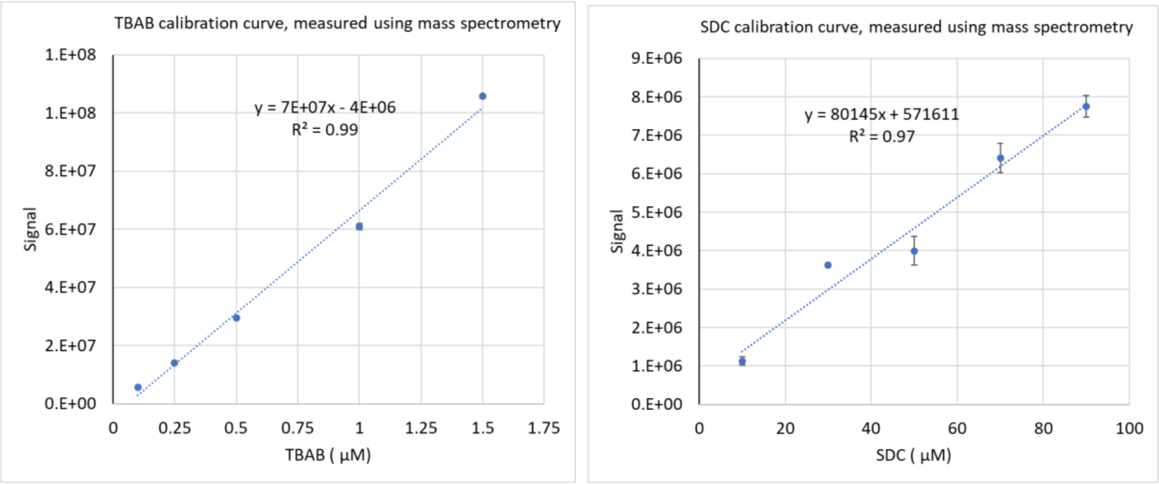
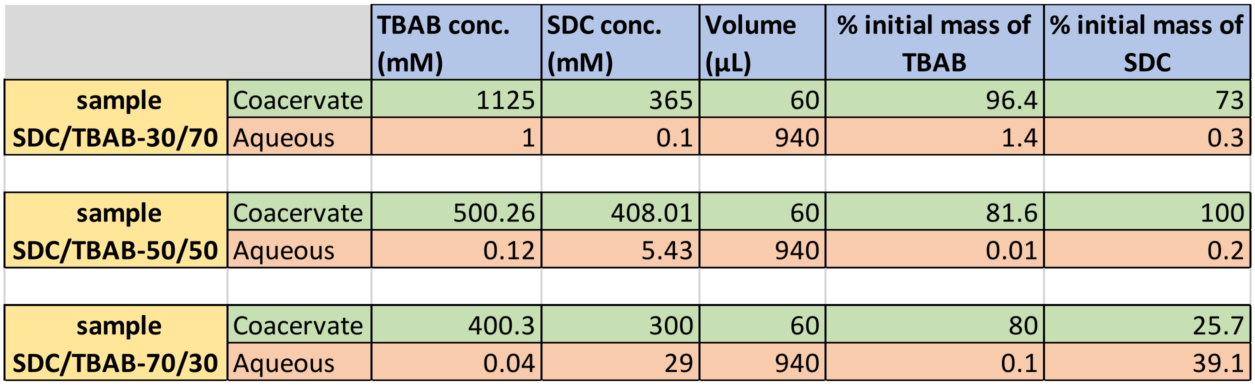


Figure 3-2. Calibration curve for TBAB and SDC, prepared using mass spectrometry measurement

From the calibration curves, the concentration of SDC and TBAB has been measured. Results show that the concentration of SDC and TBAB is much higher in the coacervate compared to aqueous phase. this difference is obvious in the SDC/TBAB-30/70 that concentration of TBAB 1000-times higher in the coacervate phase.

Table 3-3. TBAB and SDC measurements in all 3 samples with different compositions



		TBAB conc. (mM)	SDC conc. (mM)	Volume (µL)	% initial mass of TBAB	% initial mass of SDC
sample SDC/TBAB-30/70	Coacervate	1125	365	60	96.4	73
	Aqueous	1	0.1	940	1.4	0.3
sample SDC/TBAB-50/50	Coacervate	500.26	408.01	60	81.6	100
	Aqueous	0.12	5.43	940	0.01	0.2
sample SDC/TBAB-70/30	Coacervate	400.3	300	60	80	25.7
	Aqueous	0.04	29	940	0.1	39.1

3.3.4 LCMSMS analysis

After protein digestion of both phases of each sample, the peptides are desalted using a C18 cartridge and the peptides are reconstituted in 1% formic acid and the concentration of peptides after reconstitution is adjusted to 1 mg/mL. for LCMSMS analysis, 1 µL of the peptide solution has been injected to the chromatography column. The LC was Ultimate 3000 RSLC-Nano liquid chromatography systems, Dionex, coupled with an Orbitrap Fusion Lumos MS, Thermo Electron instrument. A C18 nano column was used for the separation with a column length of 75 cm, internal diameter of 75 µm, and particle size of 3 µm. The solvent system used was 0−90 min gradient run with 0−28% of solvent B, and a flow rate of 350 nL/min. Mobile phase A was 2%

(v/v) acetonitrile (ACN) and 0.1% (v/v) formic acid (FA) in water, and mobile phase B was 80% (v/v) ACN, 10% (v/v) trifluoroethanol (TFE), and 0.1% FA in water.

3.3.5 Data Analysis of mass spectrometry

The raw data collected from orbitrap instrument has been analyzed using MaxQuant (Ver. 1.6.2.3). FASTA file from UniPort based identification, trypsin digestion, oxidation of methionine, and N-terminal acetylation as variable modifications (maximum 5 modifications per peptide), carbamidomethyl as a fixed modification, 2 missed cleavages, label-free quantification with iBAQ (intensity-based absolute quantification), minimum one unique peptide for protein identification, PSM FDR (peptide-spectrum match false discovery rate) 1%, and protein (false discovery rate) FDR was 1%. Three replicates were carried out. The common proteins in two out of three runs were taken for further data analysis. Further data analysis was done using the UniProt database, Yeast Mine database, and Gene Ontology database to obtain more information for extracted proteins. Gene Ontology (GO) annotation for yeast database is based on gene code from Saccharomyces Genome Database (SGD) (http://www.geneontology.org).30 The SGD protein IDs for extracted proteins were retrieved from the UniProt database (http://www.uniprot.org). The information related to the abundance value of proteins and post-translational modification was extracted from the Yeast Mine database.

3.4 RESULTS AND DISCUSSION

For all 3 systems under study, SDC/TBAB-30/70, SDC/TBAB-50/50, and SDC/TBAB-70/30 after protein fractionation, digestion and mass spectrometric analysis, the data has been analyzed in the platform of open-source software (MaxQuant) for peptide and protein search, and for further data analysis, an in-house R program developed for analysis of results of MaxQuant. The R program that developed for this purpose, analyzes results of MaxQuant in various aspects like elimination of peptides and proteins with false identification, filtering list of proteins of each sample and writing it in a excel sheet, comparing the list of proteins of each sample as well as proteins identified in both phases of a sample, calculation of coverage of α-helix parts of the proteins, visualization including Euler diagrams, bar charts, and violin graphs, are among them. In this chapter, all aspects of proteomics data analysis that performed for each sample presented and the results of samples compared.

3.4.1 Number of Identified proteins in the samples and fractionation pattern of proteins of each sample between the phases

The first thing analysis is the number of proteins identified in each phase of each sample, number of proteins identified in each phase of each sample sample as a whole and comparison between the number of proteins identified in each sample (identified proteins in the aqueous phase + coacervate phase) and number of proteins identified in the control sample which has no phase separation. It is important to mention that for sample SDC/TBAB-50/50, we observed that almost all the proteins that initially was in the sample was extracted to the coacervate phase after formation of two-phase.

This was confirmed later on in the workflow by measuring peptide concentration of aqueous phase which showed that the peptide concentration is so low that is not with UV/Vis measurements. For this sample, only the coacervate phase has been analyzed; the proteins and all data that reported in this document from sample SDC/TBAB-50/50, are the proteins identified in the organic phase Number of Identified proteins in the samples and fractionation pattern of proteins of each sample between the phases shown in Figure 3-3.

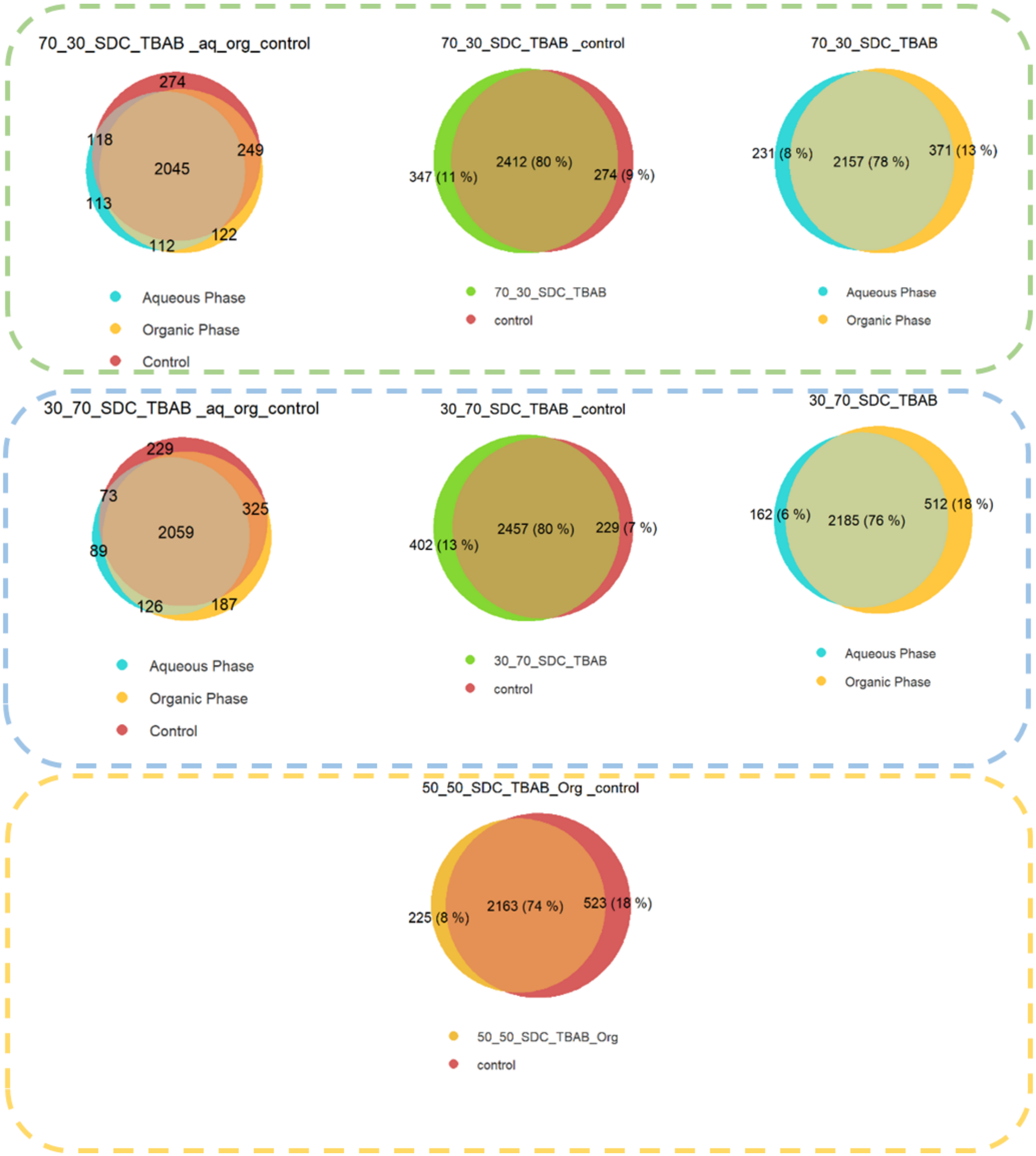


Figure 3-3. Green box shows the data for sample SDC/TBAB-70/30, from left: Euler diagram showing number of proteins in control, aqueous and coacervate phase; number of proteins in control and the sample; number of proteins in the aqueous and coacervate phase of the sample.

Blue box shows the data for sample SDC/TBAB-30/70, from left: Euler diagram showing number of proteins in control, aqueous and coacervate phase; number of proteins in control and the sample; number of proteins in the aqueous and coacervate phase of the sample.

Yellow box shows the data for sample SDC/TBAB-50/50. Euler diagram showing number of proteins in control and the sample

The number of identified proteins in the SDC/TBAB-30/70 sample is 2859 which is 6.4% higher than the number of proteins identified in the control sample which has no protein fractionation. It is also noticeable that the number of proteins identified in the coacervate phase is more than the aqueous phase and control sample. The number of identified proteins in the SDC/TBAB-70/30 sample is 2759 which is 2.7% higher than the number of proteins identified in the control sample which has no protein fractionation. It is also noticeable that the number of proteins identified in the coacervate phase is more than the aqueous phase and control sample. The number of identified proteins in the coacervate phase of SDC/TBAB-50/50 sample is 2388 which is 11% lower than the number of proteins identified in the control sample which has no protein fractionation.

The breakdown of yeast gene ontology will result in 18 major gene ontology which are chosen for the data analysis of proteins in this document. For each gene ontology, there are a certain number of proteins and in the data analysis, the number of proteins from each gene ontology that identified in each phase of each sample determined. In addition, number of identified proteins of each gene ontology in samples and control compared.

Figure 3-4 shows theses results for the sample SDC/TBAB-30/70. Among 18 gene ontologies that investigated, there was improvement in 16 gene ontologies. Figure 3-4 shows the category of gene ontologies that there was improvement in identification of proteins for the sample SDC/TBAB-30/70 as well as the number of proteins in the sample and the control. Euler diagrams also help to understand the number of proteins that are unique to the sample or control.

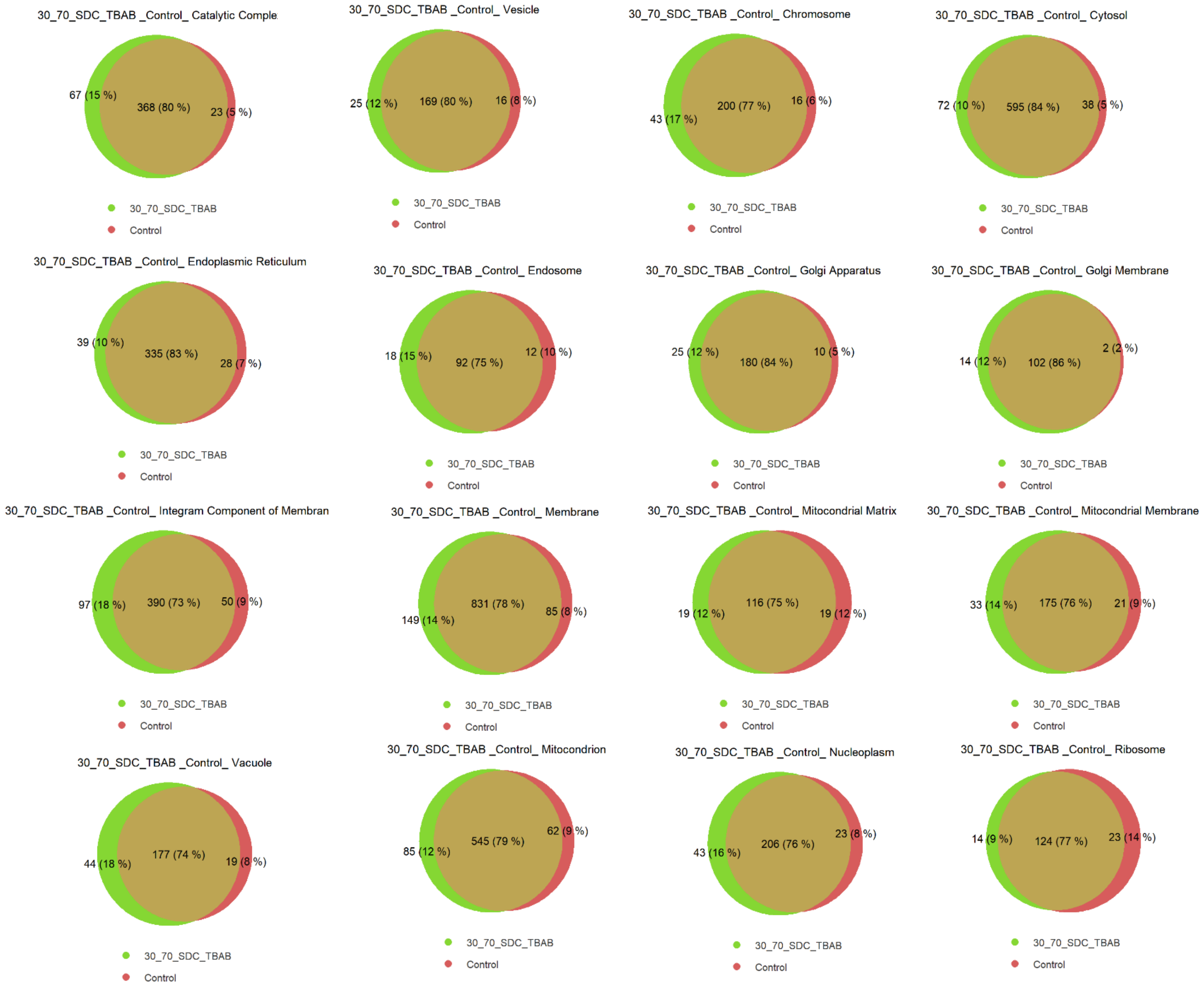


Figure 3-4. Euler diagrams showing fractionation of proteins of each gene ontology identified in the SDC/TBAB-30/70 and control

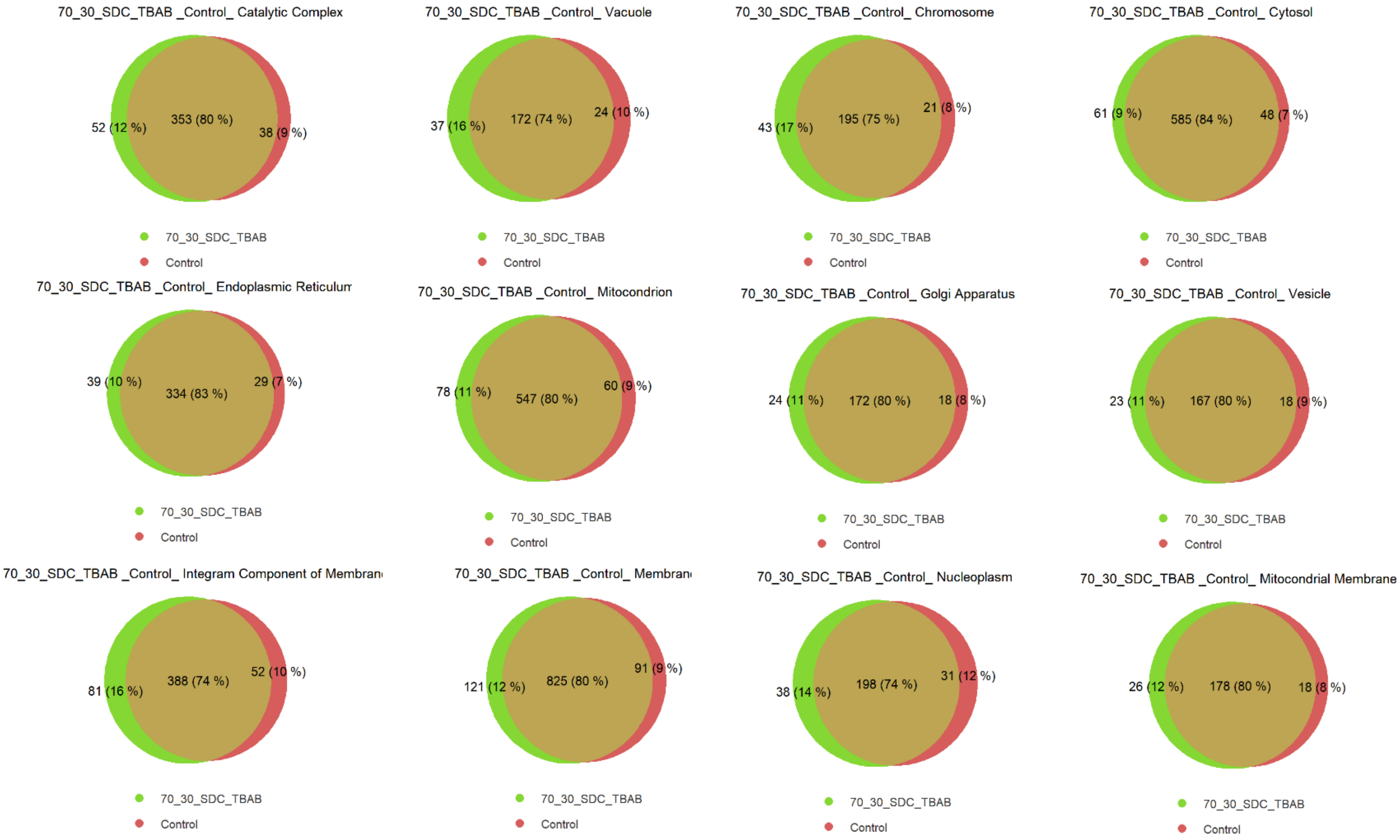


Figure 3-5. Euler diagrams showing fractionation of proteins of each gene ontology identified in the SDC/TBAB-70/30 and control

Figure 3-5 shows theses results for the sample SDC/TBAB-70/30. Among 18 gene ontologies that investigated, there was improvement in 12 gene ontologies. Figure 3-5 shows the category of gene ontologies that there was improvement in identification of proteins for the sample SDC/TBAB-70/30 as well as the number of proteins in the sample and the control. Euler diagrams also help to understand the number of proteins that are unique to the sample or control.

3.4.2 Improvement in identification of proteins with α-helix structure

To determine the which part of identified proteins is in fact α-helix structure, NetSurfP-2.0 utilized to perform a prediction. NetSurfP-2.0 is a prediction tool for secondary structures using neural network[32]. NetSurfP-2.0 is an extension of NetSurfP-1.0 which utilized deep neural network to predict secondary structures with the accuracy of 85%. In addition to accuracy, this tool presents reduced computational time compared to other methods[32].

In bottom-up proteomics, the proteins are enzymatically digested to peptides; sometimes this process produces hundreds of thousands of peptides. First, NetSurfP-2.0 is not designed to accept as many peptides at once, therefore the process of uploading the sequences and waiting for the calculations to be complete is extremely time consuming. Second, even if all sequences uploaded successfully and the results are back, it would be almost impossible to combine the results that have been produced for each individual peptide (hundreds of thousands of spread sheets) to get a coherent picture of the secondary structure of the proteins.

To solve this problem, an extension for NetSurfP-2.0 developed in-house which is specifically designed to analyze the results of bottom-up proteomics that has primarily analyzed with MaxQuant. We call this tool Yeast Proteome Secondary Structure Calculator (ypssc). This tool is written in R language at it is launched into The Comprehensive R Archive Network (CRAN) which would make it easy to use by the user even no knowledge of programming.

YPSSC, on one hand benefits forms the accuracy of NetSurfP-2.0 to calculate secondary structure and on the other hand addresses the issue of analyzing so many peptides with NetSurfP-2.0 by eliminating the need for direct analysis of the peptides from bottom-up proteomics.

Using the database provided by NetSurfP-2.0, we calculated that α-helix structure makes up to 38% of total yeast proteome sequences and 42 % of yeast membrane proteins. Using ypssc and the data from control sample in our study shows that 36% of yeast proteome is α-helix, which is very close to NetSurfP-2.0 prediction. However, for membrane proteins only 35% of sequences that have been identified are alpha helices. This shows that α-helix structures in the membrane proteins of yeast are more susceptible to be lost in the workflow of proteomics.

In α-helix structures, the polar functional groups from hydrogen bonding to from α-helix structure; this would place them in the inner part of the helix structure. The side chains of amino acids in the α-helix structure are pointing out of the helical structure[28-30]; hydrophobicity of the side chains will determine the hydrophobicity of the outer shell of the helical structure. Abundance of amino acids with hydrophobic side chains like Ala, Leu, and Met in the α-helix structure is higher than other amino acids[31] which would result in more hydrophobicity on the outer shell of the α-helix structure. Due to this property, we expect that proteins with α-helix structure would prefer to be extracted to the coacervate phase which has more hydrophobic nature compared to aqueous phase.

SDC and TBAB are known to be denaturing agent for proteins. Improved denaturation of protein in presence of those agents can improve digestion of those proteins and as a result improves identification of those proteins.

In addition, extraction, and enrichment of α-helix proteins into coacervate can improve their identification. An analysis performed to determine the number of proteins with α-helix structure in each sample and control.

Figure 3-6 is showing the Euler diagrams comparing α-helix proteins identified in each sample with control.

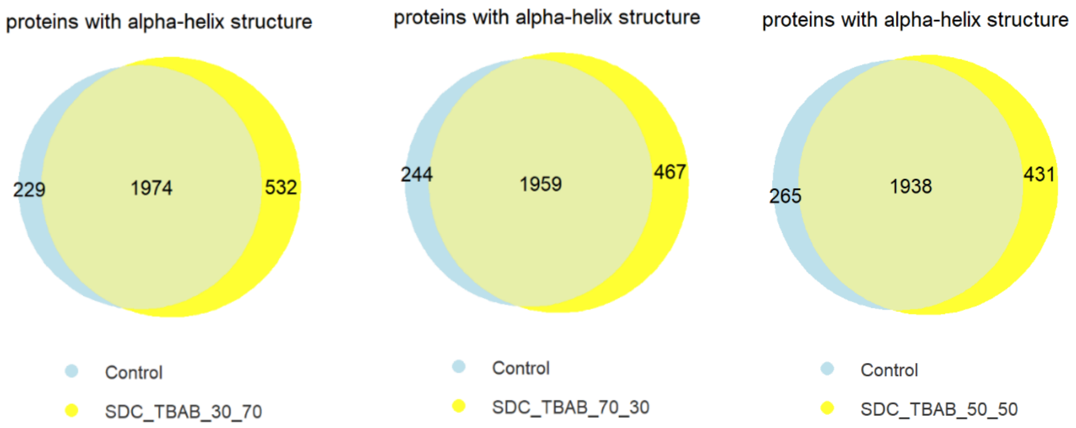


Figure 3-6. Euler diagrams comparing α-helix proteins identified in each sample with control

In all 3 samples there is considerable improvement in the number of α-helix proteins. Respectively, for SDC/TBAB-30/70, SDC/TBAB-50/50, and SDC/TBAB-70/30 there are 14%, 8% and 10% improvement in the number of identified α-helix proteins compared to control. It is interesting to find that for SDC/TBAB-50/50 sample, the number of α-helix proteins is 166 more than control; this sample is consist of only coacervate phase and this shows that enrichment of proteins to coacervate phase improved identification of α-helix proteins by 10%.

In addition, the percentage coverage of the α-helix structure that has been identified in the sample has been calculated. Major gene ontologies of yeast selected and percent coverage of α-helix for each category calculated and compared with control.

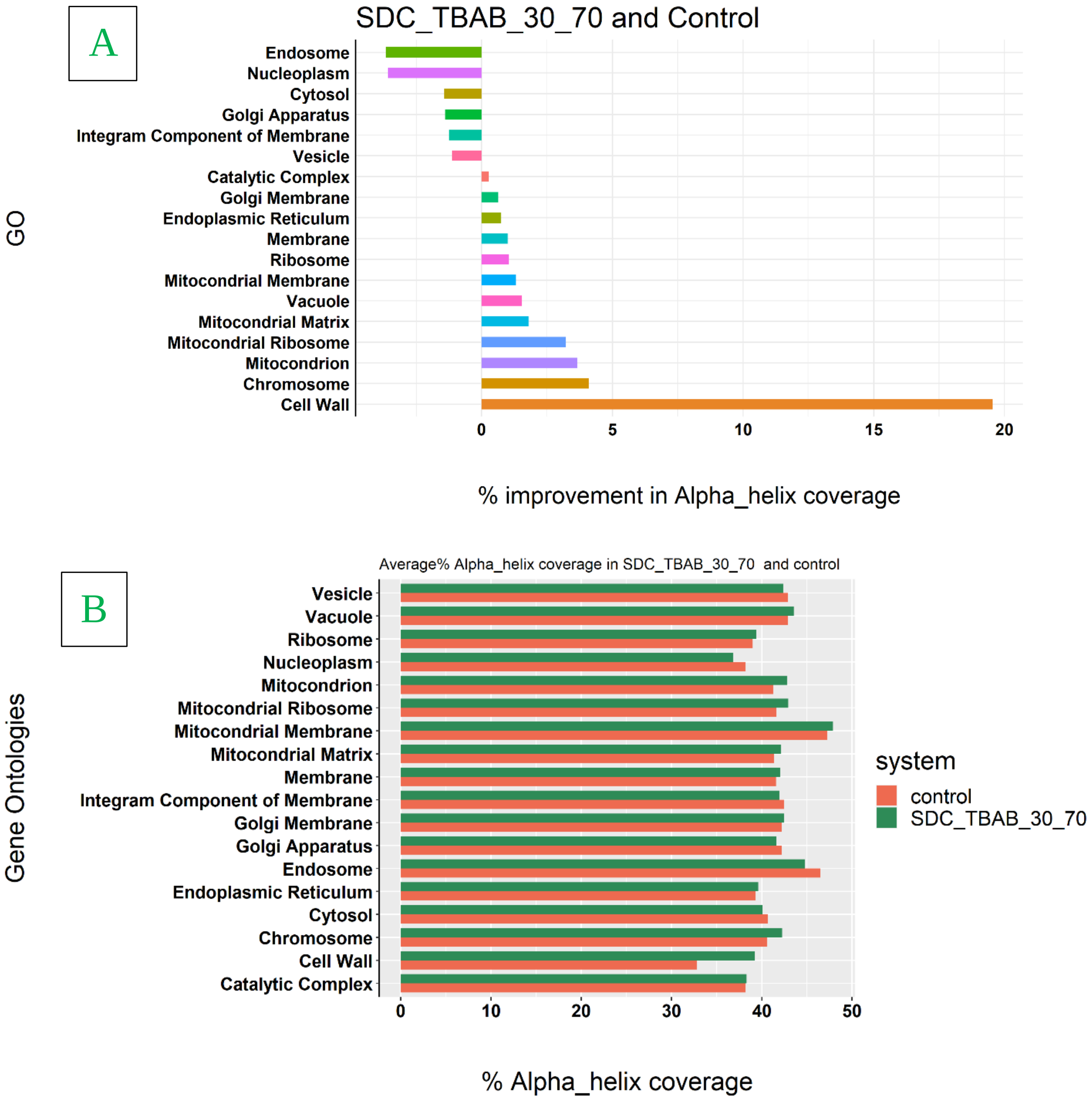


Figure 3-6. A: %improvement in α-helix coverage of main gene ontologies, a comparison between sample and the control; B: same graph but showing the absolute % α-helix coverage

Figure 3-7 shows the %improvement in coverage of alpha helix in the sample SDC/TBAB-30/70 compared to control. Results show that there is almost 20% improvement in identification of α-helix part of cell wall proteins. Cell wall in yeast is composed of proteins and sugars[33] and the combination of SDC and TBAB probably is very effective in dissolving this glycoprotein structure that wraps around the yeast cell. This is an interesting finding for the researchers that are studying the compositional of cell wall in yeast because this result show how effective this system is in dismantling the cell wall and release the proteins and make them accessible to enzymes for digestion. The second-best improvement in the coverage of α-helix part of proteins is the category of chromosome which is far less than the cell wall and it shows almost 5% improvement. For the categories of mitochondrion, endoplasmic reticulum, membrane, mitochondrial matrix, ribosome, vacuole, mitochondrial ribosome, integral component of membrane and cytosol, there are minor improvements in the coverage of α-helix part of proteins of those gene ontologies. Another important finding is that for both categories of membrane and integral component of membrane there is improvement in the α-helix part, however the improvement may not be as significant as the cell wall.

For the sample SDC/TBAB-70/30 the results of same analysis presented in the Figure 3-8. Results show that there is over 20% improvement in identification of α-helix part of cell wall proteins. The second-best improvement (after cell wall) in the coverage of α-helix part of proteins is the category of chromosome which is far less than the cell wall and it shows almost 5% improvement. For the categories of mitochondrion, endoplasmic reticulum, membrane, mitochondrial matrix, ribosome, vacuole, mitochondrial ribosome, integral component of membrane and cytosol, there are minor improvements in the coverage of α-helix part of proteins of those gene ontologies. Similar to SDC/TBAB-30/70 sample, for both categories of membrane and integral component of membrane there is improvement in the α-helix part, however the improvement may not be as significant as the cell wall.

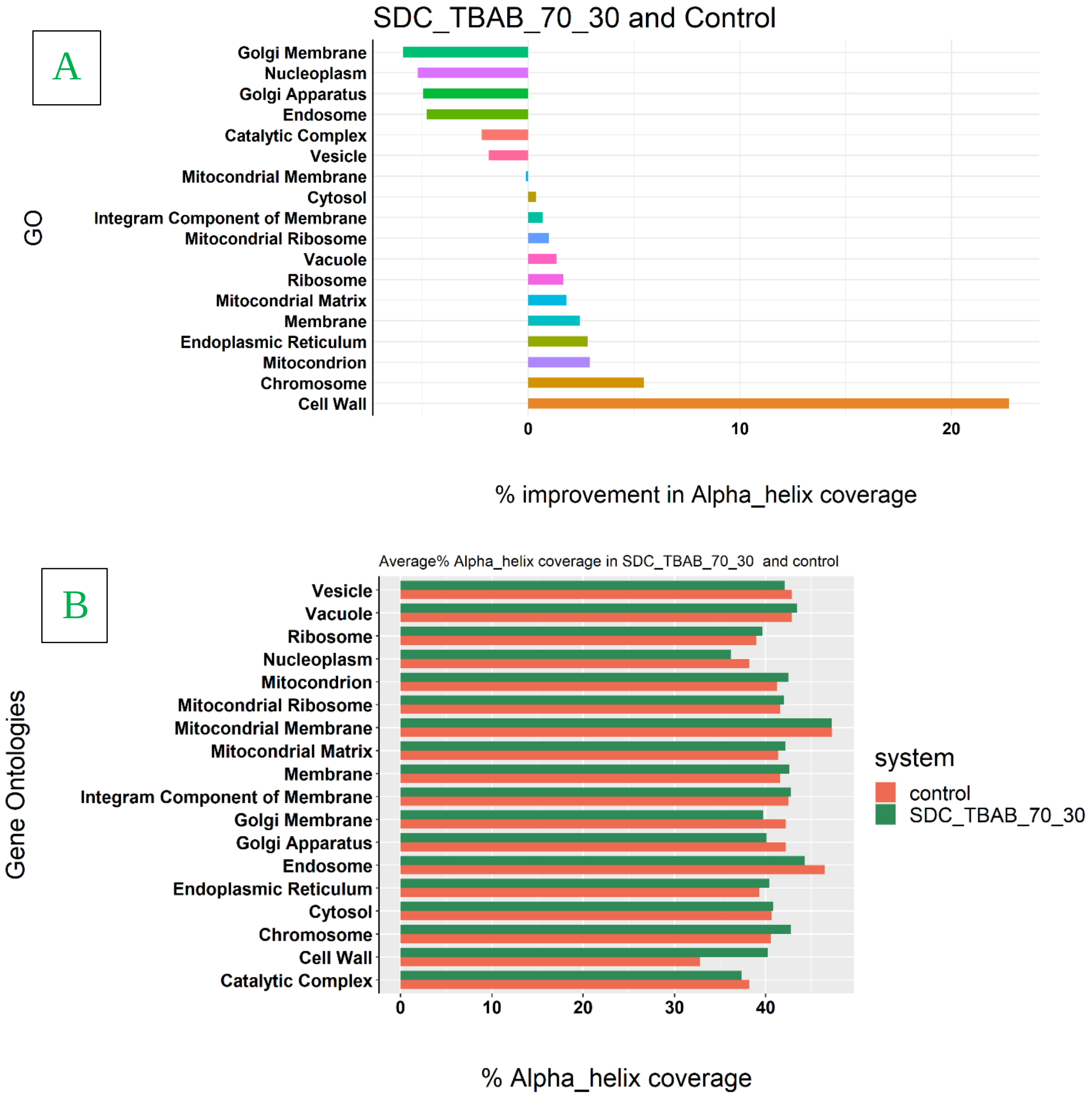


Figure 3-8. A: %improvement in α-helix coverage of main gene ontologies, a comparison between sample and the control; B: same graph but showing the absolute % α-helix coverage

For the sample SDC/TBAB-50/50 the results of same analysis presented in the Figure 3-9. Results show that there is almost 15% improvement in identification of α-helix part of cell wall proteins. The second-best improvement (after cell wall) in the coverage of α-helix part of proteins is the category of mitochondrial matrix which is far less than the cell wall and it shows over 5% improvement. For the categories of mitochondrion, membrane, mitochondrial matrix, ribosome, mitochondrial ribosome, there are minor improvements in the coverage of α-helix part of proteins of those gene ontologies. Like SDC/TBAB-30/70 and SDC/TBAB-70/30 samples, for membrane there is improvement in the α-helix part, however the improvement may not be as significant as the cell wall.

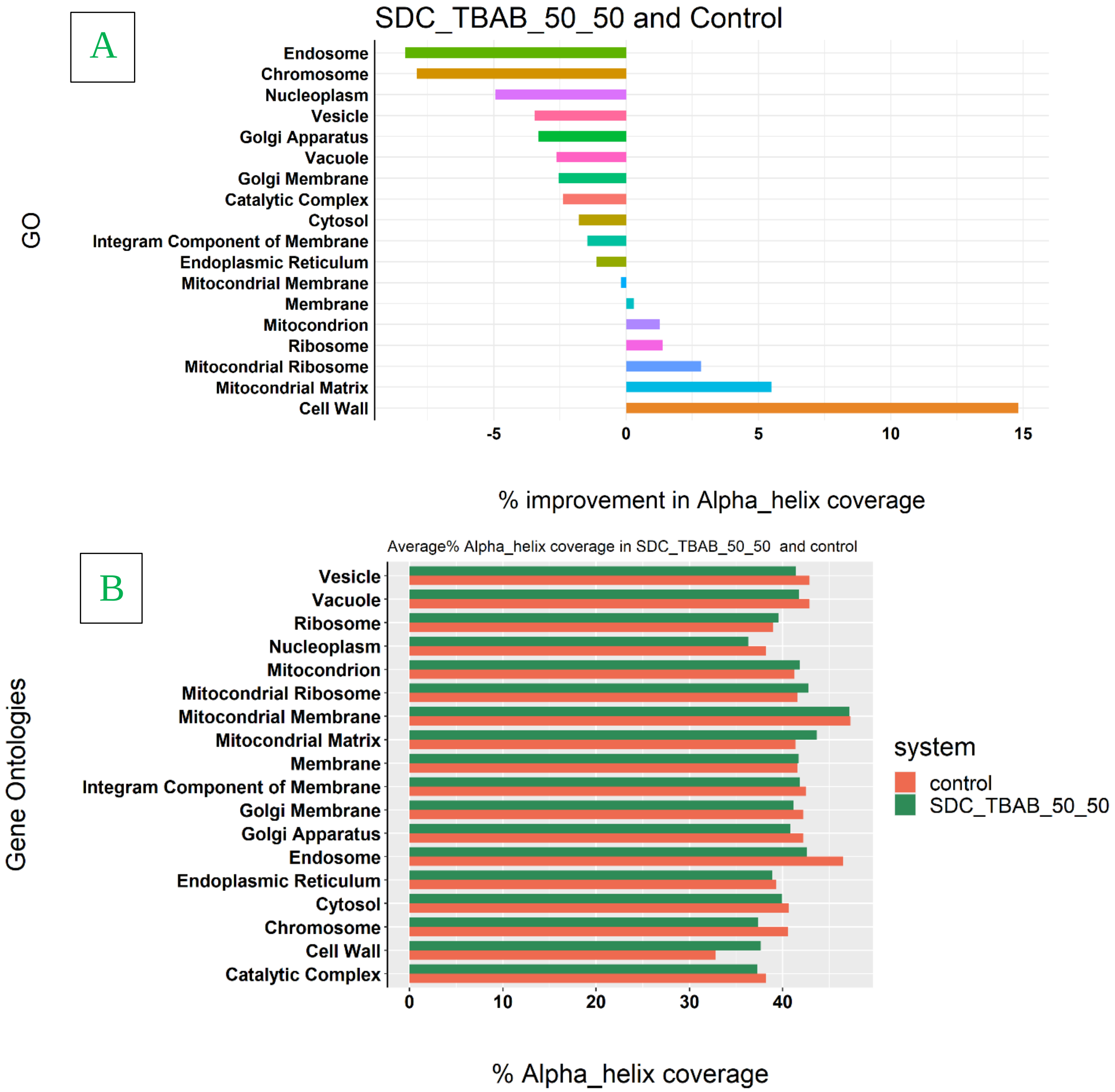


Figure 3-7. A: %improvement in α-helix coverage of main gene ontologies, a comparison between sample and the control; B: same graph but showing the absolute % α-helix coverage

3.4.3 Improvement in identification of low abundance proteins

Recently, different methods have been reported for determination of the relative proteins abundance; for example, transcriptomic analyses [19], parallel metabolic pulse labelling of genes [20], isotope clusters and stable amino acid isotope labeled peptide pairing [21] and MS techniques. The protein abundance database for Saccharomyces cerevisiae, is available online (https://yeastmine.yeastgenome.org/ [15], curated from Ghaemmaghami et al.). For each protein, different abundances are reported based on the data available in different references. In this study, we report the protein abundance as the number of molecules per cell. For each protein, the average of the abundance value from different databases was calculated and the results are reported as the database for the protein abundance in this study.

As we mentioned before, abundance of proteins in yeast varies by 4 orders of magnitude. We broke down the population of proteins into different brackets based on the abundance of those proteins; from each bracket, we compared the number of proteins that have been identified in sample and control.

For the sample SDC/TBAB-30/70, for each group of abundance, the number of identified proteins in the sample and control compared. Figure 3-10 shows the number of proteins identified in the sample and control from each group of abundance. Results show that for the abundances less than 4000 molecule per cell, there is improvement in the identification of proteins and as we get lower in abundance of proteins this improvement increases. For the bracket with lowest abundance (0-2000), there is over 40% improvement in identification of proteins.

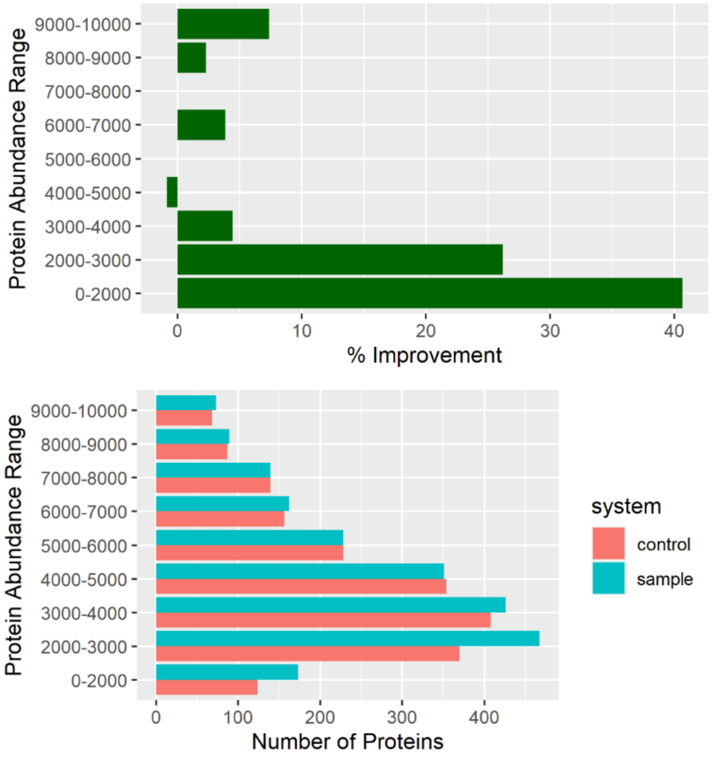


Figure 3-8. TOP: %improvement in alpha-helix coverage of main gene ontologies, a comparison between sample and the control; Bottom: same graph but showing the absolute % alpha-helix coverage

This is an important finding since low abundance proteins are the most negatively affected in the proteomics workflow and improvement in the identification these types of proteins have high importance. It is also worthy to mention that the capacity of system in identification improvement of low abundance proteins exponentially increases as the abundance of proteins gets lower than 4000 molecules per cell. For the group of abundances of 5000-6000 and 7000-8000 there is not any improvement in the number of proteins identified. Only for the category of 4000-5000 the control sample has more proteins identified compared to SDC/TBAB-30/70 sample.

Same analysis for the sample SDC/TBAB-70/30 also performed. For each group of abundance, the number of identified proteins in the sample and control compared. Figure 3-11 shows the number of proteins identified in the sample and control from each group of abundance. Results show that for the abundances less than 4000 molecule per cell, there is improvement in the identification of proteins and as we get lower in abundance of proteins this improvement increases. For the bracket with lowest abundance (0-2000), there is over 20% improvement in identification of proteins.

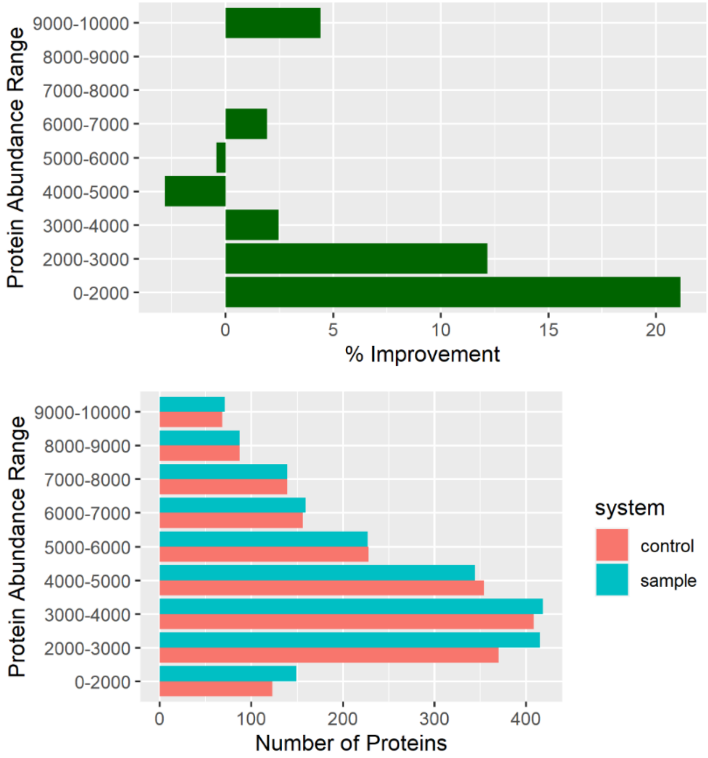


Figure 3-9. TOP: %improvement in identification of proteins from different groups of abundance, a comparison between sample and the control; Bottom: same graph but showing the absolute number of proteins identified in each abundance bracket

Like the SDC/TBAB-30/70 system that was explained earlier, the capacity of system in identification improvement of low abundance proteins exponentially increases as the abundance of proteins gets lower than 4000 molecules per cell. For the group of abundances of 7000-8000 and 8000-9000 there is not any improvement in the number of proteins identified. Only for the category of 4000-5000 the control sample has more proteins identified compared to SDC/TBAB-30/70 sample.

3.4.4 Fractionation of proteins based on pI and GRAVY

As a part of the data analysis fractionation pattern of proteins between two phases of sample SDC/TBAB-30/70 and SDC/TBAB-70/30 observed. The reason for this analysis is that in previous study that only 50 mM TBAB was used to for two-phase system, the proteins with higher pI value were extracted to the organic phase. considering the fact that most of TBAB was extracted to the coacervate phase through hydrophobic interaction with HFIP, the positive charge on the TBAB cation drived the negatively charge proteins (higher pI) to the coacervate phase. in this document since we used a mixture of positively and negatively charged surfactants, it would be interesting to know what the fractionation pattern of proteins is based on pI values.

For the sample SDC/TBAB-30/70 the concentration of SDC is 30 mM and concentration of TBAB is 70 mM. SDC and TBAB concentration measurement after forming two-phase system shows that in the sample SDC/TBAB-30/70 concentration of SDC in the coacervate phase is 365 mM and TBAB is 1125 mM. since the concentration of TBAB is almost 3 times bigger than SDC in the coacervate phase, we expect similar fractionation pattern to TBAB sample should happen. The results show that the average pI value for the proteins identified in the coacervate phase of the sample SDC/TBAB-30/70 is lower than the average pI value for the proteins identified in the aqueous phase which shows the same pattern as the TBAB sample. The same explanation for TBAB sample that mentioned earlier applies to sample SDC/TBAB-30/70 since the concentration of TBAB in the sample and coacervate is higher than SDC concentration.

For the sample SDC/TBAB-70/30 the concentration of SDC is 70 mM and concentration of TBAB is 30 mM. SDC and TBAB concentration measurement after forming two-phase system shows that in the sample SDC/TBAB-70/30 concentration of SDC in the coacervate phase is 300 mM and TBAB is 400 mM. concentration of TBAB is still higher than concentration of SDC in the coacervate phase. sice the concentration of TBAB is not as high as sample SDC/TBAB-30/70 in the coacervate phase, higher average for pI for the proteins in the coacervate phase is observed. Extraction of proteins with higher pI value to the coacervate phase increases a greater number of proteins with low pI value in the aqueous phase and this is the reason that average pI value for the proteins ion aqueous phase is lower in this sample compared to sample SDC/TBAB-30/70.

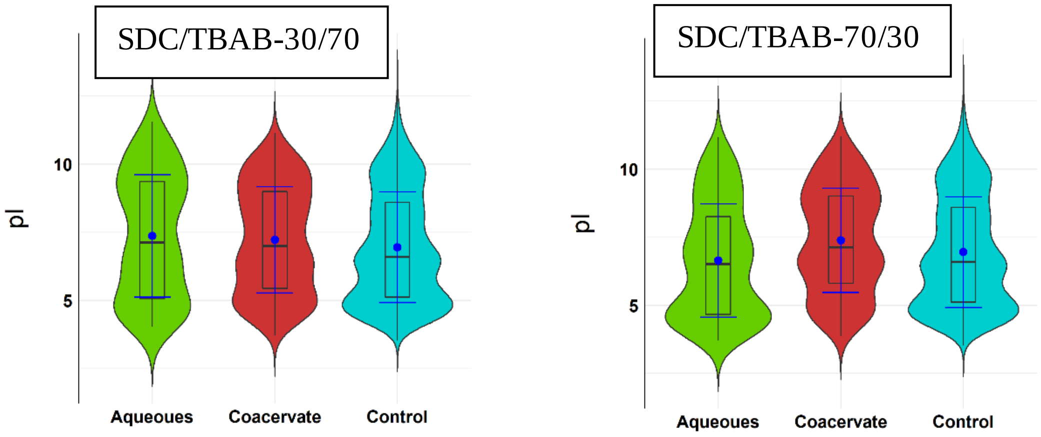


Figure 3-10. Distribution of proteins based on pI for each phase of samples compared to control

The balance between electrostatic and hydrophobic interactions determines the fractionation pattern of proteins. In this experiment demonstrated that fractionation pattern of proteins could be manipulated with the chemistry of the two-phase system. this could be a very useful approach in the studies that segregation f proteins based on the pI value is the interest.

Grand Average Hydropathy (GRAVY) is a measure of hydrophobicity of proteins. An in -depth analysis of both phases in all 3 systems performed to determine the fractionation pattern of proteins based on the GRAVY value. The surfactants and HFIP that are used in all samples can have a strong hydrophobic interaction with the proteins via their hydrophobic moiety on their structure.

However, as explained earlier, the concentration of surfactants in the aqueous phase and coacervate phase of systems changes based on the ration between concentration of surfactants that are used initially in the samples. sum of concentrations of SDC and TBAB in the aqueous phase of sample SDC/TBAB-30/70 is 1.1 mM and, in the sample, SDC/TBAB-70/30 is 29.0 mM. since the concentration of surfactants is almost 29 times bigger in the sample SDC/TBAB-70/30 compared to SDC/TBAB-30/70, the average GRAVY for the population of proteins identified only in the aqueous phase of this sample is higher compared to the sample sample SDC/TBAB-30/70. Similar explanation also applies to the coacervate phase as well.

Sum of surfactant concentrations in the coacervate phase of samples SDC/TBAB-30/70 and SDC/TBAB-70/30 are 1490 mM and 700 mM respectively. Higher concentration of surfactants in the coacervate phase of SDC/TBAB-30/70 is the reason that average GRAVY of proteins only identifies in the coacervate of this sample is higher than the corresponding number in the sample SDC/TBAB-70/30.

A comparison between the average GRAVY of the proteins that are only identified in the aqueous phase and the coacervate phase of SDC/TBAB-30/70 shows lower value for the aqueous phase compared to the coacervate phase. The reason for this, we think, is that the concentration of surfactants (SDC+TBAB) in the coacervate phase is ×1000 higher compared to the aqueous phase; this would make the coacervate more hydrophobic which would draw more hydrophobic proteins into the coacervate phase and excluding more hydrophilic proteins to the aqueous phase. Figure 3-13 is representation of this explanation.

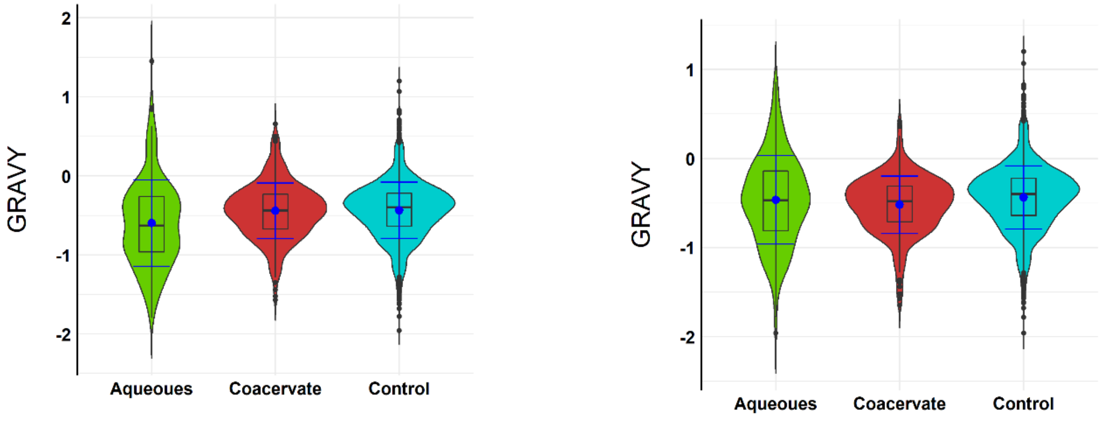


Figure 3-11. Distribution of proteins based on GRAVY for each phase of samples compared to control

3.5 CONCLUSION

In continuation of our line of research in coacervates forming form surfactants, we have tested various surfactants and the results show that each surfactant provides a certain type of protein fractionation pattern. In this study, we have examined yeast protein fractionation pattern in the biphasic system induced by presence of mixed surfactants, anionic and cationic, and we found that there is a big difference in the results; the ratio between SDC and TBAB determines the fractionation pattern of proteins based on pI. in cases that the concentration of SDC and TBAB is the same in the sample, no fractionation takes place, and all the proteins are extracted in the coacervate phase. Figure 3-14 shows that the fractionation of proteins based on pI is dependent on the ratio between SDC and TBAB in the sample.

The other fact that is noticeable in the results is, the use of SDC and TBAB, regardless of the concentration ratio between them, will improve number of proteins with α-helix structure. This is probably due to solubilizing power of the SDC and TBAB for denaturing the proteins with alpha helix structure and result in improved digestion.

Regarding the low abundance proteins, we found that all samples, except SCD/TBAB-50/50, are providing better identification for proteins with abundance less than 4000 molecules per cell. Overall, the results of SDC/TBAB-30/70 provides better improvement in all aspects of proteomics discussed in this document.

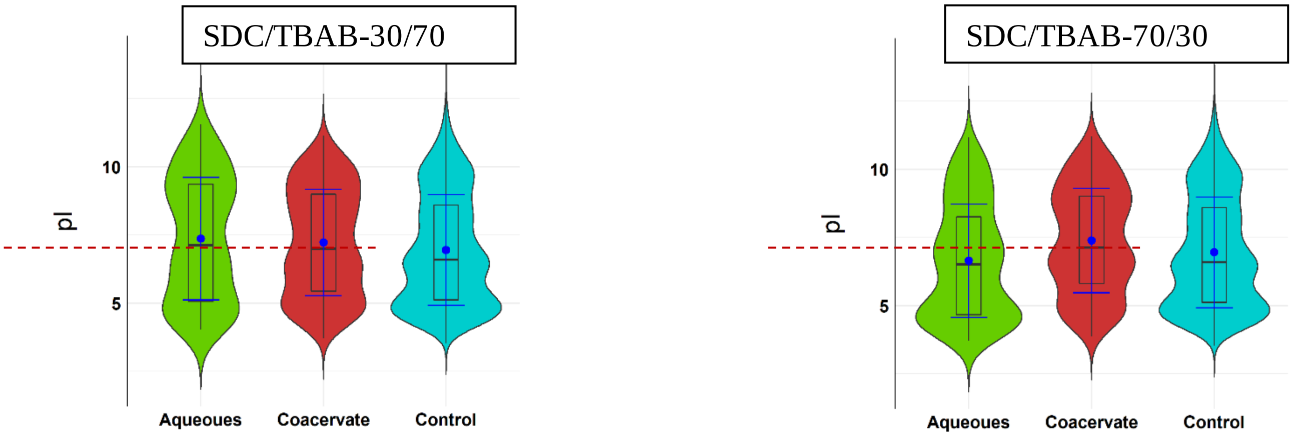


Figure 3-12. Fractionation pattern of proteins based on pI changes by changing the ration between surfactants

3.6 REFERENCES

1. Manes, N. P.; Nita-Lazar, A., Application of targeted mass spectrometry in bottom-up proteomics for systems biology research. J Proteomics **2018,** 189, 75-90.

2. Dwivedi, R. C.; Spicer, V.; Harder, M.; Antonovici, M.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Krokhin, O. V., Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. Anal Chem **2008,** 80 (18), 7036-42.

3. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R., Protein analysis by shotgun/bottom-up proteomics. Chem Rev **2013,** 113 (4), 2343-94.

4. Ly, L.; Wasinger, V. C., Protein and peptide fractionation, enrichment and depletion: tools for the complex proteome. Proteomics **2011,** 11 (4), 513-34.

5. Tao, D.; Qiao, X.; Sun, L.; Hou, C.; Gao, L.; Zhang, L.; Shan, Y.; Liang, Z.; Zhang, Y., Development of a highly efficient 2-D system with a serially coupled long column and its application in identification of rat brain integral membrane proteins with ionic liquids-assisted solubilization and digestion. J Proteome Res **2011,** 10 (2), 732-8.

6. Zhu, M. Z.; Li, N.; Wang, Y. T.; Liu, N.; Guo, M. Q.; Sun, B. Q.; Zhou, H.; Liu, L.; Wu, J. L., Acid/Salt/pH Gradient Improved Resolution and Sensitivity in Proteomics Study Using 2D SCX-RP LC-MS. J Proteome Res **2017,** 16 (9), 3470-3475.

7. Yu, P.; Petzoldt, S.; Wilhelm, M.; Zolg, D. P.; Zheng, R.; Sun, X.; Liu, X.; Schneider, G.; Huhmer, A.; Kuster, B., Trimodal Mixed Mode Chromatography That Enables Efficient Offline Two-Dimensional Peptide Fractionation for Proteome Analysis. Anal Chem **2017,** 89 (17), 8884-8891.

8. Futcher, B.; Latter, G. I.; Monardo, P.; McLaughlin, C. S.; Garrels, J. I., A sampling of the yeast proteome. Mol Cell Biol **1999,** 19 (11), 7357-68.

9. Baggerman, G.; Vierstraete, E.; De Loof, A.; Schoofs, L., Gel-based versus gel-free proteomics: a review. Comb Chem High Throughput Screen **2005,** 8 (8), 669-77.

10. Qian, W. J.; Jacobs, J. M.; Liu, T.; Camp, D. G.; Smith, R. D., Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications. Mol Cell Proteomics **2006,** 5 (10), 1727-44.

11. Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; Apweiler, R.; Haab, B. B.; Simpson, R. J.; Eddes, J. S.; Kapp, E. A.; Moritz, R. L.; Chan, D. W.; Rai, A. J.; Admon, A.; Aebersold, R.; Eng, J.; Hancock, W. S.; Hefta, S. A.; Meyer, H.; Paik, Y. K.; Yoo, J. S.; Ping, P.; Pounds, J.; Adkins, J.; Qian, X.; Wang, R.; Wasinger, V.; Wu, C. Y.; Zhao, X.; Zeng, R.; Archakov, A.; Tsugita, A.; Beer, I.; Pandey, A.; Pisano, M.; Andrews, P.; Tammen, H.; Speicher, D. W.; Hanash, S. M., Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics **2005,** 5 (13), 3226-45.

12. Barnea, E.; Sorkin, R.; Ziv, T.; Beer, I.; Admon, A., Evaluation of prefractionation methods as a preparatory step for multidimensional based chromatography of serum proteins. Proteomics **2005,** 5 (13), 3367-75.

13. Kim, J. Y.; Lee, J. H.; Park, G. W.; Cho, K.; Kwon, K. H.; Park, Y. M.; Cho, S. Y.; Paik, Y. K.; Yoo, J. S., Utility of electrophoretically derived protein mass estimates as additional constraints in proteome analysis of human serum based on MS/MS analysis. Proteomics **2005,** 5 (13), 3376-85.

14. Pavlou, M. P.; Kulasingam, V.; Sauter, E. R.; Kliethermes, B.; Diamandis, E. P., Nipple aspirate fluid proteome of healthy females and patients with breast cancer. Clin Chem **2010,** 56 (5), 848-55.

15. Fuller, B. F.; Ottens, A. K., Separation of the neuroproteome by ion exchange chromatography. Methods Mol Biol **2009,** 566, 193-200.

16. Rocchiccioli, S.; Citti, L.; Boccardi, C.; Ucciferri, N.; Tedeschi, L.; Lande, C.; Trivella, M. G.; Cecchettini, A., A gel-free approach in vascular smooth muscle cell proteome: perspectives for a better insight into activation. Proteome Sci **2010,** 8, 15.

17. Liu, H.; Lin, D.; Yates, J. R., Multidimensional separations for protein/peptide analysis in the post-genomic era. Biotechniques **2002,** 32 (4), 898, 900, 902 passim.

18. Simpson, D. C.; Ahn, S.; Pasa-Tolic, L.; Bogdanov, B.; Mottaz, H. M.; Vilkov, A. N.; Anderson, G. A.; Lipton, M. S.; Smith, R. D., Using size exclusion chromatography-RPLC and RPLC-CIEF as two-dimensional separation strategies for protein profiling. Electrophoresis **2006,** 27 (13), 2722-33.

19. Moore, A. W.; Jorgenson, J. W., Comprehensive three-dimensional separation of peptides using size exclusion chromatography/reversed phase liquid chromatography/optically gated capillary zone electrophoresis. Anal Chem **1995,** 67 (19), 3456-63.

20. Khaledi, M. G.; Jenkins, S. I.; Liang, S., Perfluorinated alcohols and acids induce coacervation in aqueous solutions of amphiphiles. Langmuir **2013,** 29 (8), 2458-64.

21. Nejati, M. M.; Khaledi, M. G., Perfluoro-alcohol-induced complex coacervates of polyelectrolyte-surfactant mixtures: phase behavior and analysis. Langmuir **2015,** 31 (20), 5580-9.

22. McCord, J. P.; Muddiman, D. C.; Khaledi, M. G., Perfluorinated alcohol induced coacervates as extraction media for proteomic analysis. J Chromatogr A **2017,** 1523, 293-299.

23. Koolivand, A.; Clayton, S.; Rion, H.; Oloumi, A.; O'Brien, A.; Khaledi, M. G., Fluoroalcohol - Induced coacervates for selective enrichment and extraction of hydrophobic proteins. J Chromatogr B Analyt Technol Biomed Life Sci **2018,** 1083, 180-188.

24. Koolivand, A.; Azizi, M.; O'Brien, A.; Khaledi, M. G., Coacervation of Lipid Bilayer in Natural Cell Membranes for Extraction, Fractionation, and Enrichment of Proteins in Proteomics Studies. J Proteome Res **2019,** 18 (4), 1595-1606.

25. Khanal, D. D.; Tasharofi, S.; Azizi, M.; Khaledi, M. G., Improved Protein Coverage in Bottom-Up Proteomes Analysis Using Fluoroalcohol-Mediated Supramolecular Biphasic Systems With Mixed Amphiphiles for Sample Extraction, Fractionation, and Enrichment. Anal Chem **2021,** 93 (20), 7430-7438.

26. Azizi, M.; Tasharofi, S.; Koolivand, A.; Oloumi, A.; Rion, H.; Khaledi, M. G., Improving identification of low abundance and hydrophobic proteins using fluoroalcohol mediated supramolecular biphasic systems with quaternary ammonium salts. J Chromatogr A **2021,** 1655, 462483.

27. Klausen, M. S.; Jespersen, M. C.; Nielsen, H.; Jensen, K. K.; Jurtz, V. I.; Sønderby, C. K.; Sommer, M. O. A.; Winther, O.; Nielsen, M.; Petersen, B.; Marcatili, P., NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. Proteins **2019,** 87 (6), 520-527.

28. Robinson, S. W.; Afzal, A. M.; Leader, D. P., Handbook of Pharmacogenetics and Stratified Medicine. Academic Press: 2014; pp 259-287.

29. Boyle, A. L., Applications of de novo designed peptides. In Peptide Applications in Biomedicine, Biotechnology and Bioengineering, Koutsopoulos, S., Ed. Woodhead Publishing: 2018; pp Pages 51-86.

30. Banerjee, J.; Radvar, E.; Azevedo, H. S., Self-assembling peptides and their application in tissue engineering and regenerative medicine. In Peptides and Proteins as Biomaterials for Tissue Regeneration and Repair, Woodhead Publishing: 2018; pp Pages 245-281.

31. Clark, D. P.; Pazdernik, N. J.; McGehee, M. R., Molecular biology. Third / David P. Clark, Nanette J. Pazdernik, Michelle R. McGehee. ed.; Academic Cell: Amsterdam, 2018.

32. Klausen, M. S.; Jespersen, M. C.; Nielsen, H.; Jensen, K. K.; Jurtz, V. I.; Sønderby, C. K.; Sommer, M. O. A.; Winther, O.; Nielsen, M.; Petersen, B.; et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins* **2019**, *87* (6), 520-527. DOI: 10.1002/prot.25674.

33. Xie X, Lipke PN. On the evolution of fungal and yeast cell walls. *Yeast.* **2010**;27(8):479-488. doi:10.1002/yea.1787

CHAPTER 4

YEAST PROTEOME SECONDARY STRUCTURE CALCULATOR

4.1 Primary and Secondary Structures of Proteins

The chain of amino acids that form the backbone of proteins is the primary structure of the proteins. Secondary structures of a proteins are a spatial conformation of the primary structures. These structures are usually formed because of hydrogen bonds between amide groups of amino acids and carbonyl groups of other amino acids in the primary structure. Among secondary structures, α-helices and β-sheets are the dominant structures. Linus Pauling et al. in 1951 proposed α-helical structure which is demonstrated in Figure 4-1. In this structure the group of hydrogen bonds are placed in the center of the helical structure and the side chains of the amino acids in the proteins backbone, are pointing outward[1].

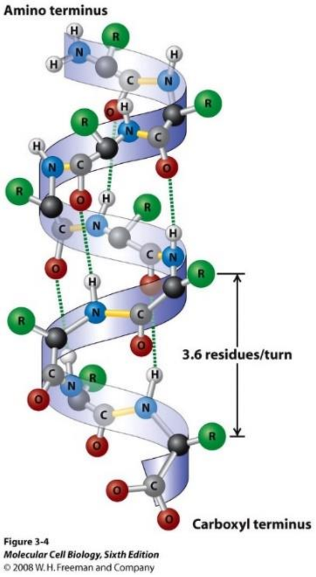


Figure 3-4
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

Figure 4-1. Typical B-helix structure formed form primary structure of a protein

After α-helices, β-structures are the most common secondary structures of the proteins and it consist of β-strands and β-sheets. β -structures are formed as a result of hydrogen bonding between amino and carbonyl groups of amino acids as well[1].

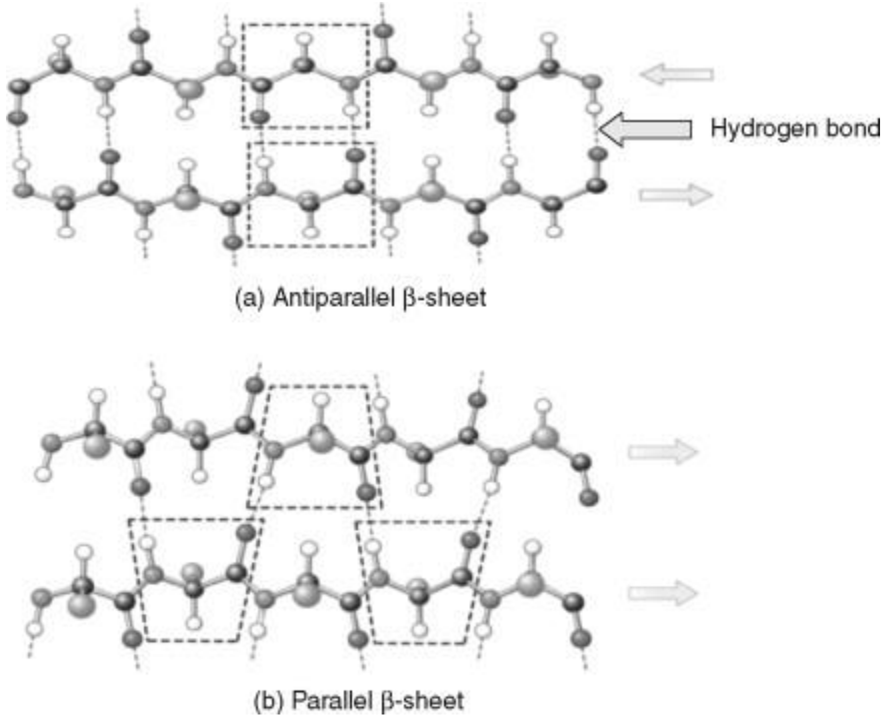


Figure 4- 2. Hydrogen bonding that forms parallel and anti-parallel β-sheet[1]

4.2 Experimental Determination of Secondary Structures of Proteins

Experimental determination of secondary structures is not trivial. There are two methods for determination of protein structures: first, using data from Nuclear Magnetic Resonance (NMR); second: X-ray diffraction of the crystals that are formed from the proteins. Both methods are difficult and time consuming. Therefore, the unknown secondary structure of proteins is often predicted using the data form the structural information of proteins that are known and determined using experimental methods.

Although Circular Dichroism (CD) Spectroscopy, provides information about the ratio between α-helices and β-sheets, it does not provide direct information about the secondary structure of the protein[2].

4.3 Importance of Prediction of Secondary Structures of Proteins

Secondary structure of the proteins can be used to predict the tertiary and quaternary structure (3D) since predicting tertiary and quaternary structure solely using primary sequence may not be sufficient[3]. Information about 3D structure of proteins is necessary for the purpose of drug discovery and for treatments that targets proteins [4,5,6].

Since determination of secondary structure of proteins using experimental methods Is extremely limited and difficult, the ability to predict secondary structures of proteins from the primary sequence is particularly important.

Anfinsen experiment demonstrated that secondary structural characteristic of a proteins could be directly determined from its primary sequence[7]. Many methods have been developed for prediction of secondary structure which had a great positive impact on biology and chemistry[8,9,10,11,12]. However, the accuracy of these methods have been under question[13]. Experimental methods cannot keep up with the increase in demand for predictions of structural features; that would only leave the option of improving the prediction tools[14]. NetSurfP-1.0 is a prediction tool for secondary structures using neural network[15]. NetSurfP-2.0 is an extension of NetSurfP-1.0 which utilized deep neural network to predict secondary structures with the accuracy of 85%. In addition to accuracy, this tool presents reduced computational time compared to other methods[16].

NetSurfP-2.0 is designed to be user friendly and efficient in calculation time of large number of sequences. In addition to that the output of the calculation is available in many formats that would make further data analysis even easier. This tool is available as a web-sever (http://www.cbs.dtu.dk/services/NetSurfP-2.0/) which can accept up to 4000 sequences at a time. NetSurfP-2.0 is an amazing tool for the prediction of secondary structures from the primary structure. However, this tool like other prediction tool come with its own drawbacks.

NetSurP-2.0. is not designed to process large number of peptides and provide a coherent picture of secondary structure of proteins corresponding to those peptides. A tool developed as an extension to NetSurP-2.0 that can process large number of peptides form bottom-up proteomics and calculates the percent coverage of secondary structures of the proteins involved in the study; this tool is named yeast proteome secondary structure calculator (ypssc) In collaboration with developers of NetSurP-2.0, we analyzed whole yeast proteome in NetSurP-2.0 and created a database for secondary structures of yeast proteins. The tool we developed, ypssc, uses this database to find secondary structures of peptides and proteins identified in the sample. Without ypssc, the task of secondary structure identification form bottom-up proteomics would be almost impossible, however, ypssc enables the user to process a large bottom-up proteomics data (like the proteomics studies that mentioned earlier) in less than 1 min in a regular desktop computer. ypssc is written in R language, and it is part of Comprehensive R Archive Network (CRAN) as a package which enables the user to use this advanced tool with no knowledge of programming. ypssc, on one hand benefits forms the accuracy of NetSurfP-2.0 to calculate secondary structure and on the other hand addresses the issue of analyzing so many peptides with NetSurfP-2.0 by eliminating the need for direct analysis of the peptides from bottom-up proteomics.

Instead of direct analysis of peptides by NetSurfP-2.0 which raises the problem of combining the results of peptides to proteins, the whole yeast proteome has been analyzed once by NetSurfP-2.0 and kept as Secondary Structure Database for Yeast Proteome (SSDYP). Then the peptides form the experiment are matched and compared to this database to extract secondary structure of the peptides.

The SSDYP contains structural information for all amino acids of whole yeast proteome (over 3000,000 amino acids) which contains over 6700 proteins. For a hypothetical protein, the SSDYP contains the ID of the protein, amino acids with numbers and structural information for each amino acid. Figure 4-3 shows a hypothetical protein in the SSDYP.

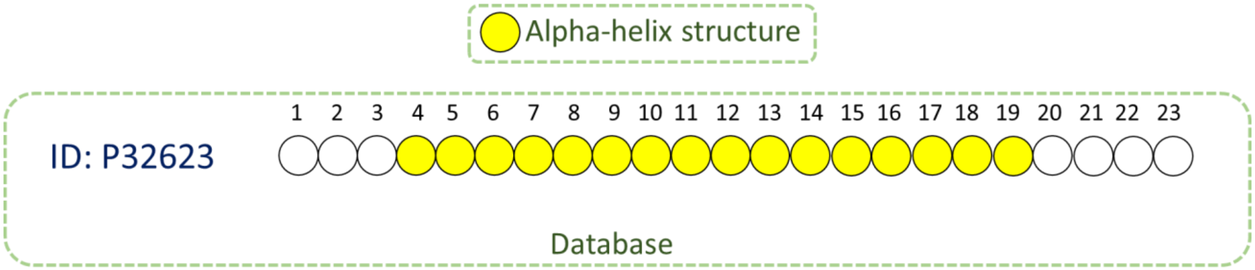


Figure 4-3. A protein in the SSDYP is represented by the ID of the protein, amino acids (circles) with numbers and structural information about the amino acids (yellow color which represents α-helical structure in this case).

Focusing on the hypothetical protein, in the real sample, there are many peptides identified from the hypothetical protein. Some parts of protein maybe covered many times, and some parts could be lost and not identified. Figure 4-3 shows a typical situation that peptides of a protein identified in the sample. ypssc first finds all the peptides that belongs to the hypothetical protein and arrange them based on the numbers of the amino acids; then it removes the parts of the protein that have been identified more than once in multiple peptides and collapses the population of identified peptides in the sample into one sequence that represents the coverage of the hypothetical protein. The result would show that which part of the protein is identified, and which part is missing.

Then, ypssc matches the the sequence that identified in the sample with SSDYP to find the structural information about amino acids. In this example we are looking at α-helix structures. Figure 4-5 shows this process. In Figure 4-6, the amino acids from the sample that have been matched with the α-helix part of the SSDYP are colored green. The ration between number of α-helix amino acids in the sample to SSDYP, determines the coverage of α-helix in the sample.

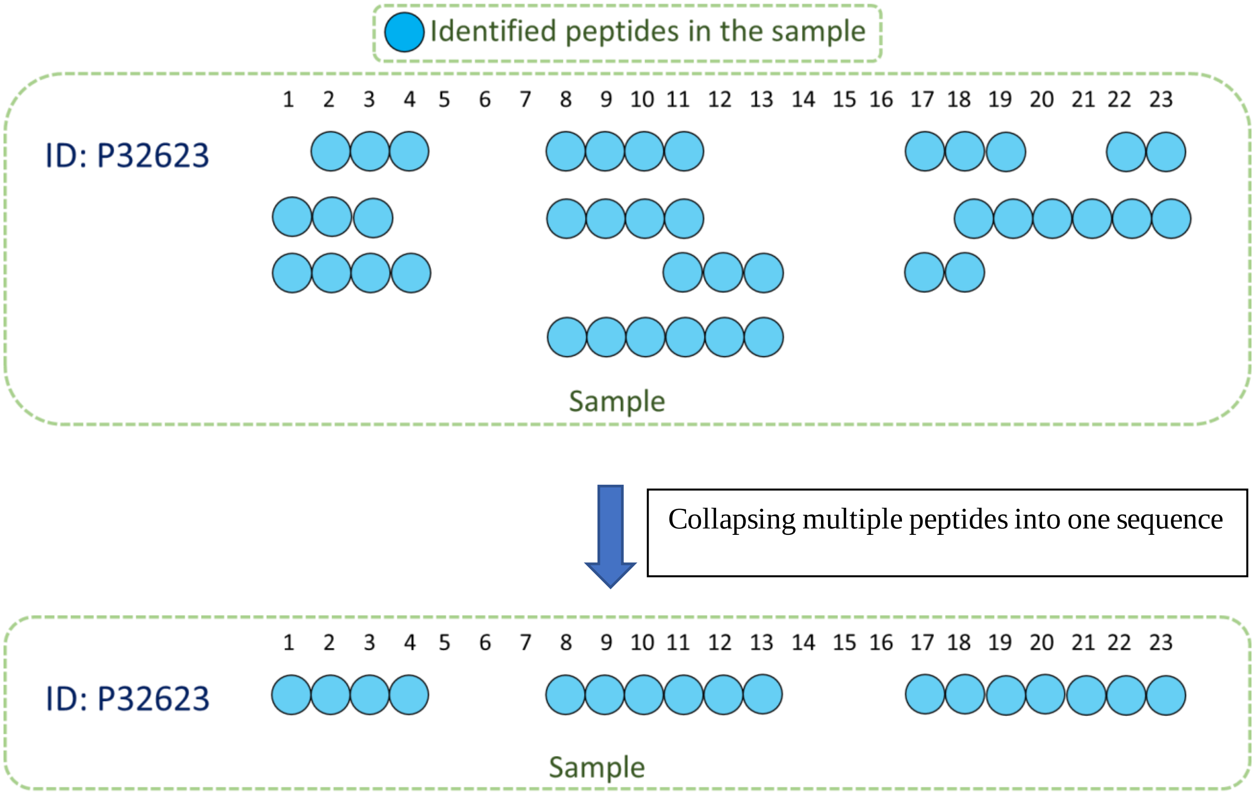


Figure 4-4. ypssc finds all the peptides that belongs to the hypothetical protein and arrange them then it collapses the population of identified peptides in the sample to one sequence that represents the coverage of the hypothetical protein.

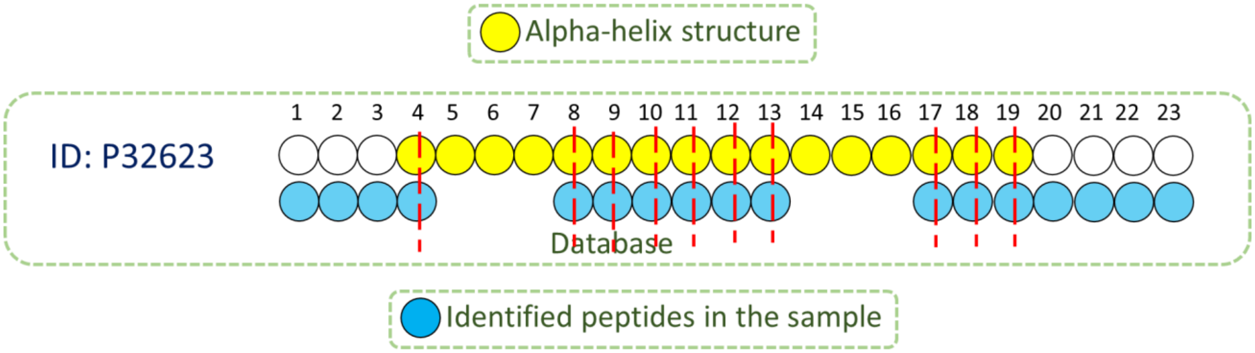


Figure 4-5. matching the sequences that found in the sample and database to find the structural information about the amino acids

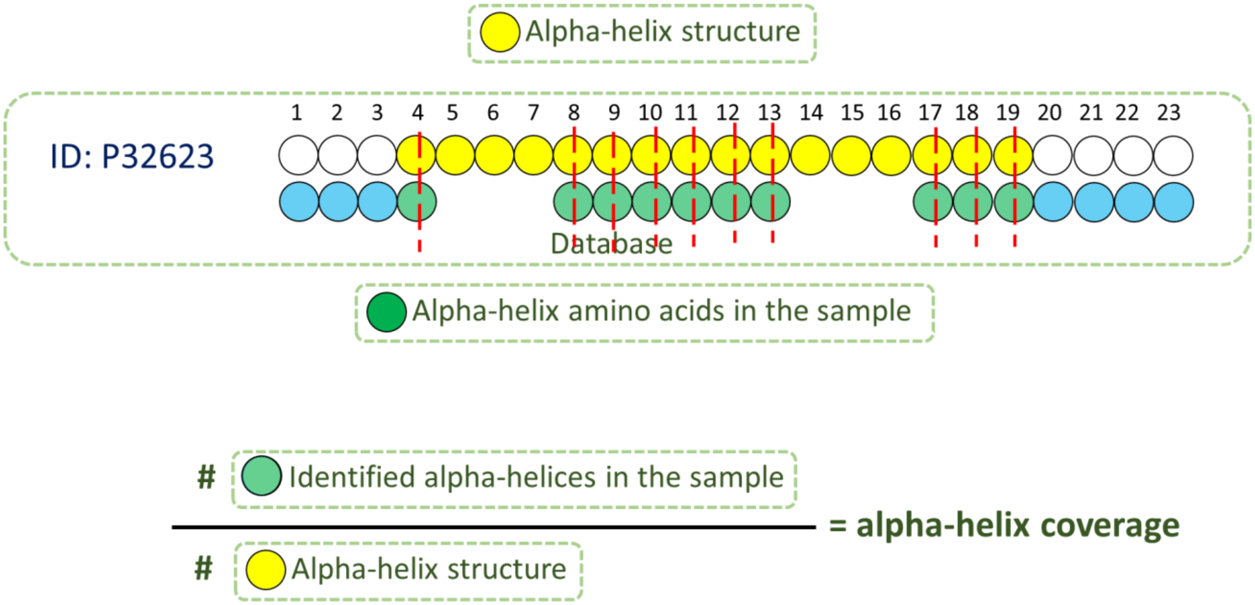


Figure 4-6. Final calculation by ypssc. In this step the program calculated the coverage of α-helix structures.

The same method has been used to determine β-sheets. Part of sequence that is not β-sheet or α-helix is primary structure which in this document mentioned as “chain structure”.

Another advantage of this methods is the ability to analyze many samples in a very short time (less than 5 min in a regular desktop computer) instead of running them repeatedly in NetSurfP-2.0 which is a complicated algorithm, and it will take very long time to do the calculations.

4.4 Input of the ypssc

MaxQuant is a quantitative proteomics software designed to analyze large mass-spectrometric data. The input of MaxQuant is a raw file (.raw) from high-resolution mass spectrometers. After analysis of the raw file in MaxQuant, the program generates a folder named “combined”. In this folder there is another folder named “txt” which contains many files with text format (.txt). One of the files called “peptides” which is the input of the ypssc to calculate secondary structures. ypssc has been designed such a way that can analyzed and extract information regarding the sample regardless of the name that user chosen for the sample.

4.5 Output of the ypssc

The output of the program is a csv file (.csv) that contains 5 columns, and the number of rows depends on the number of proteins in the sample. First column contains the ID of the identified α-helix proteins in the sample, second column contains the number of identified amino acids from the corresponding protein, third column contains number of identified amino acids with secondary structure, fourth column contains the number of amino acids that the protein originally has in the SSDYP, and fifth column contains the number of amino acids with secondary structure that the protein originally has in the SSDYP. These columns should provide all information that the user needs to know about the protein and its structural information as well as structural information about the parts of the protein that has been identified in the sample.

In addition to ypssc, we developed a program written to perform specific data analysis that we are interested in our group. This program is in Appendix 4-1

4.6 The Code

```
##################################### auxil functions

# >>

#' @title readFileInput

#' @param pathFileInput input file path from which bla bla bla xxxx xxxx xxxx xxxx xxxx xxxx
xxxx.

# <<

##################################### readFileInput()

# >>

readFileInput <- function( pathFileInput ) {

    # Reading csv file >>

    df = read.csv( pathFileInput )

    # Removing the columns that are not needed and finding the columns containing sample
information >>

    df = df[ , -which( names(df) %in% c(    "Sequence","N.term.cleavage.window",

                                            "C.term.cleavage.window","Amino.acid.before",

                                            "First.amino.acid","Second.amino.acid",

                                            "Second.last.amino.acid","Last.amino.acid",

                                            "Amino.acid.after","A.Count","R.Count","N.Count",

                                            "D.Count","C.Count","Q.Count","E.Count",

                                            "G.Count","H.Count","I.Count","L.Count",

                                            "K.Count","M.Count","F.Count","P.Count",

                                            "S.Count","T.Count","W.Count","Y.Count",

                                            "V.Count","U.Count","O.Count","Length",

                                            "Missed.cleavages","Mass",

                                            "Leading.razor.protein","Gene.names",

                                            "Protein.names","Unique..Groups.",

                                            "Unique..Proteins.","Charges","PEP",

                                            "Score","Experiment.ST168.THF.A",

                                            "Experiment.ST168.THF.O",

                                            "Experiment.ST169.DMSO.A","Experiment.ST169.DMSO.O",

                                            "Experiment.ST170.But.A","Experiment.ST170.But.O",

                                            "id","Protein.group.IDs","Mod..peptide.IDs",
                                            "Evidence.IDs","MS.MS.IDs","Best.MS.MS",
                                            "Oxidation..M..site.IDs","Taxonomy.IDs",
                                            "MS.MS.Count" ) ) ]
    names = names(df)
    sampleNames         = names[ grepl("Intensity.", names) ]
    sampleNamesUpdate   = gsub( '\\.|Intensity.', ' ', sampleNames )
    names_list          = vector()
    i  = 1
    pb = winProgressBar( title = "progress bar",
                         min   = 0,
                         max   = length(sampleNames),
                         width = 300 )
    for ( i in 1 : length(sampleNames) ) {
        temp       = paste(sampleNamesUpdate[i],' \n \n ')
        temp
        names_list = paste( names_list, temp )
        # Sys.sleep(0.9)
        Sys.sleep(0.1)
        setWinProgressBar( pb, i, title = paste( sampleNamesUpdate[i], '    ',
round(i/length(sampleNames)*100, 0), "% done") )
    }
    close(pb)


    # Conformation about sample names from user >>
    sampleNameConfirmation = dlgInput(paste("Identified sample names in the uploaded file:\n \n
\n", names_list,
                                            "\nIf it is correct, please enter 'Yes'"))$res
    class(sampleNameConfirmation)
    if ( sampleNameConfirmation == "yes" ) {
        tkmessageBox( title   = "Message",
                      message = "Your analysis in in progress",
                      icon    = "info",
                      type    = "ok" )
        as.character(names(df))
    } else if( sampleNameConfirmation=="no") {

        tkmessageBox( title   = "Message",

                      message = "Please put a 'samples.CSV' file containing just sample names in
one column",

                      icon    = "info",

                      type    = "ok" )

        Sample_names = read.csv('samples.csv')

    }

    # Returning multiple variables as a R-list >>

    dataFileInput                   = list()

    dataFileInput$df                = df

    dataFileInput$sampleNames       = sampleNames

    dataFileInput$sampleNamesUpdate = sampleNamesUpdate

    return( dataFileInput )

}

# <<

##################################### readFileInput()
##############################################################################

##################################### creatOutputDir()
#############################################################################

# >>

creatOutputDir <- function( pathDirOutput ) {



    dateTimeCurrent = format( Sys.time(), "%Y%m%d_%H%M%S" )        # << get current date and time

    nameDirOutput   = paste0( "results_AHC_", dateTimeCurrent )     # << name of the output folder

    pathDirOutput   = paste0( pathDirOutput, "/", nameDirOutput )  # << path of the output folder

    dir.create( pathDirOutput )                                    # creating new folder for
output files

    setwd( pathDirOutput )              # << setting working dir to "pathDirOutput" to write
output files



    return( dateTimeCurrent )

}

##################################### creatOutputDir()

##################################### removeRows()

# >>
```

```
removeRows <- function( df, dateTimeCurrent ) {

    removeDoubious = dlgInput( paste0("Do you want to remove the rows containing doubious 
proteins?\n",

                                      "Rows that have 2 or more protiens assigned to one 
identified peptide are called doubious\n",

                                      "Answer with yes or no") )$res

    df = filter( df, !grepl( ';', df$Proteins) )

    write.csv( df, paste0( dateTimeCurrent, " ", 'df.csv' ), row.names = FALSE)

    removeReverse  = dlgInput( paste0("Do you want to remove rows that contains peptides that 
matched to decoy that has reverse ",

                                      "sequnce of real protein?\n",

                                      "Theses proteins are usually removed.\n",

                                      "Answer with yes or no") )$res

    df = filter( df, !grepl( '\\+', df$Reverse) )

    removeReverse  = dlgInput( paste0("Do you want to remove rows that contains peptides that are 
showing signs of contamination?\n",

                                      "Theses proteins are usually removed.\n",

                                      "Answer with yes or no") )$res

    df = filter( df, !grepl( '\\+', df$Potential.contaminant) )

    removeReverse  = dlgInput( paste("Do you want to remove rows that contains peptides that are 
not showing any intensity?\n",

                                     "Theses proteins are usually removed.\n",

                                     "Answer with yes or no") )$res

    df = filter( df, df$Intensity > 0 )

    return( df )

}

##################################### removeRows()

##################################### auxil functions

##################################### αHelixCalculator()

αHelixCalculator = function( pathFileInput = "C:/Users/Desktop/peptides_second rep.csv",

                                 pathDirOutput = "C:/Users/Downloads" ) {

    print("Started")

    startTime = Sys.time()

    # Getting current working directory

    originalWorkingDir = getwd()

    # Checking if 'pathDirOutput' is provided
    if ( is.null( pathDirOutput ) ) {
        pathDirOutput = getwd()
    }
    # Reading the input sample file
    dataFileInput     = readFileInput( pathFileInput )
    df                = dataFileInput$df
    sampleNames       = dataFileInput$sampleNames
    sampleNamesUpdate = dataFileInput$sampleNamesUpdate
    # Create output folder
    dateTimeCurrent = creatOutputDir( pathDirOutput )
    # Removing the rows that are not needed
    df = removeRows( df, dateTimeCurrent )
    # Writing `dataBase_numOfAA`
    write.csv( dataBase_numOfAA,
               "dataBase_numOfAA.csv",
               row.names = FALSE )
    # A helix calculation for dataBase
    αHelixCalculation( df, sampleNames, sampleNamesUpdate, dateTimeCurrent )
    endTime   = Sys.time()
    timeTaken = endTime - startTime
    print( paste0( "Time taken for the AHC run: ", format(timeTaken) ) )
    setwd( originalWorkingDir )
    return( invisible(NULL) )
}
##################################### αHelixCalculation()
αHelixCalculation <- function( df, sampleNames, sampleNamesUpdate, dateTimeCurrent ) {
    dataBase_α   = select(dataBase_α, c(1,2))
    dataBase_reduced = dataBase_α
    num_Pro_aaa    = unique(dataBase_reduced$id)
    protein        = vector()
    num_aaa_pro_DB = vector()
    pb_1 = winProgressBar( title = "progress bar",
                           min   = 0,
                           max   = length(num_Pro_aaa),
                           width = 300 )
    i = 1
    for( i in 1 : length(num_Pro_aaa) ) {
        item                = num_Pro_aaa[i]
        proteins            = filter(dataBase_reduced, id == item)
        num_aaa_pro_DB_temp = length(proteins$id)
        num_aaa_pro_DB      = c(num_aaa_pro_DB_temp,num_aaa_pro_DB)
        protein             = c(unique(proteins$id),protein)
        proteins            = vector()
        num_aaa_pro_DB_temp = vector()
        setWinProgressBar( pb_1, i,
                           title = paste( 'A-helix calculation for database      ',
                                          round(i/length(num_Pro_aaa)*100, 0),
                                          "% done") )
    }
    close(pb_1)
    # Calculating the number of amino acids for α ####
    aaa              = data.frame( id      = protein,
                                   num_aaa = num_aaa_pro_DB )
    cal_for_database = left_join(  dataBase_numOfAA,
                                   aaa,
                                   by = 'id' )
    Sys.sleep(0.5)
    # Samples ####
    i = 1
    for( i in 1 : length(sampleNames) ) {
        temp = which( names(df) == sampleNames[i] )
        # Peptides in the sample >>
        sample_peptides = filter( df, df[,temp] > 0 )
        write.csv( sample_peptides,
                   paste0( dateTimeCurrent,
                           " ", 'List of peptides in',
                       sampleNamesUpdate[i], '.csv' ),
               row.names = FALSE )
    sample = paste( as.character(sampleNamesUpdate[i]), '_ peptides' )
    assign( sample, sample_peptides )
    # Proteins in the sample >>
    sample_proteins = unique( sample_peptides$Proteins )
    write.csv( sample_proteins,
               paste0( dateTimeCurrent,
                       " ", 'List of proteins in',
                       sampleNamesUpdate[i], '.csv' ),
               row.names = FALSE )
    sample = paste( as.character(sampleNamesUpdate[i]), '_ proteins' )
    assign( sample, sample_proteins )
    # Calculating α helix coverage for samples >>
    proteins_in_s = vector()
    aa_in_s       = vector()
    aaa_in_s      = vector()
    pb_2 = winProgressBar( title = "progress bar",
                           min   = 0,
                           max   = length(sample_proteins),
                           width = 300 )
    j = 1
    for( j in 1 : length(sample_proteins) ) {
        item      = sample_proteins[j]
        Pro_chunk = filter( sample_peptides, sample_peptides$Proteins == item )
        k         = 1
        list_aa_s = vector()
        for( k in 1 : length(Pro_chunk$Proteins) ) {

            start          = Pro_chunk$Start.position[k]
            end            = Pro_chunk$End.position[k]
            list_aa_s_temp = seq(start:end)
            list_aa_s_temp = list_aa_s_temp+start-1
            list_aa_s      = c( list_aa_s_temp, list_aa_s )
            list_aa_s_temp = vector()
        }
        proteins_temp = item
        proteins_in_s = c( proteins_temp, proteins_in_s )
        proteins_temp = vector()
        aa_in_s_temp  = length( unique(list_aa_s) )
        aa_in_s       = c( aa_in_s_temp, aa_in_s )
        aa_in_s_temp  = vector()
        protein_chunk_dataBase = filter( dataBase_reduced, id == item )
        aaa_in_s_temp = unique(list_aa_s)%in%protein_chunk_dataBase$n
        aaa_in_s_temp = sum(aaa_in_s_temp)
        aaa_in_s      = c( aaa_in_s_temp, aaa_in_s )
        aaa_in_s_temp = vector()
        results = data.frame( id                              = proteins_in_s,
                              num_amino_acids_in_sample       = aa_in_s,
                              num_α_amino_acids_in_sample = aaa_in_s )

        results = left_join( results, cal_for_database, by = 'id' )
        # write.csv( results,
        #            paste0( dateTimeCurrent,
        #                    " ", "α_helix analysis of",
        #                    sampleNamesUpdate[i],
        #                    ".csv" ),
        #            row.names = FALSE )
        setWinProgressBar( pb_2, j,
                           title = paste( 'A-helix calculation for',
                                          sampleNames[i],
                                          '    ',
                                          round(j/length(sample_proteins)*100, 0),
                                          "% done"))
    }
    write.csv( results,
                   paste0( dateTimeCurrent,
                           " ", "α_helix analysis of",
                           sampleNamesUpdate[i],
                           ".csv" ),
                   row.names = FALSE )
        close(pb_2)
    }
    return( invisible(NULL) )
}
##################################### βSheetCalculator()
# >>
βSheetCalculator = function( pathFileInput = "C:/Users/Shashank/Desktop/peptides_second rep.csv",
                                pathDirOutput = "C:/Users/Shashank/Downloads" ) {
    print("Started")
    startTime = Sys.time()
    # Getting current working directory
    originalWorkingDir = getwd()
    # Checking if 'pathDirOutput' is provided
    if ( is.null( pathDirOutput ) ) {
        pathDirOutput = getwd()
    }
    # Reading the input sample file
    dataFileInput     = readFileInput( pathFileInput )
    df                = dataFileInput$df
    sampleNames       = dataFileInput$sampleNames
    sampleNamesUpdate = dataFileInput$sampleNamesUpdate
    # Create output folder >>>>>>>>
    dateTimeCurrent = creatOutputDir( pathDirOutput )
    # Removing the rows that are not needed
    df = removeRows( df, dateTimeCurrent )
    # Writing `dataBase_numOfAA`
    write.csv( dataBase_numOfAA,
               "dataBase_numOfAA.csv",
               row.names = FALSE )

    # B-sheet calculation for dataBase

    βSheetCalculation ( df, sampleNames, sampleNamesUpdate, dateTimeCurrent )

    # End

    endTime   = Sys.time()

    timeTaken = endTime - startTime

    print( paste0( "Time taken for the AHC run: ", format(timeTaken) ) )

    # Setting working directory back to original


    setwd( originalWorkingDir )

    return( invisible(NULL) )

}

# <<

##################################### βSheetCalculator()
########################################################################

##################################### βSheetCalculation()

# >>

βSheetCalculation <- function( df, sampleNames, sampleNamesUpdate, dateTimeCurrent ) {

    dataBase_β-   = select(dataBase_β, c(1,2))

    dataBase_reduced = dataBase_β

    num_Pro_baa    = unique(dataBase_reduced$id)

    protein        = vector()

    num_baa_pro_DB = vector()

    pb_1 = winProgressBar( title = "progress bar",

                           min   = 0,

                           max   = length(num_Pro_baa),

                           width = 300 )

    i = 1

    for( i in 1 : length(num_Pro_baa) ) {

        item                = num_Pro_baa[i]

        proteins            = filter(dataBase_reduced, id==item)

        num_baa_pro_DB_temp = length(proteins$id)

        num_baa_pro_DB      = c(num_baa_pro_DB_temp,num_baa_pro_DB)

        protein             = c(unique(proteins$id),protein)
        proteins            = vector()
        num_baa_pro_DB_temp = vector()
        setWinProgressBar( pb_1, i,
                           title = paste( 'B-sheet calculation for database     ',
                                          round( i/length(num_Pro_baa)*100, 0 ),
                                          "% done") )
    }
    close(pb_1)
    # Calculating the number of amino acids for β-####
    baa              = data.frame( id      = protein,
                                   num_baa = num_baa_pro_DB )
    cal_for_database = left_join(  dataBase_numOfAA,
                                   baa,
                                   by = 'id' )

    Sys.sleep(0.5)
    # Samples ####
    i = 1
    for( i in 1 : length(sampleNames) ) {
        temp = which( names(df) == sampleNames[i] )
        # Peptides in the sample >>
        sample_peptides = filter( df, df[,temp] > 0 )
        write.csv( sample_peptides,
                   paste0( dateTimeCurrent,
                           " ", 'List of peptides in',
                           sampleNamesUpdate[i], '.csv' ),
                   row.names = FALSE )
        sample = paste( as.character(sampleNamesUpdate[i]), '_ peptides' )
        assign( sample, sample_peptides )
        # Proteins in the sample >>
        sample_proteins = unique(sample_peptides$Proteins)
        write.csv( sample_proteins,
                   paste0( dateTimeCurrent,
                       " ", 'List of proteins in',
                       sampleNamesUpdate[i], '.csv' ),
               row.names = FALSE )
    sample = paste( as.character(sampleNamesUpdate[i]), '_ proteins' )
    assign( sample, sample_proteins )
    # Calculating β-sheet coverage for samples >>
    proteins_in_s = vector()
    aa_in_s       = vector()
    baa_in_s      = vector()
    pb_2 = winProgressBar( title = "progress bar",
                           min   = 0,
                           max   = length(sample_proteins),
                           width = 300 )
    j = 1
    for( j in 1 : length(sample_proteins) ) {
        item      = sample_proteins[j]
        Pro_chunk = filter( sample_peptides, sample_peptides$Proteins == item )
        k         = 1
        list_aa_s = vector()

        for( k in 1 : length(Pro_chunk$Proteins) ) {
            start          = Pro_chunk$Start.position[k]
            end            = Pro_chunk$End.position[k]
            list_aa_s_temp = seq(start:end)
            list_aa_s_temp = list_aa_s_temp+start-1
            list_aa_s      = c( list_aa_s_temp, list_aa_s )
            list_aa_s_temp = vector()
        }
        proteins_temp = item
        proteins_in_s = c( proteins_temp, proteins_in_s )
        proteins_temp = vector()
        aa_in_s_temp  = length( unique(list_aa_s) )
        aa_in_s       = c( aa_in_s_temp, aa_in_s )
```

```
            aa_in_s_temp  = vector()
            protein_chunk_dataBase = filter( dataBase_reduced, id == item )
            baa_in_s_temp = unique(list_aa_s)%in%protein_chunk_dataBase$n
            baa_in_s_temp = sum(baa_in_s_temp)
            baa_in_s      = c( baa_in_s_temp, baa_in_s )
            baa_in_s_temp = vector()
            results = data.frame( id                              = proteins_in_s,
                                  num_amino_acids_in_sample      = aa_in_s,
                                  num_β_amino_acids_in_sample = baa_in_s )
            results = left_join( results, cal_for_database, by='id' )
            # write.csv( results,
            #            paste0( dateTimeCurrent,
            #                    " ", "β-sheet analysis of",
            #                    sampleNamesUpdate[i],
            #                    ".csv" ),
            #            row.names = FALSE )
            setWinProgressBar( pb_2, j,
                               title = paste( 'B-sheet calculation for ',
                                              sampleNames[i],
                                              '    ',
                                              round( j/length(sample_proteins)*100, 0 ),
                                              "% done") )
        }
        write.csv( results,
                   paste0( dateTimeCurrent,
                           " ", "β-sheet analysis of",
                           sampleNamesUpdate[i],
                           ".csv" ),
                   row.names = FALSE )
        close(pb_2)
    }
    return( invisible(NULL) )
}
```

```
# <<
##################################### βSheetCalculation()
##################################### chainCalculator()
# >>
chainCalculator <- function( pathFileInput = "C:/Users/Shashank/Desktop/peptides_second rep.csv",
                             pathDirOutput = "C:/Users/Shashank/Downloads" ) {
    print("Started")
    startTime = Sys.time()
    # Getting current working directory
    originalWorkingDir = getwd()
    # Checking if 'pathDirOutput' is provided
    if ( is.null( pathDirOutput ) ) {
        pathDirOutput = getwd()
    }
    # Reading the input sample file
    dataFileInput     = readFileInput( pathFileInput )
    df                = dataFileInput$df
    sampleNames       = dataFileInput$sampleNames
    sampleNamesUpdate = dataFileInput$sampleNamesUpdate
    # Create output folder >>>>>>>>
    dateTimeCurrent = creatOutputDir( pathDirOutput )
    # Removing the rows that are not needed
    df = removeRows( df, dateTimeCurrent )
    # Writing `dataBase_numOfAA`
    write.csv( dataBase_numOfAA,
               "dataBase_numOfAA.csv",
               row.names = FALSE )
    # Chain calculation for dataBase
    chainCalculation     ( df, sampleNames, sampleNamesUpdate, dateTimeCurrent )
    # End
    endTime   = Sys.time()
    timeTaken = endTime - startTime
    print( paste0( "Time taken for the AHC run: ", format(timeTaken) ) )
    # Setting working directory back to original
    setwd( originalWorkingDir )
    return( invisible(NULL) )
}
# <<
##################################### chainCalculator()
##################################### chainCalculation()
# >>
chainCalculation <- function( df, sampleNames, sampleNamesUpdate, dateTimeCurrent ) {
    dataBase_chain   = select(dataBase_chain,c(1,2))
    dataBase_reduced = dataBase_chain
    num_Pro_caa    = unique(dataBase_reduced$id)
    protein        = vector()
    num_caa_pro_DB = vector()
    pb_1  =  winProgressBar( title = "progress bar",
                             min   = 0,
                             max   = length(num_Pro_caa),
                             width = 300)
    i = 1
    for( i in 1 : length(num_Pro_caa) ) {
        item                = num_Pro_caa[i]
        proteins            = filter( dataBase_chain, id == item )
        num_caa_pro_DB_temp = length(proteins$id)
        num_caa_pro_DB      = c(num_caa_pro_DB_temp,num_caa_pro_DB)
        protein             = c(unique(proteins$id),protein)
        proteins            = vector()
        num_caa_pro_DB_temp = vector()
        setWinProgressBar( pb_1, i,
                           title = paste( 'chain calculation for database     ',
                                          round( i/length(num_Pro_caa)*100, 0 ),
                                          "% done") )
    }
    close(pb_1)
```

```
# Calculating the number of amino acids for chain ####
caa              = data.frame( id      = protein,
                               num_caa = num_caa_pro_DB )
cal_for_database = left_join(  dataBase_numOfAA,
                               caa,
                               by = 'id' )
Sys.sleep(0.5)
# Samples ####
i = 1
for( i in 1 : length(sampleNames) ) {
    temp = which( names(df) == sampleNames[i] )
    # Peptides in the sample >>
    sample_peptides = filter( df, df[,temp] > 0 )
    write.csv( sample_peptides,
               paste0( dateTimeCurrent,
                       " ", 'List of peptides in',
                       sampleNamesUpdate[i], '.csv' ),
               row.names = FALSE )
    sample = paste( as.character(sampleNamesUpdate[i]), '_ peptides' )
    assign( sample, sample_peptides )
    # Proteins in the sample >>
    sample_proteins = unique(sample_peptides$Proteins)
    write.csv( sample_proteins,
               paste0( dateTimeCurrent,
                       " ", 'List of peptides in',
                       sampleNamesUpdate[i], '.csv' ),
               row.names = FALSE )
    sample = paste( as.character(sampleNamesUpdate[i]), '_ proteins' )
    assign( sample, sample_proteins )
    # Calculating β-sheet coverage for samples >>
    proteins_in_s = vector()
    aa_in_s       = vector()
    caa_in_s      = vector()
```

```
pb_2  =  winProgressBar( title = "progress bar",
                         min   = 0,
                         max   = length(sample_proteins),
                         width = 300 )
j = 1
for( j in 1 : length(sample_proteins) ) {
    item      = sample_proteins[j]
    Pro_chunk = filter( sample_peptides, sample_peptides$Proteins == item )
    k = 1
    list_aa_s = vector()

    for( k in 1 : length(Pro_chunk$Proteins) ) {
        start          = Pro_chunk$Start.position[k]
        end            = Pro_chunk$End.position[k]
        list_aa_s_temp = seq(start:end)
        list_aa_s_temp = list_aa_s_temp+start-1
        list_aa_s      = c( list_aa_s_temp, list_aa_s )
        list_aa_s_temp = vector()
    }
    proteins_temp = item
    proteins_in_s = c( proteins_temp, proteins_in_s )
    proteins_temp = vector()
    aa_in_s_temp  = length( unique(list_aa_s) )
    aa_in_s       = c( aa_in_s_temp, aa_in_s )
    aa_in_s_temp  = vector()
    protein_chunk_dataBase = filter( dataBase_reduced, id == item )
    caa_in_s_temp = unique(list_aa_s)%in%protein_chunk_dataBase$n
    caa_in_s_temp = sum(caa_in_s_temp)
    caa_in_s      = c( caa_in_s_temp, caa_in_s )
    caa_in_s_temp = vector()
    results = data.frame( id                             = proteins_in_s,
                          num_amino_acids_in_sample      = aa_in_s,
                          num_chain_amino_acids_in_sample= caa_in_s )
```

```
            results = left_join( results, cal_for_database, by = 'id' )
            # write.csv( results,
            #            paste0( dateTimeCurrent,
            #                    " ", "chain analysis of",
            #                    sampleNamesUpdate[i],
            #                    ".cs" ),
            #            row.names = FALSE )
            setWinProgressBar( pb_2, j,
                               title = paste( 'chain calculation for ',
                                              sampleNames[i],
                                              '    ',
                                              round( j/length(sample_proteins)*100, 0 ),
                                              "% done") )
        }
        write.csv( results,
                   paste0( dateTimeCurrent,
                           " ", "chain analysis of",
                           sampleNamesUpdate[i],
                           ".cs" ),
                   row.names = FALSE )
        close(pb_2)
    }
    return( invisible(NULL) )
}
# <<
##################################### chainCalculation()
```

4.7 REFERENCES

(1) Gromiha, M. M. *Protein Bioinformatics*; Elsevier Science, 2010.

(2) Skipper, L. *Encyclopedia of Analytical Science*; Elsevier Science, 2005.

(3) *Advances in Protein Chemistry and Structural Biology*; Elsevier Science, 2015.

(4) Staker, B. L.; Buchko, G. W.; Myler, P. J. Recent contributions of structure-based drug design to the development of antibacterial compounds. *Curr Opin Microbiol* **2015**, *27*, 133-138. DOI: 10.1016/j.mib.2015.09.003.

(5) Whittle, P. J.; Blundell, T. L. Protein structure--based drug design. *Annu Rev Biophys Biomol Struct* **1994**, *23*, 349-375. DOI: 10.1146/annurev.bb.23.060194.002025.

(6) Nero, T. L.; Parker, M. W.; Morton, C. J. Protein structure and computational drug discovery. *Biochem Soc Trans* **2018**, *46* (5), 1367-1379. DOI: 10.1042/BST20180202.

(7) ANFINSEN, C. B.; HABER, E.; SELA, M.; WHITE, F. H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci U S A* **1961**, *47*, 1309-1314. DOI: 10.1073/pnas.47.9.1309.

(8) Cuff, J. A.; Clamp, M. E.; Siddiqui, A. S.; Finlay, M.; Barton, G. J. JPred: a consensus secondary structure prediction server. *Bioinformatics* **1998**, *14* (10), 892-893. DOI: 10.1093/bioinformatics/14.10.892.

(9) Hanson, J.; Yang, Y.; Paliwal, K.; Zhou, Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **2017**, *33* (5), 685-692. DOI: 10.1093/bioinformatics/btw678.

(10) Heffernan, R.; Yang, Y.; Paliwal, K.; Zhou, Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics* **2017**, *33* (18), 2842-2849. DOI: 10.1093/bioinformatics/btx218.

(11) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **1999**, *292* (2), 195-202. DOI: 10.1006/jmbi.1999.3091.

(12) McGuffin, L. J.; Bryson, K.; Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **2000**, *16* (4), 404-405. DOI: 10.1093/bioinformatics/16.4.404.

(13) Schnoes, A. M.; Brown, S. D.; Dodevski, I.; Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* **2009**, *5* (12), e1000605. DOI: 10.1371/journal.pcbi.1000605.

(14) Kodama, Y.; Shumway, M.; Leinonen, R.; Collaboration, I. N. S. D. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* **2012**, *40* (Database issue), D54-56. DOI: 10.1093/nar/gkr854.

(15) Petersen, B.; Petersen, T. N.; Andersen, P.; Nielsen, M.; Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* **2009**, *9*, 51. DOI: 10.1186/1472-6807-9-51.

(16) Klausen, M. S.; Jespersen, M. C.; Nielsen, H.; Jensen, K. K.; Jurtz, V. I.; Sønderby, C. K.; Sommer, M. O. A.; Winther, O.; Nielsen, M.; Petersen, B.; et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins* **2019**, *87* (6), 520-527. DOI: 10.1002/prot.25674.

CHAPTER 5

DESALTING PROTEINS USING AQUEOUS-BASED ASSOCIATED SOLVENTS

Used with permission from Sajad Tasharofi, Durga Khanal, Jonathan Thacker, Morteza G. Khaledi

5.1 ABSTRACT

An innovative method for desalting proteins introduced. The two-phase that formed because of this method extracts proteins into the bottom phase and extracts the salts into the top phase. this two-phase formed addition of small portion of HFIP and butanone to aqueous solution of proteins. The top phase (aqueous phase) is composed of 96% water and the bottom phase (organic phase) is composed of 80% HFIP and butanone. Three proteins, lysozyme, ubiquitin, and RNase salinated with 5 different salts and desalted using this method. Desalting lysozyme from a 100 mM $MgCl_2$ resulted in signal intensity that is 84% of the signal of lysozyme in aqueous solution without any salt in it. Similarly, for desalting ubiquitin from a 100 mM $MgCl_2$ resulted in signal intensity that is 86% of the signal of ubiquitin in aqueous solution without any salt in it. It is noteworthy that these proteins have no signal in the solutions with 100 mM salt in them.

5.2 INTRODUCTION

Analysis of intact proteins often requires samples that are essentially salt-free. SDS-PAGE and mass spectrometry with electro spray ionization (ESI) as ionization source, are examples that require protein desalting. Salt in the protein sample interferes with electrophoresis and in mass spectrometry suppresses ionization of proteins at the ion source, therefore, decreases protein signal and in high concentration of salt makes the protein undetectable. In top-down mass spectrometry of proteins, protein desalting is particularly important since presence of salt, even at low levels, can adversely affect signal intensity and suppress detection of post translational modifications (PTMs) and proteoforms[1]. There are many methods for desalting the protein samples and the choice of the methods depends on many factors including: volume of the sample, protein concentration in the sample, sample matrix, sensitivity of protein (pH and organic solvents), etc.

Dialysis was one of the first methods developed for separation of small ions form proteins using a semi-permeable membrane and electrical current. In the case of protein desalting, the dialysis membrane retains proteins and passes the salts. However, protein loss due to adsorption to the membrane is relatively high and precautions should be taken for desalting of samples with low amounts of proteins. In addition, dialysis efficiency is very dependent on the temperature and viscosity of the protein solution[2].

Ultrafiltration is another desalting method that is widely used to concentrate and desalt protein samples. Like dialysis, this method also relies on a permeable membrane (filter) to desalt protein samples; the filter is permeable to small molecules and not for large molecules, like proteins. These membranes are available in different molecular cut-off sizes and volumes. Large amount of sample loss may occur for proteins that have near or smaller than the filter pore size. Ultrafiltration is a suitable method for protein desalting; however, like dialysis sample loss due to adsorption to the filter makes it unsuitable for dilute protein samples. On the other hand, for concentrated protein samples, formation of thick layers of proteins can clog the filter and cease the flow of salt solution through the filter[3]. Partially clogged filters require extended filtration time which increases the chance of post translational modification or any chemical changes in the structure of the protein.

Precipitation of proteins has been used for desalting and concentrating low amount of proteins. Methanol in combination with acetone[4] or methanol in combination with chloroform[5] are the most common solvents for precipitating and separating proteins from salts. The biggest concern using these methods is coprecipitation of salt with protein and protein loss especially when the sample contains low amounts of proteins.

Size exclusion LC columns are also used for purification of proteins from salts and small molecules. This method is particularly effective for desalting proteins in the native forms. However, LC separation leads to sample dilution. Porath demonstrated that gel-forming polymers have the ability to hold small ions but not big molecules like serum proteins [6]. Hedlund discussed desalting of proteins with size exclusion chromatography (SEC)[7].

Among all protein desalting methods, Reversed Phase sorbents are the most common way of desalting protein samples[8]. This method has been very effective for desalting low amount of proteins with little or no sample loss. Naldrett et al. used a C18 membrane to desalt proteins like BSA[9]. Elution of proteins from the sorbent requires substantial amount of organic solvent which will dilute the protein sample. Another requirement is the use of low pH media to ensure adequate protein adsorption on the sorbent. Acidic media can lead to hydrolysis of proteins. Paul et al. demonstrated desalting of proteins using HILIC column[10].Desalting methods mentioned above are not suitable for all samples in top-down proteomics; a need for alternative for sample clean-up method still exists[1].

In this manuscript we introduce a new method for desalting proteins. using fluoroalcohol-Organic solvent Two-phase System (FOAS) that can address the drawbacks of the current methods. In this method, protein solutions are desalted in a two-phase system that is composed of water (75% V/V), an organic solvent (butanone, 7.5% V/V) and Hexafluoro-2-propanol (HFIP, 7.5% V/V). The top phase is called aqueous phase because it is mostly composed of water (96% V/V) and the bottom phase is called H-O phase because it is mostly composed of HFIP and organic solvent (HFIP, 35% V/V; organic solvent, 40% V/V).; this would differentiate between these methods and the conventional methods for protein precipitation which uses chloroform as a main component.

Chloroform is not miscible with water and when it's added to the aqueous solution of protein (in addition to methanol or acetone) makes a two-phase system. This two-phase system forms because of vast difference between hydrophobicity of water and chloroform (water and oil). Formation of this two-phase system precipitates the proteins in the interface of the two phases and salts remain in the top phase that is mainly composed of water.

In FOAS, however, organic solvent (at the percentages we use in this experiment) and HFIP are miscible in water and the two-phase forms for a different reason; the two-phase system forms because of association of HFIP, organic solvent and water. Although the H-O phase is mostly composed of HFIP and organic solvent, there is considerable amount of water present in this phase (20% V/V) which makes the H-O phase more capable of dissolving proteins and not precipitating them. Presence of HFIP in the H-O phase makes it denser than the aqueous phase which settles at the bottom of the centrifuge tube. Aqueous phase on the top, which contains most of the salt, easily separated from the H-O phase and goes to waist and the bottom phase, with desalted proteins will remain. Unlike size exclusion filters and C18 methods, this method does not require expensive material (filters, C18 columns) and lab equipment (high speed centrifuge). In this document the formation, characteristics and the results of desalting using the FOAS explained to the full extend.

5.3 EXPERIMENTAL

5.3.1 Materials

Millipore-DI water was used for sample preparation.1,1,1,3,3,3-Hexafluoro-2-propanol (HFIP) was obtained from Oakwood Chemical, USA. 2-butanone was purchased form Alpha Aesar with the purity of 99%. Formic acid (99%) was purchased from Alfa Aesar. Ubiquitin with purity of >98% was purchased from Sigma-Aldrich, ribonuclease A was purchased form Fisher Scientific, Lysozyme was purchased form Fisher Scientific. $MgCl_2$ was purchased form Alpha Aesar.

5.3.2 Two-phase formation and protein desalting

All samples have been prepared in a 1.5 mL microcentrifuge tubes; each sample has a total volume of 1 mL and is composed of 850 µL water, 75 µL organic solvent, 75 µL HFIP. Concentration of proteins in the samples varies based on the experiment. Butanone was used as organic solvents in this experiment. Proteins are dissolved in water and then butanone added to the protein solution. The solution then vortexed for few seconds to mix the components and then HFIP added to the mixture. Addition of HFIP induces the two-phases and then the the solution is vortexed/agitated for 1 min and then centrifuged for 4 min at 2000 g. after centrifugation the two-phase are segregated in the tube with a very distinct borderline between the phases. The top phase (aqueous) has a volume of ~ 900 µL and the H-O phase (organic phase) has a volume of ~100 µL. The aqueous phase then separated from the organic phase using a 200 µL pipette. This step should be performed very accurately because if the extraction of aqueous phase is not complete, the remaining little droplets form the aqueous phase can heavily contaminate the desalted proteins in the H-O phase due to high concentration of salt in the aqueous phase. To prevent contamination, od H-O phase by the droplets of aqueous phase that are beading on the inner surface of centrifuge tube, the H-O phase extracted with a pipette and transferred to a new microcentrifuge tube. The H-O phase then diluted with 100 µL of water to induce another phase separation to back-extract the protein form organic phase to the aqueous phase. The sample then vortexed/agitated for 1 min and centrifuged for 1 min at 2000g to separate the phases in the centrifuge tube. The solution in the top phase that contains the desalted protein, analyzed through flow injection to mass spectrometer (MS).

Control sample was 10 µM of standard proteins in a salt-free aqueous solution. Known amounts of various salts were added to the standard proteins solutions followed by formation of the two-phase system to separate the salt and protein in separate phases. All samples were prepared in a 1.5 mL polypropylene microcentrifuge tubes; all samples have a total volume of 1 mL, and were composed of 850 µL water, 70 µL organic solvent, 80 µL HFIP and 100 mM of different salts. Concentration of protein in each sample is 10 µM and concentration of salt is 100 mM except mentioned otherwise.

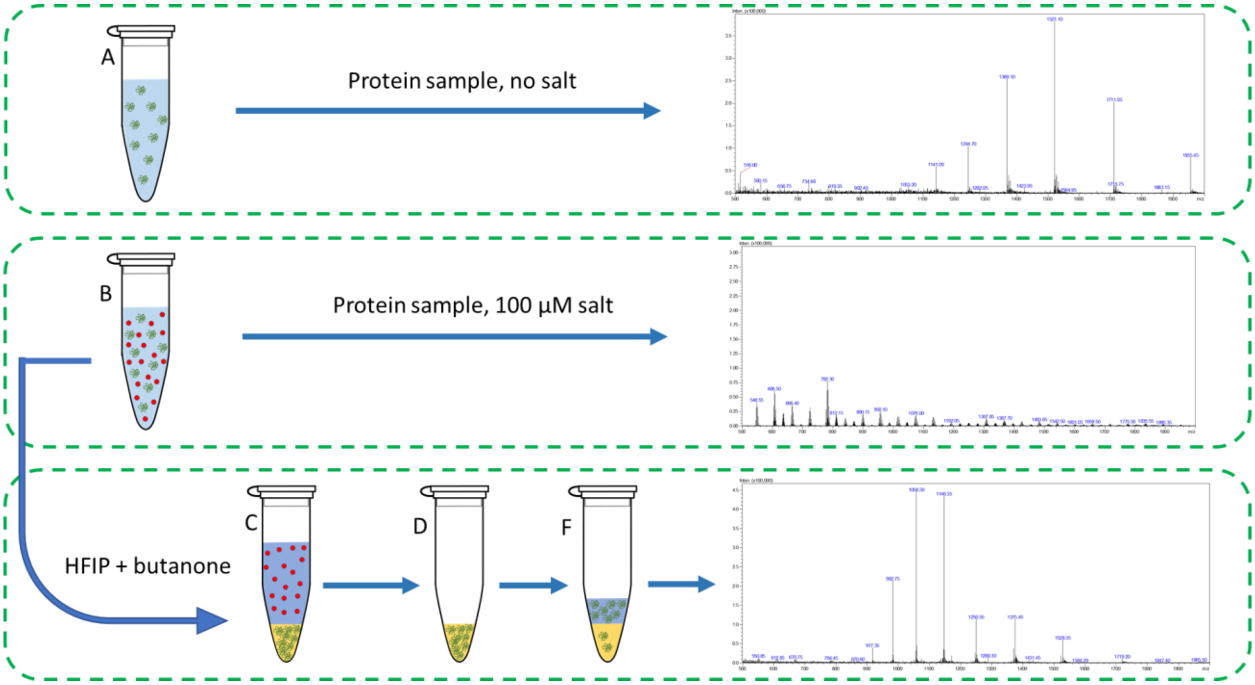


Figure 5-1. Workflow of forming the FOAS to desalt the sample. The green box on top shows MS analysis of aqueous solution of a protein with no salt (A). The middle box shows addition of salt (red dots in the tube B) to the protein solution (A) eliminates the signal in MS. The bottom box shows that using FOAS can desalt the sample in few fast steps. After addition of HFIP and butanone to the tube B, two-phase formed (C), the top phase in tube C contains most of the salt. The bottom phase in tube C contains desalted proteins which is separated (D) and diluted with water to induce a two-phase to back-extract the protein to the aqueous phase and analyze with MS.

5.3.3 GCMS analysis

In addition to the phase diagram for the butanone system, we also analyzed the chemical composition of each phase using GC-MS. In this experiment, the GCMS-2010SE (Shimadzu) instrument was used with a capillary column (30 m × 0.25 mm × 0.25 µm), and the stationary phase 5% Phenyl-Arylene 95% Dimethylpolysiloxane. The carrier gas was helium, and the column flow rate was 0.9 mL/min. The analysis was done in single ion mode (SIM) for better sensitivity. The experimental parameters are listed in Table 1. Calibration plots were made and the concentrations of the solvents in each phase were determined. The top aqueous phase and the bottom organic phase were diluted several times prior to analysis.

Table 5-1. GCMS method program and parameters

Samples	Sample injection			Carrier gas flow (mL/min)		Oven profile			Ion source Temp (ºC)
	Inj. temp.(ºC)	Inj. Vol. (µL	Split ratio	Carrier gas	Column flow (mL/min)	Rate (ºC/min)	Oven temp. (ºC)	Hold time (min)	
Solvent mixture	200					-	100	0.5	250
		1	0.3888889	Helium	0.9	25	200	0	

5.3.4 LCMS analysis

Desalted proteins after second extraction were directly analyzed by a LCMS-2020 single quadrupole mass spectrometer manufactured by Shimadzu. The LC pump is LC-20AD XR, the auto sampler is SIL-20AC XR. Auto sampler was directly connected to ESI source of mass spectrometer using 1 m long peek tubing with 1/16 "OD and 0.005" ID. Injection volume is 5 µL and the flow rate for the carrier is 0.1 mL/min. The carrier used for flow injection is 50% ACN with 0.1% formic acid. Scan range of mass spectrometer is from 500 - 2000 Da, event time is 0.5 sec, interface voltage is 4.5 kV, interface temperature is 350 °C, DL temperature is 250 °C, heat block temperature is 400 °C and nebulizing gas flow is 1.5 L/min. Samples are carried to mass spectrometer via direct infusion using flow injection.

5.3.5 Salt concentration measurement

We chose conductivity measurement for measuring total ion concentration in the sample. Ring-disk electrodes (RDE) are suitable for measurement of solution with low volume, however, the commercial RDEs are too large for the microliter size sample volumes in FOAS. A RDE was made using a metal rod, a metal tube, and insulating plastic tubes to be placed between the metal rod and the metal tube and as the outer layer. The structure of the electrode demonstrated in Figure 5-2.

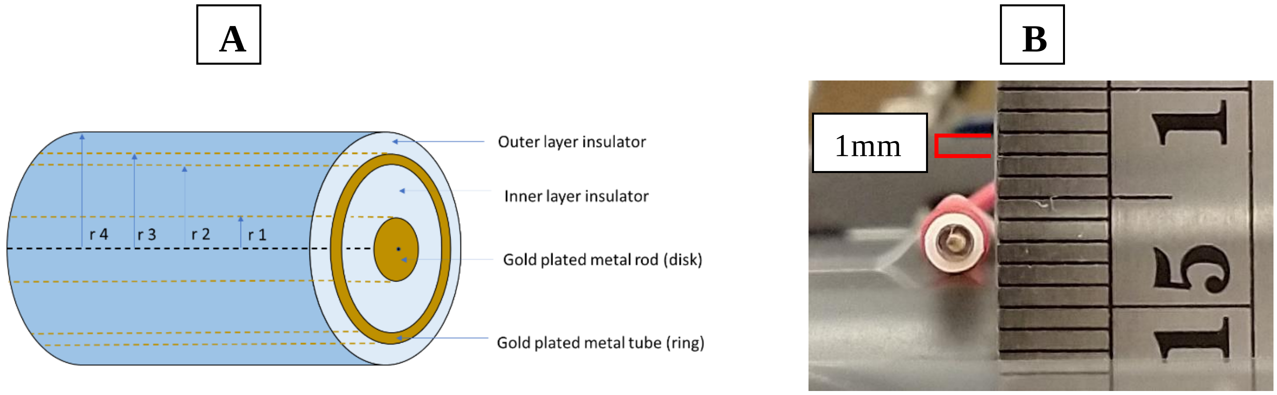


Figure 5-2. structure of the RDE, r_1 is the outer diameter (OD) of the disc electrode, r_2 is the inner diameter (ID) of the of ring electrode and r_3 is the OD of the ring electrode.

Specific conductance of a solution, σ, is directly related to concentration of charged solutes in the solution. Specific conductance is also directly related to measured conductance, *G* as it is explained in the equation 1.

$$\sigma = G\mathrm{K} \tag{1}$$

K in equation 1 is the cell constant in conventional conductivity cells which is composed of two parallel electrodes with the same area which are placed apart at a certain distance. Since the electric field is contained in the space between the electrodes, conductivity measurement using this electrode is representative of conductivity of the solution that is placed in the space between the electrodes and not dependent on the solution beyond that space. However, in RDE arrangement, the electric field extends to the space beyond the surface of the electrode. This would make the conductivity measurements using this electrode dependent on the depth of the solution (D) extending from the surface of the electrode. As D extends to infinity, G will reach to its limiting value, G_∞. The measurements for this experiment are performed at the depth of the solution at which 99% of G_∞ is reached. Dasgupta et al. calculated the depth of the solution (from the surface of the electrode) should be at least 1.66 mm to get 0.99 G_∞ value[11].

Using the RDE and the methods mentioned earlier, calibration curves for conductivity of different salts performed. Figure 5-3 shows theses calibration curves with the linear fit.

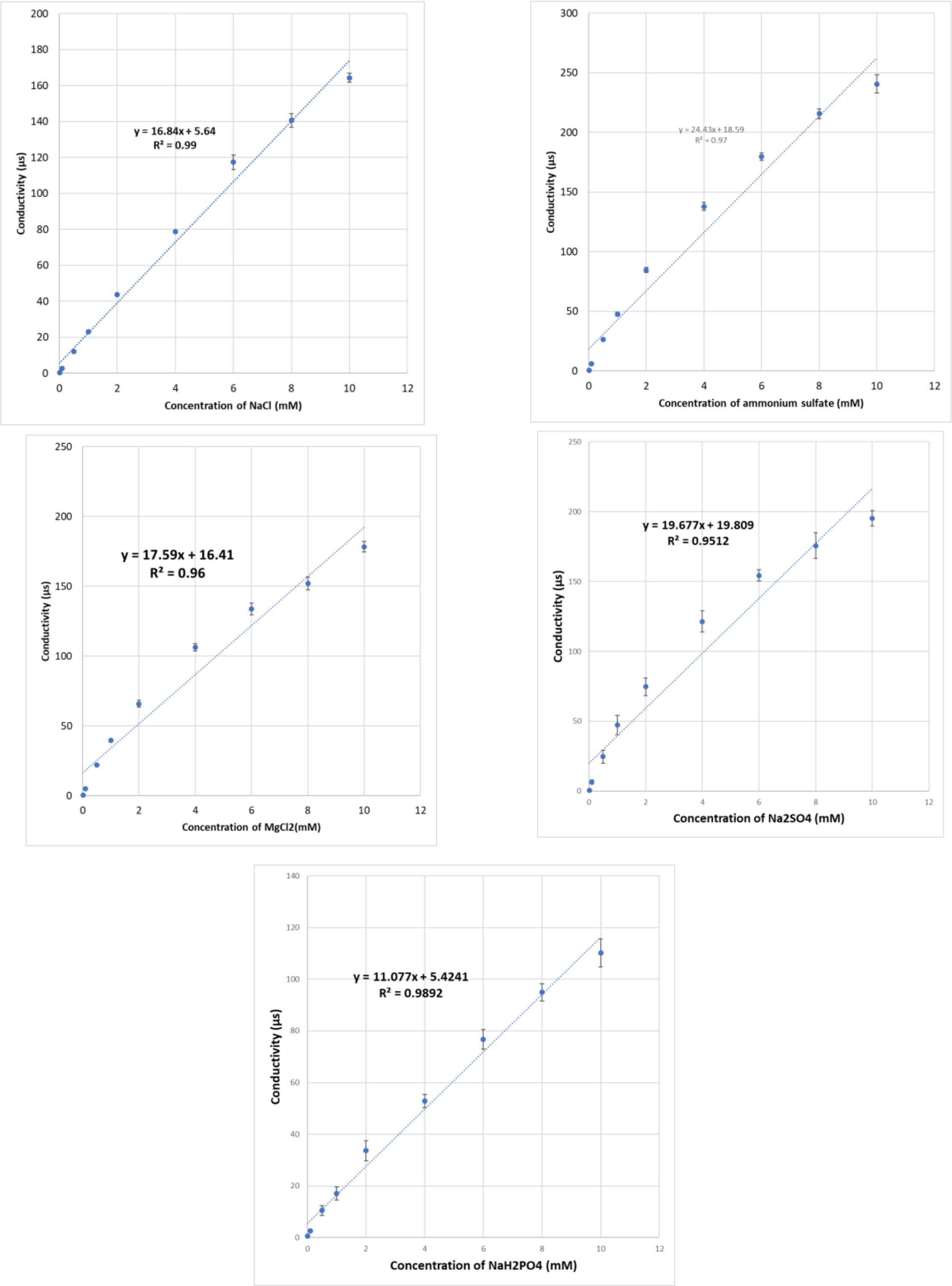


Figure 5-3. Calibration curves for conductivity measurements of salts that are used in the experiment.

5.4 RESULTS AND DISCUSSIONS

5.4.1 Phase diagram and compositional analysis of the FOAS

We determined the phase transition to two-phase system as a function of butanone and HFIP concentration, The volume fraction of the Organic (H-O) phase, (defined as the volume of the H-O over the total volume, 1mL) or the phase ratio (volume of the H-O phase/volume of aq phase) increases with an increase in HFIP concentration, however, it remains nearly the same or increases slightly with an increase in butanone concentration at a constant percentage of HFIP.

Figure 5-4 is a contour plot representing changes in the volume of organic phase in butanone/HFIP/water system.

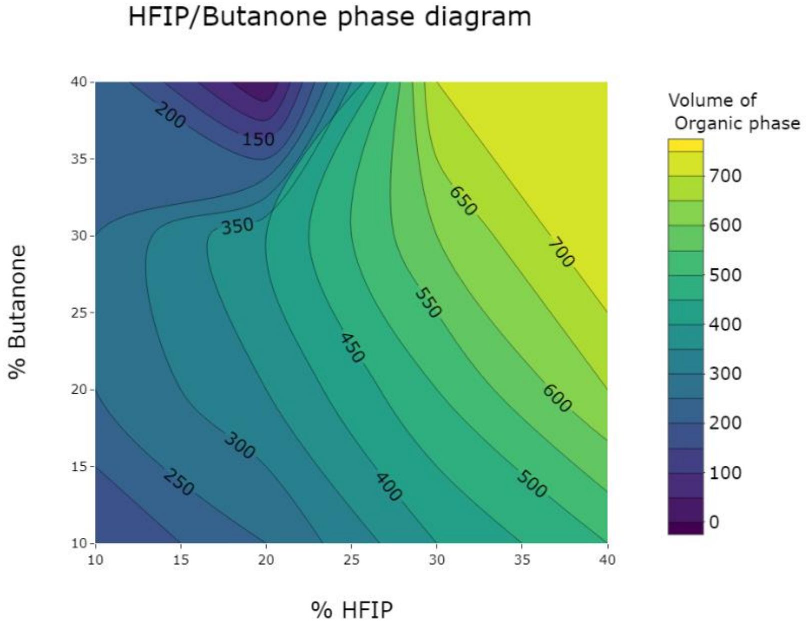


Figure 5-4. phase diagram of Butanone/HFIP/ water system; different colors in the contour plot correspond to different volumes of the Organic (H-O) phase

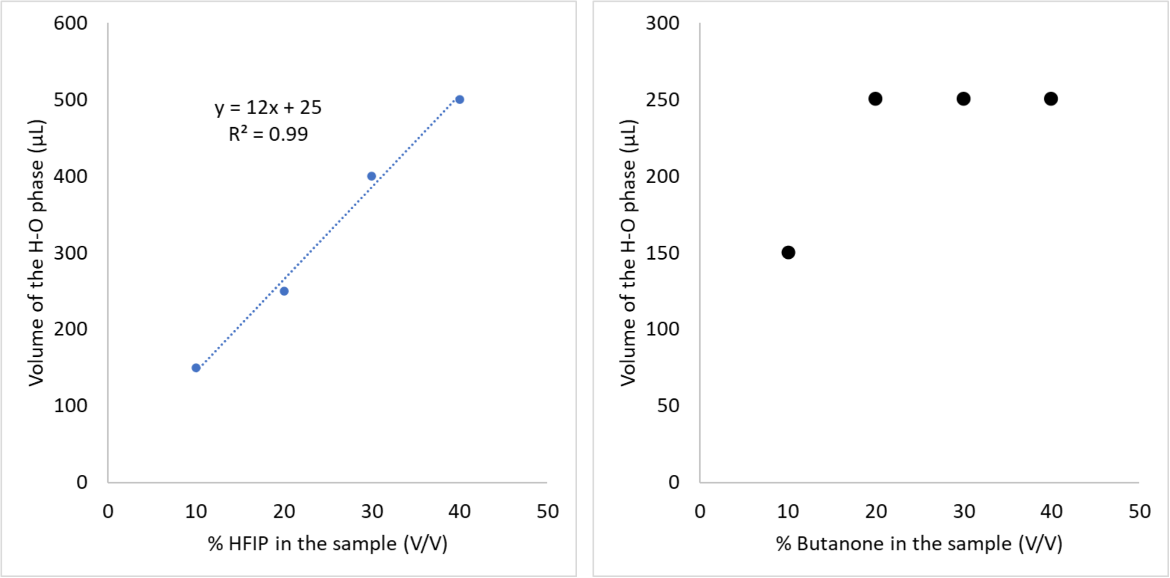


Figure 5-5. left: shows changes in the volume of the H-O phase when % of butanone kept constant and % HFIP changes from 10-40%

Right: shows changes in the volume of the H-O phase when % of HFIP kept constant and % butanone changes from 10-40%. In this case the volume of H-O phase does not change by increasing % butanone more than 20%.

5.4.2 Compositional analysis of the FOAS

Samples with different percentages of butanone and HFIP made (from 5% to 40%; for butanone and HFIP) and after formation of the two-phase, the two phases separated. The aqueous phase diluted 5-times and the organic phase diluted 40-times. Then samples from each phase analyzed with GCMS to determine the percentage of butanone and HFIP in each phase. Table 2 shows the results of analysis.

Table 5-2. Samples with different percentages of butanone and HFIP made and after formation of the two-phase, each phase analyzed to determine the percentage of butanone and HFIP

% HFIP in the sample	% butanone in the sample	% HFIP in the aqueous phase	% HFIP in the organic phase	% butanone in the aqueous phase	% butanone in the organic phase
5	5	1.6	35.0	2.1	41.0
8	8	1.7	36.0	3.1	44.0
10	10	1.9	37.4	1.9	40.4
10	20	0.6	26.6	7.2	57.0
10	30	1.5	22.1	16.1	68.5
10	40	1.1	17.4	20.2	75.1
20	10	4.0	42.3	0.1	24.0
20	20	1.6	34.5	2.3	39.0
20	30	1.0	29.6	6.0	52.0
30	10	5.6	42.8	0.0	13.3
30	20	2.9	38.8	0.5	28.6
30	30	2.4	37.4	3.7	43.3
30	40	0.9	33.3	5.2	56.4
40	10	6.5	42.3	0.0	9.0
40	20	4.0	42.6	0.1	23.3
40	30	2.7	41.7	1.6	36.7
40	40	1.3	37.0	2.5	45.3

Figure 5-6 is contour plots of the data in table 2 (excluding 5% and 8% data) which provides a visual representation of how composition in each phase changes as we change the percentage of butanone and HFIP in the sample.

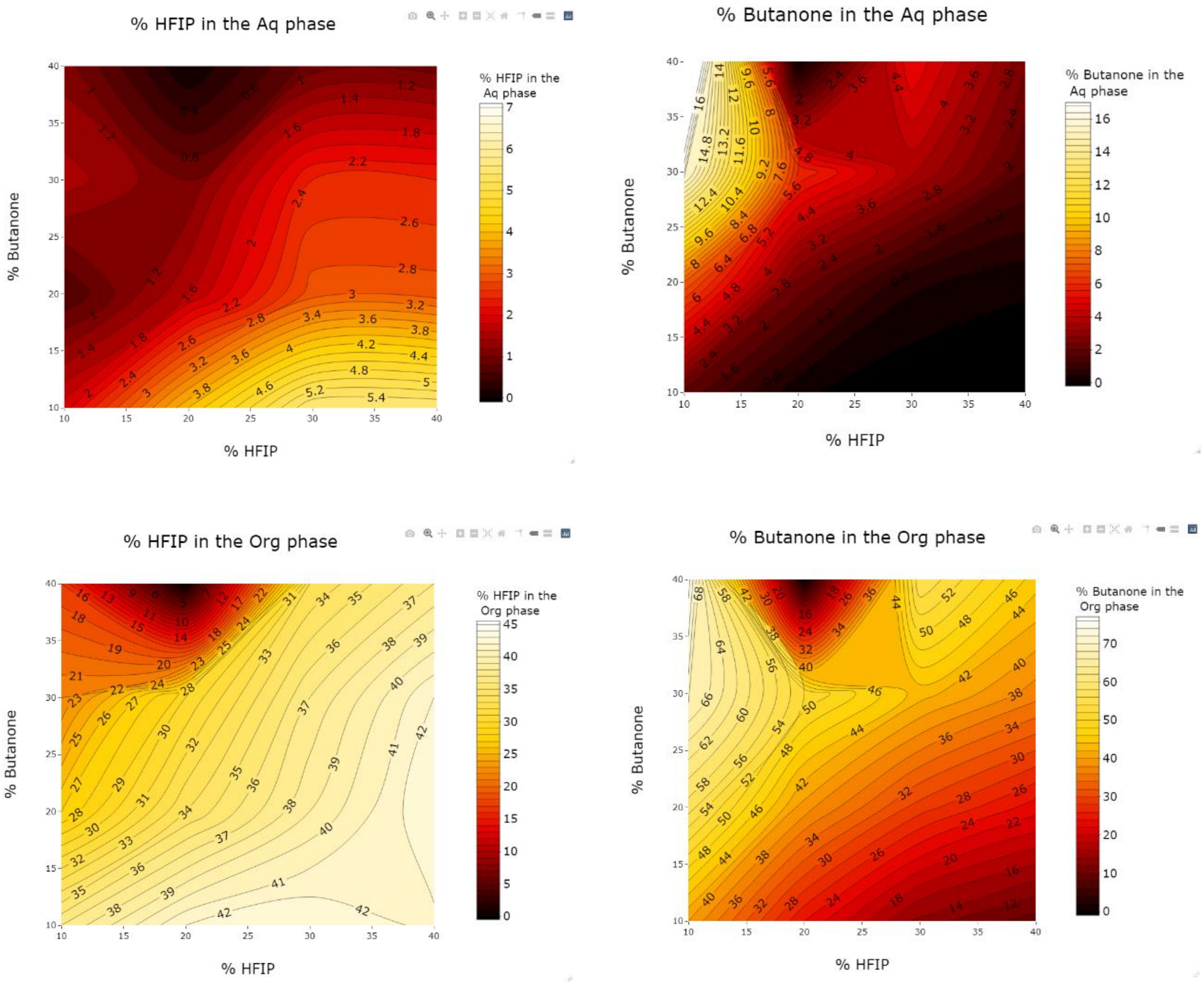


Figure 5-6. Compositional analysis for determine percentage of HFIP and butanone in aqueous and organic phase

5.4.3 Fractionation of salt and protein in the FOAS

Based on the results of compositional analysis, almost 80% of the H-O phase is composaed of organic solvents and 20% water. The aqueous phase, however, is mostly composed of water (almost 96%). This vast difference between composition of aqueous and organic phase makes this system able to successfully extrcat and dissolve the proteins into H-O phase and retain the salt in the aqueous phase. To examine this, we prepared a salinated solution (NaCl) of Bovine Serum Albumin (BSA) in 850 µL of water and made a two-phase system total volume of1 mL by adding 75 µL HFIP and 75 µL butanone to it. The concentration of protein in the sample is 400 µg/mL and concentration of salt is 100 mM. The aqueous and the H-O phases were separated, and the salt and protein concentrations were determined in each phase. BSA concentration measured using built-in method in Thermo Scientific™ Nanodrop™ One. Before protein concentration measurement 20 µL of each phase dried in separate microcentrifuge tubes and reconstituted in water to prevent interference of organic solution in the FOAS with Nanodrop measurement. For salt concentration measurement 50 µL of each phase dried in separate microcentrifuge tubes and reconstituted in water; the sample from organic phase for salt concentration measurement diluted 2-times and sample from aqueous phase for salt concentration measurement diluted 10-times.

The results show that the concentration of BSA in the H-O phase is 10-times higher than that in the aqueous phase, and conversely the concentration of salt in the aqueous phase is 45-times higher than that in the H-O phase. Figure 5-7 shows the fractionation of salt and protein between the two phases.

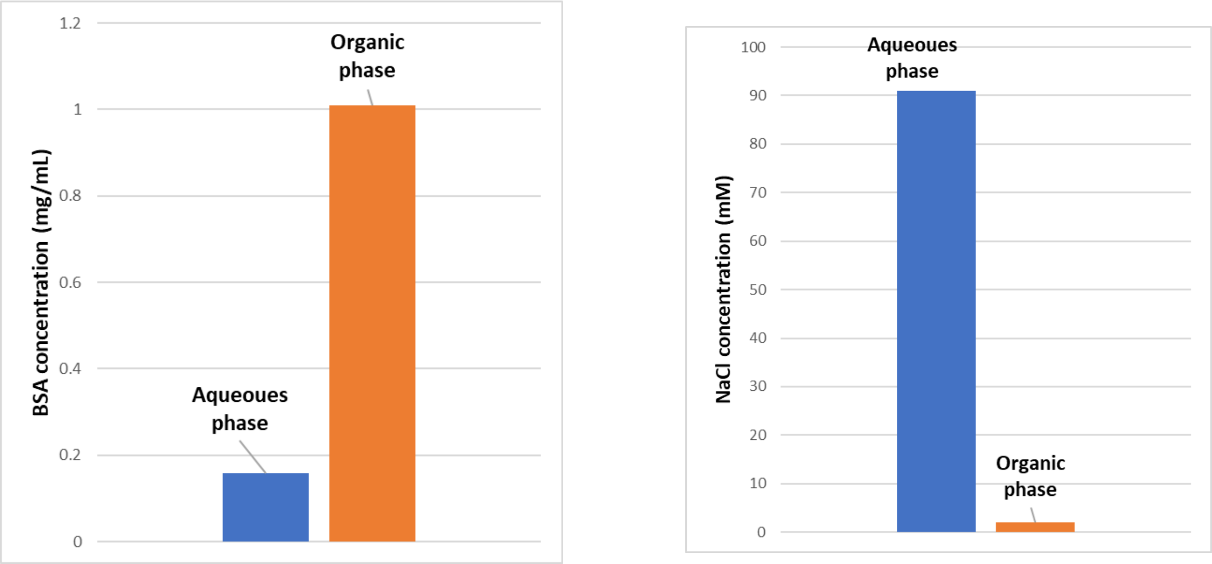


Figure 5-7. Fractionation of salt and protein in the phases in the sample of BSA with 100 mM NaCl

To examine whether this is true for other protein samples or not, samples of 5 other proteins prepared and the same measurement for salt repeated and results show that the same fractionation pattern for salt and protein has been observed. Table 1 shows concentration of salt in both phases for different proteins.

Table 5-3. concentration of salt in both phase of different samples with different proteins

	salt con in Aq (mM)	salt con in Org (mM)	Vol of Aq (µL)	Vol of org (µL)	Total salt in Aq (mmol)	Total salt in Org (mmol)	% salt mass in org
Gramicidin	140	4	840	160	117	0.7	0.6
BSA	126	5	860	140	109	0.8	0.7
Yeast	131	12	860	140	112	1.7	1.5
LGB	133	4	860	140	114	0.6	0.5
Ovalbumin	110	13	870	130	96	1.7	1.7
Myoglobin	127	9	870	130	111	1.2	1.1

It would be interesting to know if initial concentration of salt in the sample can change this fractionation pattern for salt and protein between two phases. To examine this, different sample salinated BSA solutions (400 µg /mL) were prepared with concentration of salt (NaCl) from 100 mM up to 2000 mM. After making the two-phase, phases separated, and concentration of salt and protein measured using the same methods that explained earlier. four samples salinated BSA solutions with conc of NaCl 100, 200, 1000 and 2000 mM was made. For these samples the aqueous phase diluted 10-times, 20-times, 200-times, and 200-times respectfully.

Interestingly, the results show that concentration of salt in the H-O phase is independent of initial concentration of salt in the sample; this is an interesting finding because it shows that this system can exclude salt from H-O phase regardless of the concentration of salt in the protein sample that is subject to desalting using this system.

Another interesting result is that concentration of protein in the aqueous phase decreases as initial concentration of salt in the increases, indicating that protein is more likely to be slat-out form the aqueous phase that contains most of the salt and for the purpose of desalting proteins, it is deposed.

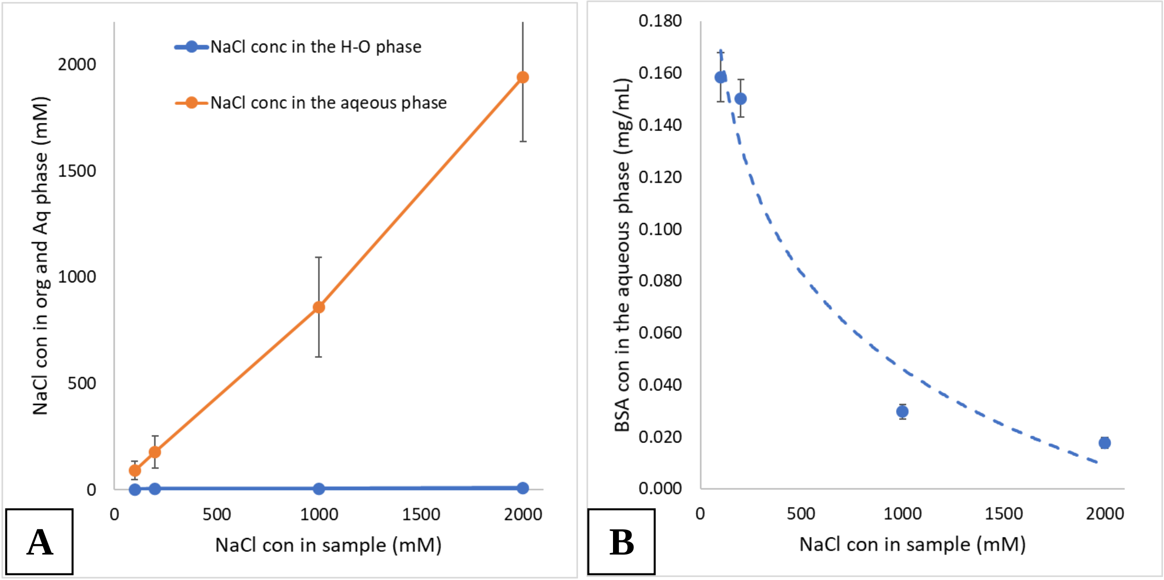


Figure 5-8. A: upon increasing NaCl conc in the sample, NaCl conc increases in the aqueous phase and remains the same in the organic phase. B: upon increasing NaCl conc in the sample, BSA conc in the aqueous phase decreases.

In addition to BSA, the same experiment repeated for 3 more proteins: RNase, lysozyme, and myoglobin. The results for these 3 proteins are very much like the results of BSA. For all 3 proteins under study, as the NaCl conc increases in the sample, concentration of salt shows almost no change in the H-O phase but for the aqueous phase it increases linearly as NaCl conc increases in the sample.

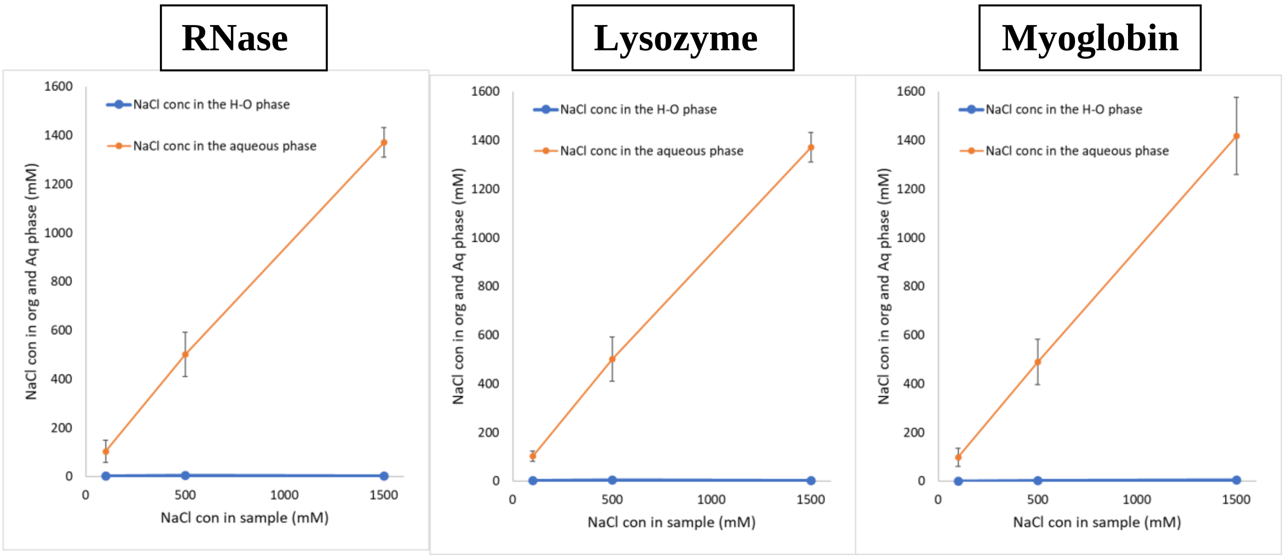


Figure 5-9. Upon increasing NaCl conc in the sample, like BSA, the same fractionation pattern for NaCl observed for RNase, lysozyme, and myoglobin as well.

5.4.4 MS Signal suppression due to salt

As mentioned earlier, salts can suppress ionization of proteins in the ESI that leads to reduction of proteins MS signal intensity. Figure 5-10 shows the magnitude of such effect on the MS-ESI detection of RNase. In the Figure 5-10 we compared signal intensity of 1 µM RNase in water and in the presence of NaCl at different concentrations. As shown in Figure 5-10, at 1 mM NaCl, the protein signal dropped almost 10-times. At concentrations greater than 1 mM, the protein signal diminished entirely. Similar results were observed for Lysozyme and Ubiquitin, where the protein signal completely disappeared in the presence of 100 mM NaCl. Figure 5-11 shows signal elimination by 100 mM NaCl in lysozyme and ubiquitin solutions.

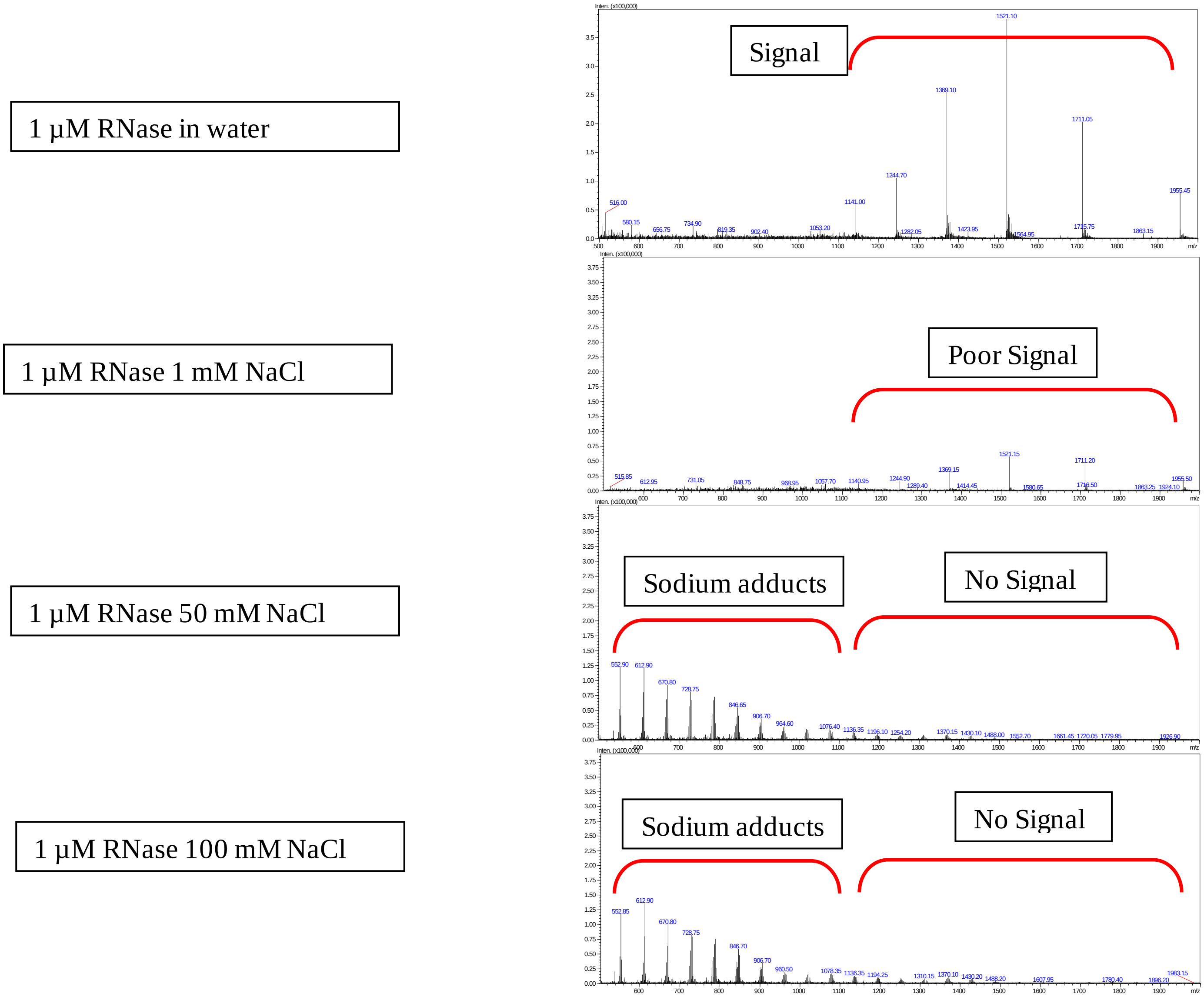


Figure 5-10. signal suppression in ESI-MS. As we increase the concentration of salt in a sample of RNase the signal intensity decreases and at high concentrations, only sodium adducts are visible in mass spectra.

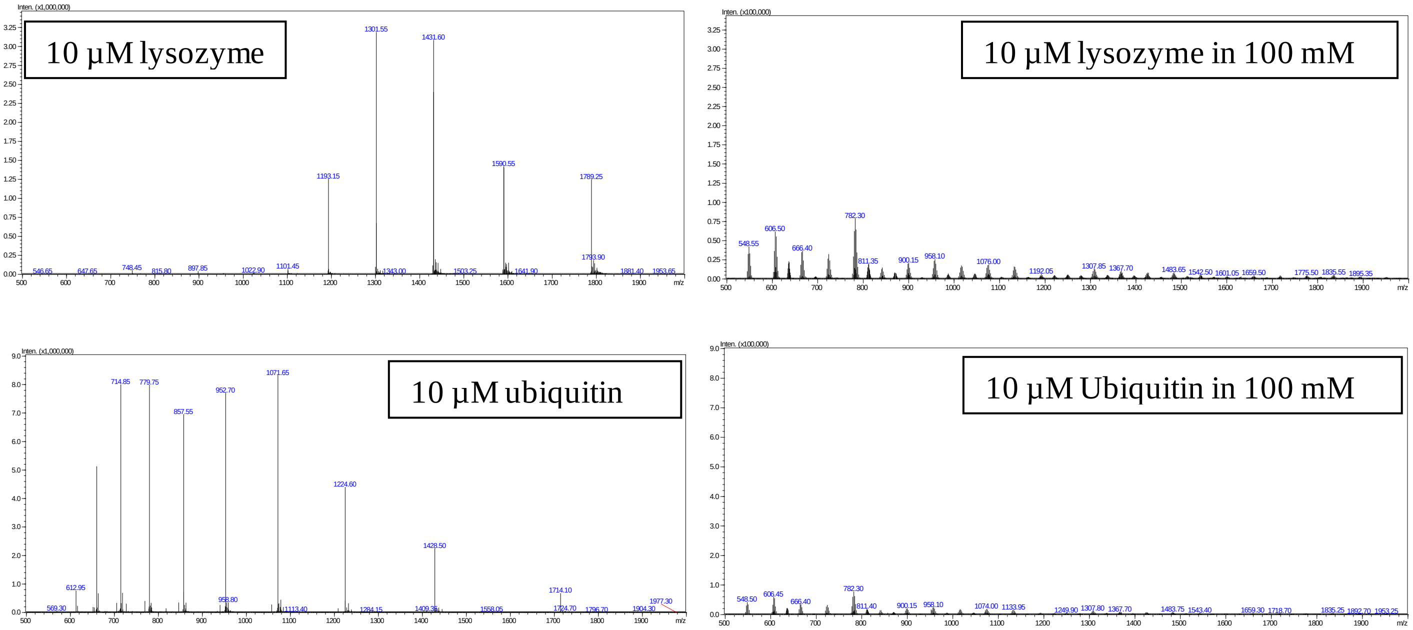


Figure 5-11. In addition to RNase, signal intensity for lysozyme and ubiquitin also goes to zero as we add 100 mM NaCl in the sample of proteins

5.5.5 Protein desalting using two-phase system

As described earlier, in the BSA sample, salt is extracted in the aqueous phase while most of the protein is extracted into H-O (Organic) phase. The upside of FOAS is the fact that 99% of initial mass of salt in the sample remained in the aqueous phase which leaves the proteins extracted in the Organic phase, almost salt free. The next step would be desalting a salinated protein solution and detecting it in the MS. For this purpose, a sample of RNase prepared based on the methods described in experimental section. Concentration of protein in this sample is 10 µM and it is salinated with 100 mM NaCl. Based on the workflow described in Figure 5-1, the two-phase formed and the H-O phase containing desalted protein separated from the aqueous phase and 5 µL of it directly injected to MS through flow injection. Figure 5-12 shows the workflow and the MS result of analysis. Unfortunately, the protein was not detectable in the although the presence of protein in the H-O phase was confirmed by protein conc measurement using Thermo Scientific™ Nanodrop™ One. Since the concentration of organic solvents are high in the H-O phase, they could be source of interference with ionization of protein in ESI. To examine whether organic solvents used in this sample can do such interference, solutions of RNase in 20% butanone (because butanone has limited solubility in water) and 90% HFIP (because RNase in not soluble in 100% HFIP) prepared and analyzed in MS. Results show that the solvents are not the source of interference since RNase is very well detected those solutions. Figure 5-13 shows that presence of organic solvents does not suppress detection of RNase.

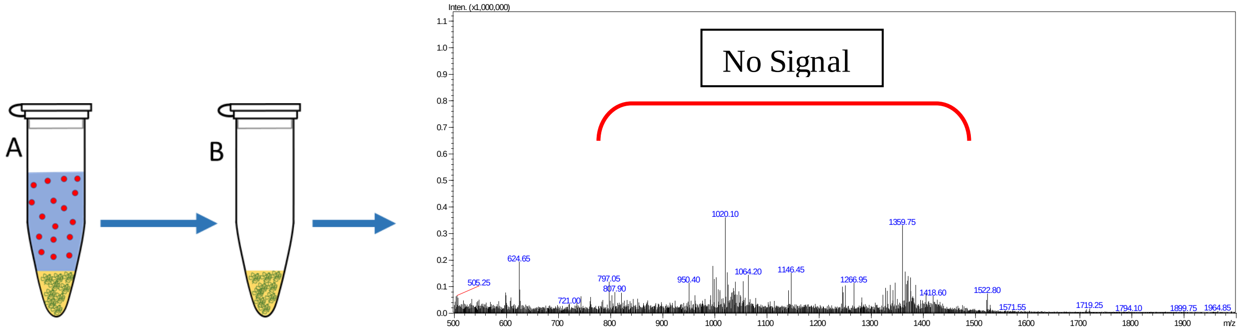


Figure 5-12. Tube A: RNase sample is desalted using FOAS; Tube B: The H-O phase from Tube A analyzed in MS. The analysis shows that the protein in not detectable in MS which is possibly because of poor ionization in ESI.

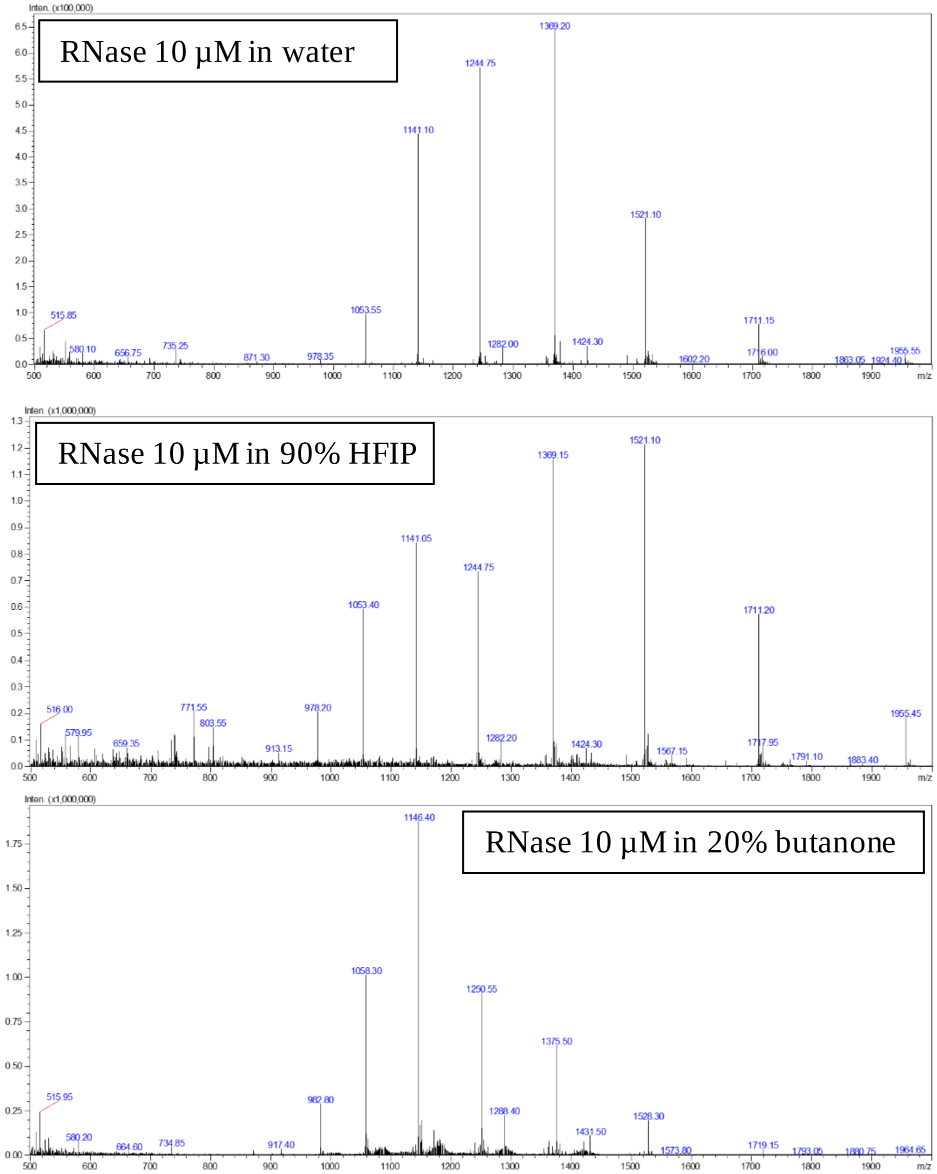


Figure 5-13. RNase is detectable in MS in all the solvents used to form FOAS.

The role of organic solvents in interfering with ionization and detection of RNase was ruled out. A possible reason for lack of detection of protein in H-O phase would be complex structures that is formed from conglomeration of HFIP, butanone, water, and protein. The structure is probably stable enough to prevent ionization of proteins in ESI.

A possible solution for this is evaporation of HFIP and butanone form the H-O phase to break the complex structure and release the protein. for this purpose, after formation of the two -phase system, the H-O phase separated and diluted with water and placed in Eppendorf Vaccufuge® plus to dry the solvents. The remaining solution (~200 µL) injected to MS and results analyzed. Figure 5-14 shows the works flow and the results of partial evaporation of HFIP, and butanone form the H-O phase. the results show that although the protein is detectable, the signal intensity is very low. Partial evaporation of H-O phase, to some degree, helped with the detection of protein in MS.

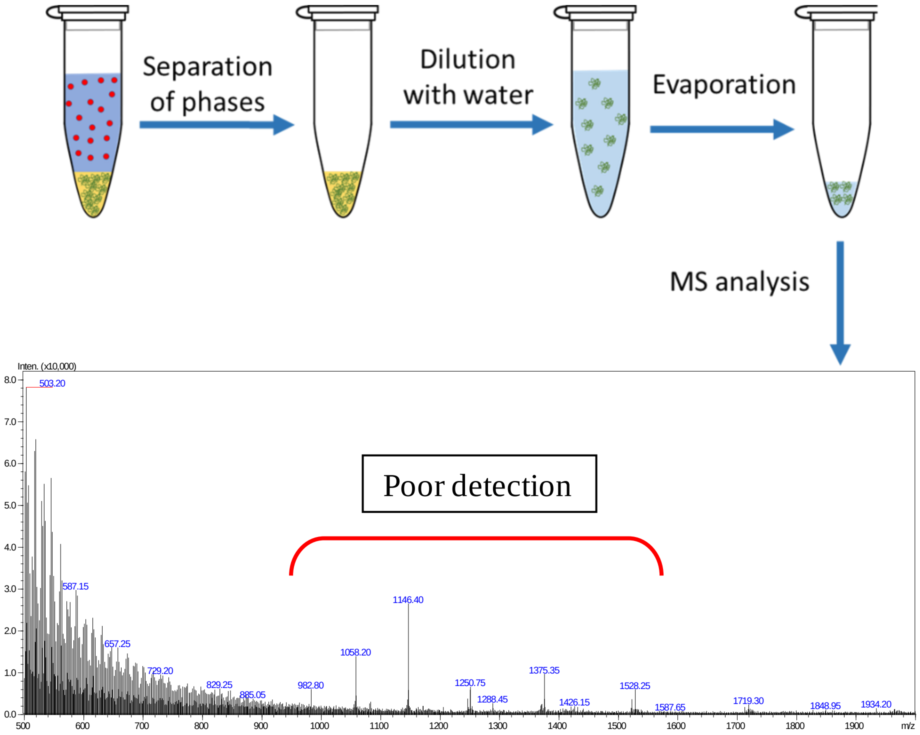


Figure 5-14. Partial evaporation of H-O phase to release the protein from possible complex structures. After forming two phase system, the H-O phase separated and diluted with water and placed in vacuum centrifuge to dry the solvents. The remaining solution injected to MS and analyzed.

Although partial evaporation of H-O phase made detection of protein possible, the signal intensity is poor, and the results are not satisfactory.

An alternative method for releasing proteins form the possible complex structures examined. In this method after formation of two-phase system the H-O phase separated and diluted with ~100 µL of water; addition of water induced another two-phase. Then, the solution vortexed/agitated for 1 min and centrifuged for 1 min at 2000g. The top phase (aqueous phase) separated and analyzed in MS. We call this process double extraction since proteins one separated from the salt in the first two-phase system that was formed and then protein a back-extracted to the aqueous phase through formation of second two-phase system. The results show that not only the protein is detected in MS, but the signal intensity is also almost 20-times higher than partial evaporation method.

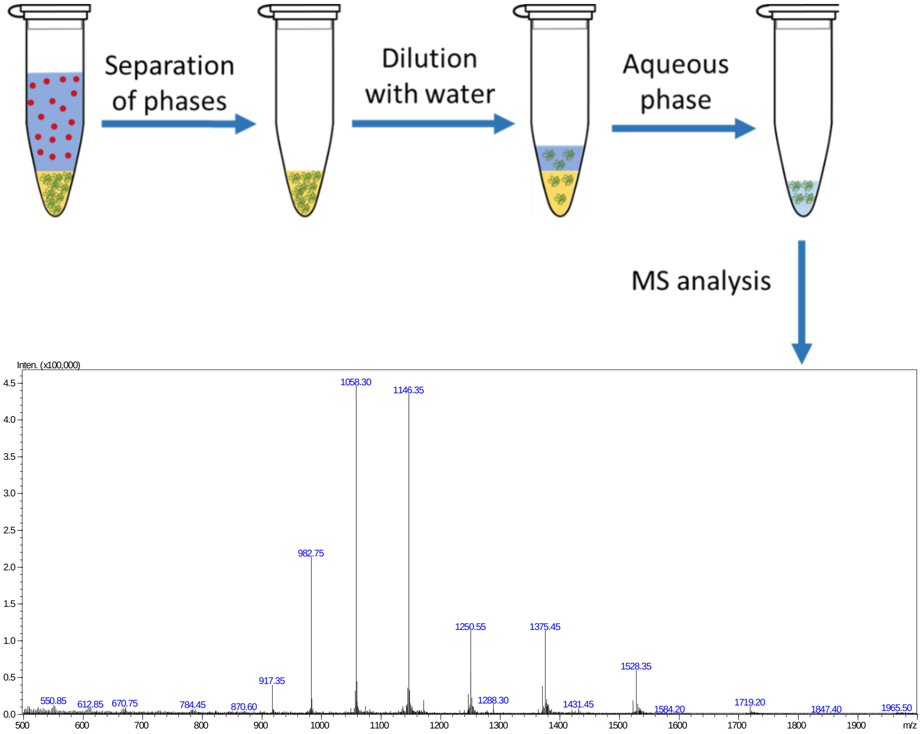


Figure 5-15. Workflow for double extraction. This method is consisting of two extraction steps; in the first extraction proteins are separated form salt and extracted to H-O phase. the H-O phase then diluted with water to form a secondary two-phase system and back-extract the protein to the aqueous phase. Using this method, the result of MS shows the protein is detectable with a good signal intensity.

Figure 5-16 is comparison between partial evaporation and double extraction. There is a big difference between partial extraction signal intensity compared to double extraction.

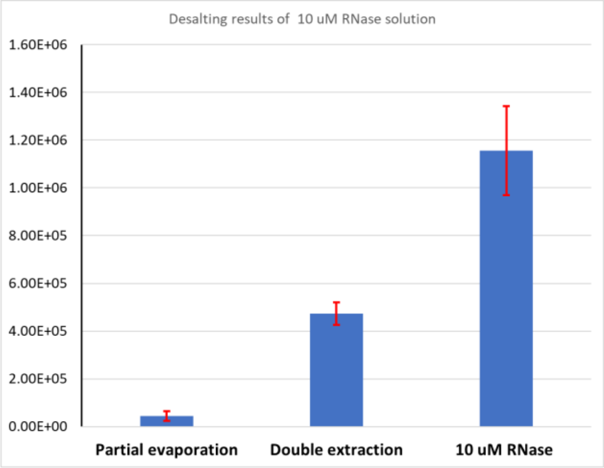


Figure 5-16. comparison between signal intensity of the protein extracted using partial evaporation and double extraction. Double extraction results in much better protein signal.

5.5.6 Optimizing the extraction method

Initially, the protein under study (RNase) was not detectable in the H-O phase after performing desalting process using FOAS. Two solutions were proposed for that problem and results shows that double extraction produces far superior results in detection of protein compared to partial extraction. In this section, the focus is on improving the methodology and workflow to improve protein desalting and extraction even further. One of the parameters in the workflow that is interesting to investigate is the vortex/agitation step that is performed twice in the workflow of double extraction. In all the experiments presented so far, vortex used in the workflow to mix the sample after formation of two-phase. In the double extraction experiment also, vortex used to mix the sample after forming two-phase in both steps of extraction. It is interesting to know whether the intensity of solution mixing can affect protein fractionation between the two phases. Vortex does not agitate the solution; it rather creates a vortex in the microcentrifuge tube and provide very gentle solution mixing. As an alternative to vortex, a more vigorous mixing like agitation via a tissue lyser experimented to find out whether that has effect on protein fractionation or not. After forming the two-phase system instead of vortex, the solution is placed in tissue lyser and shaked with the frequency of 20 Hz. For the the second step of the fractionation the same procedure repeated. In addition to replacing vortex with agitation in both steps of extraction, combination of both vortex and agitation for the first and second step also tested. Figure 5-17 shows all 4 possibilities that rises using these two methods of mixing for two steps of extraction. The time used for either agitation or vortex is the same and it is 1 min.

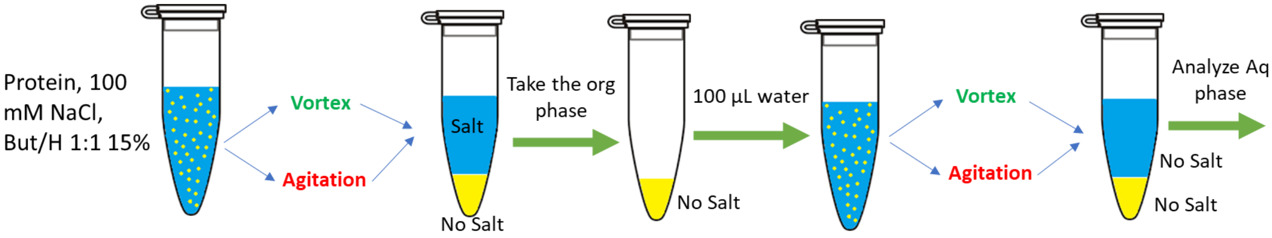


Figure 5-17. Extraction method possibilities in the workflow of double extraction. There are 4 possibilities considering 2 methods of extraction for each step of extraction.

All 4 combinations that is possible, as explained earlier, tested in triplicate. The results show that performing vortex for the first step and agitation for the second step produces far superior results in term of protein's signal intensity in MS compared to all other combinations. The worst results have been produced when the opposite is done; that means instead of vertexing first and agitating second, we agitate first and vortex the second the results are far worst. Other combinations lay in between in this spectrum. Figure 5-18 compares the protein signal intensity after double extraction using vortex and agitation in combination and alone for two steps of protein extraction.

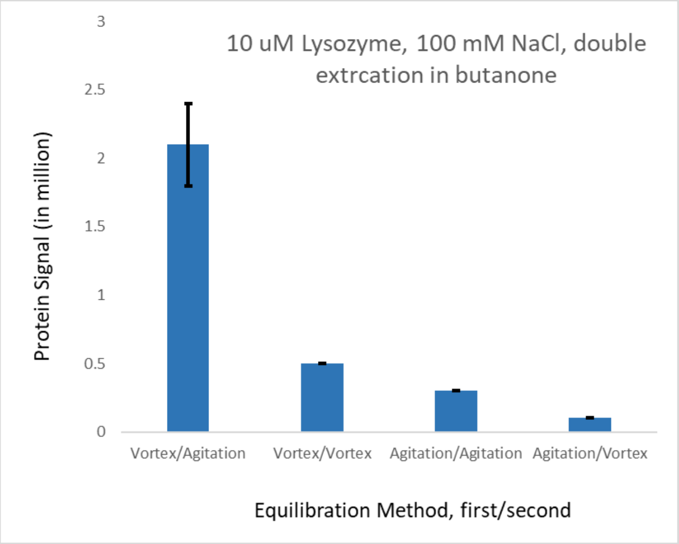


Figure 5-18. All possibilities for protein extraction using two methods of vertexing and agitation for each step of double extraction. Vertexing for the first step and agitation for the second step provides far superior results compared to all other possibilities

The exact reason for this observation is not clear; however, form the intensities for each method, it is obvious that using agitation in the first step of protein extraction does not favor protein extraction to the organic phase since the worst and second worst intensities belong to methods that uses agitation for the first step of extraction. In contrast, vortex for the first step of extraction favors protein extraction to the H-O phase. Keeping vortex as a best method of mixing for the first step of protein extraction which lead to best and second-best results in protein intensities, it is obvious form the results that there is a vast difference between using vortex or agitation for the second step of protein extraction with agitation producing far better results compared to vortex. More vigorous physical force maybe needed for the second step of extraction to back-extract the protein to the aqueous phase. With this optimization, experiments on desalting different proteins from different salts performed which is discussed in the next section.

5.5.7 Desalting different proteins from different salts

In this section results of desalting different proteins sample that salinated with different salts, are presented. Proteins chosen for the experiments are Lysozyme, RNase, and ubiquitin which are salinated with $MgCl_2$, NaCl, NaH_2PO_4, Na_2SO_4 and $(NH_4)_2SO_4$ and desalted using optimized double extraction method. Signal intensities of proteins from each sample after desalting compared to the control sample which has the same protein in water at same concentration of samples but does not contain any salt.

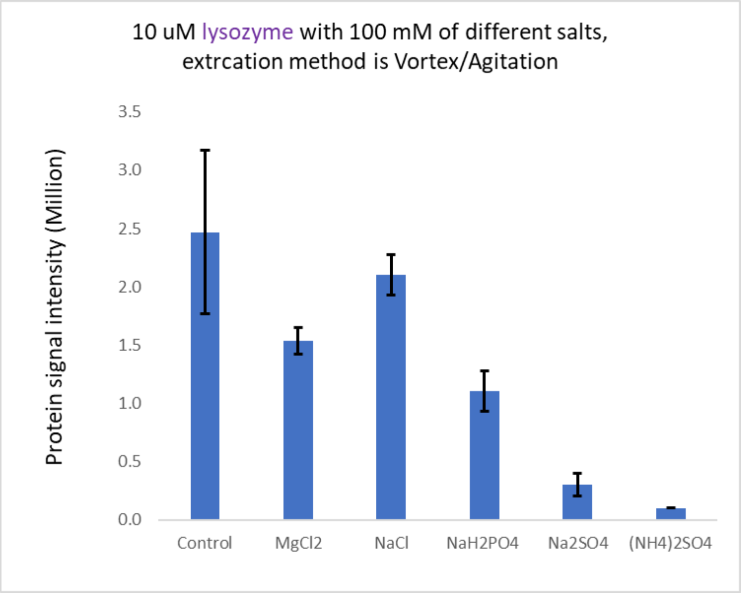
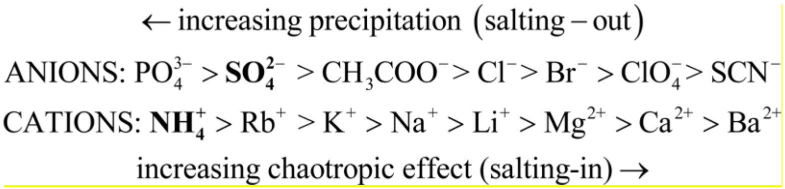


Figure 5-19. Signal intensity of lysozyme after desalting using double extraction compared to control with no salt.

For lysozyme, the results of desalting for $MgCl_2$, NaCl are far better than other salts. For NaCl and $MgCl_2$ sample the signal intensity after desalting the protein is 84% and 60% (respectively)of the protein in the control which is not salinated. Note that the concentration of NaCl in this sample is 100 mM which means before desalting, the protein signal is not detectable and after desalting the signal is almost as strong as the control sample. It is noticeable that salts that contain sulfate as anion, have inferior results compared to other salts. This is because of sulfate characteristic which is known as a good protein salting out agent[12]. Preferential solvation is the main reason for salting out effect in which the layer of water at the vicinity of protein is deprived of salt. Addition of salt to a solution of proteins, increases surface tension of water and as a result hydrophobic interaction between protein and water increases which leads to protein decreasing their size by folding and aggregating. Aggregation causes protein precipitation which in this case is called salting out effect. Capacity of ions in increasing surface tension of water follows Hofmeister series which is demonstrated in Figure 5-19. As it is shows in the Figure 5-19, sulfate and ammonium are amongst the highest in their ability to increase surface tension of water.



$$\leftarrow \text{increasing precipitation } (\text{salting} - \text{out})$$
$$\text{ANIONS: } PO_4^{3-} > \mathbf{SO_4^{2-}} > CH_3COO^- > Cl^- > Br^- > ClO_4^- > SCN^-$$
$$\text{CATIONS: } \mathbf{NH_4^+} > Rb^+ > K^+ > Na^+ > Li^+ > Mg^{2+} > Ca^{2+} > Ba^{2+}$$
$$\text{increasing chaotropic effect (salting-in)} \rightarrow$$

Hofmeister series describes the ability of ions in increasing surface tension of water.

and precipitate out of the solution.

Increasing the surface tension of the water and strengthening hydrophobic interaction between protein and water may drive the protein the the only hydrophobic surface that exist in the microcentrifuge tube which id the inner wall of the microcentrifuge tube. This can cause large amount of protein loss due to non-specific adsorption.

For desalting ubiquitin, samples of 10 µM ubiquitin salinated solution prepare with $MgCl_2$, NaCl, NaH_2PO_4, Na_2SO_4 and $(NH_4)_2SO_4$. In addition, an aqueous solution of 10 µM ubiquitin solution prepared as control. After salinated solutions of ubiquitin prepared, HFIP and butanone added to the solution and the two-phase formed. The solution then vortexed and centrifuged for 1 min. The aqueous phase disposed and the H-O phase which has the volume of ~100 µL, transferred to a fresh microcentrifuge tube and diluted with 100 µL water to form a secondary two phase. the solution then agitated using a tissue lyser for 1 in and then centrifuged. the top phase separated and analyzed in MS.

Figure 5-20 shows the results of analysis of control sample along with desalted ubiquitin samples.

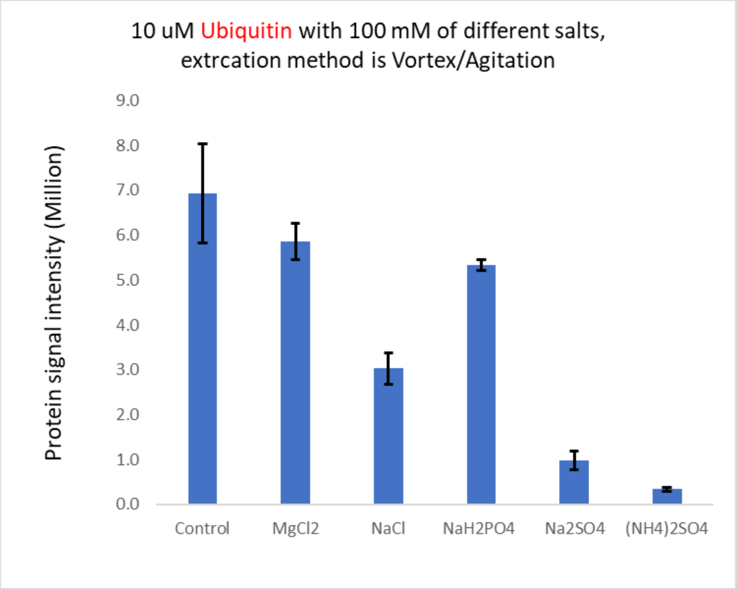


Figure 5-20. Signal intensity of ubiquitin after desalting using double extraction compared to control with no salt.

Among the samples, the desalted sample from $MgCl_2$ shows highest signal intensity of the protein in MS. This sample after desalting has 86% of signal intensity of control sample which does not contain any salt. This is where without desalting the sample with $MgCl_2$ salt in it, there would be no signal for ubiquitin (shown earlier). The second strong signal after $MgCl_2$ sample is the NaH_2PO_4 in which after desalting ubiquitin form a solution that contains 100 mM NaH_2PO_4, the signal intensity of ubiquitin is 60% of the signal intensity of 10 µM ubiquitin in water (no salt). Signal intensity of salinated ubiquitin sample with NaCl is 43% of control sample which is 10 µM ubiquitin in water (no salt); this shows lower intensity compared to signal intensity of salinated Lysozyme with NaCl, which was 60% of control sample after desalting. Like lysozyme experiment, for desalted ubiquitin samples that was salinated Na_2SO_4 and $(NH_4)_2SO_4$, the signal intensity of ubiquitin is lower compared to ubiquitin samples that were salinated with other salts.

This result is consistent with the observation for the lysozyme samples that was salinated with Na_2SO_4 and $(NH_4)_2SO_4$.

For desalting RNase, samples of 10 µM RNase salinated solution prepare with $MgCl_2$, NaCl, NaH_2PO_4, Na_2SO_4 and $(NH_4)_2SO_4$. In addition, an aqueous solution of 10 µM RNase solution prepared as control. After salinated solutions of RNase prepared, HFIP and butanone added to the solution and the two-phase formed. The solution then vortexed and centrifuged for 1 min. The aqueous phase disposed and the H-O phase which has the volume of ~100 µL, transferred to a fresh microcentrifuge tube and diluted with 100 µL water to form a secondary two phase. the solution then agitated using a tissue lyser for 1 in and then centrifuged. the top phase separated and analyzed in MS. Figure 5-21 shows the results of analysis of control sample along with desalted RNase samples.

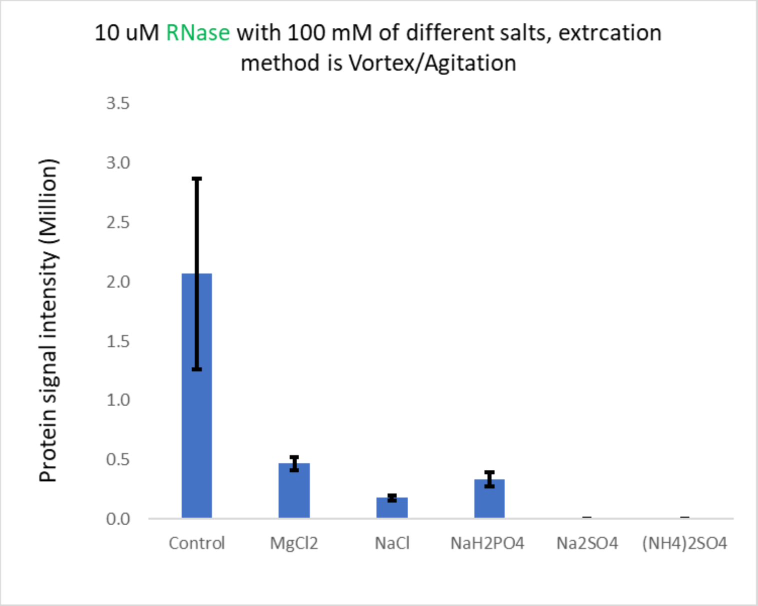
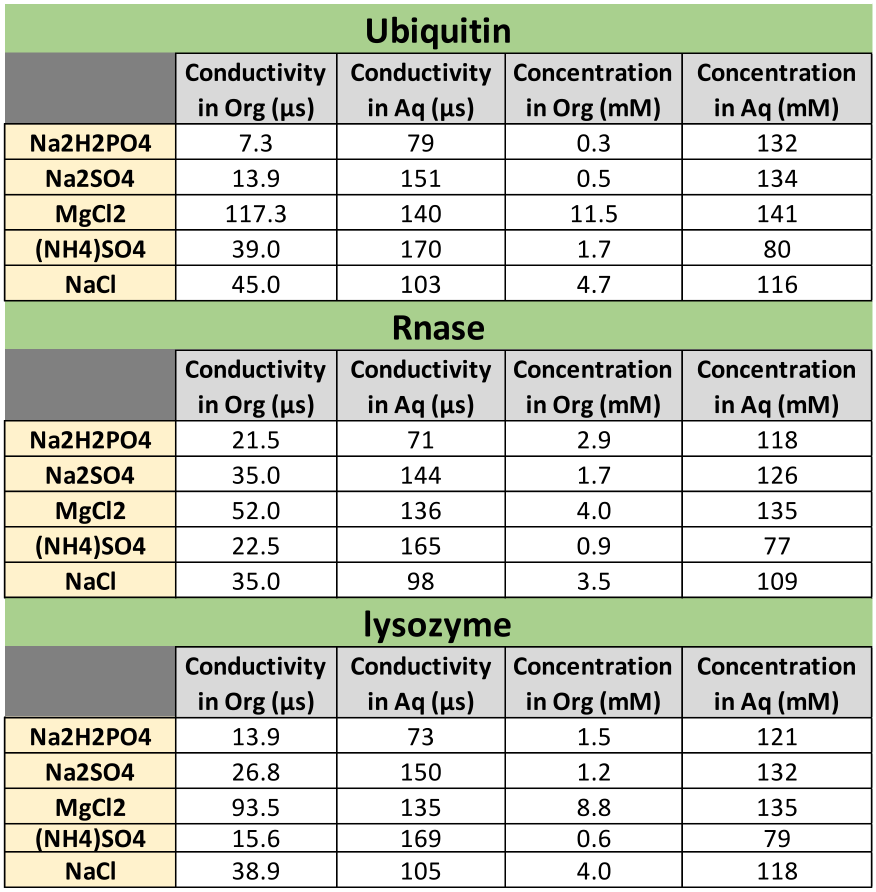


Figure 5-21. signal intensity of RNase after desalting using double extraction compared to control with no salt.

Among the samples, the desalted RNase sample that was salinated with $MgCl_2$ shows the highest signal intensity in MS. This sample after desalting has 24% of signal intensity of control sample which does not contain any salt. This is where without desalting, for the sample with 100 mM $MgCl_2$ salt, there would be no signal for ubiquitin (shown earlier). The second strong signal after $MgCl_2$ sample is the NaH_2PO_4, which is after desalting ubiquitin from salinated ubiquitin solution that contains 100 mM NaH_2PO_4, the signal intensity of ubiquitin is 14% of the control solution signal intensity which is of 10 µM ubiquitin in water (no salt). Signal intensity of salinated ubiquitin solution with NaCl is 10% of control sample which is 10 µM ubiquitin in water (no salt); this shows lower intensity compared to signal intensity for desalted Lysozyme and ubiquitin from salinated solutions of those proteins with NaCl, which was 60% and 43% of control respectively. Unlike desalted lysozyme and ubiquitin which was salinated with Na_2SO_4 and $(NH_4)_2SO_4$, for RNase sample no signal was detected after desalting salinated solutions of RNase with Na_2SO_4 and $(NH_4)_2SO_4$. A possible explanation for this is that at for the same concentration, RNase has lower relative signal intensity compared to lysozyme and ubiquitin. Increasing concentration of RNase in the control and salinated samples can normalize the signal intensity of RNase in MS to provide a better comparison with lysozyme and ubiquitin.

After analysis of all desalted samples in MS, concentration of salt in each phase of all samples that presented earlier measured. Table 6 contains conductivity measurements for aqueous and H-O phase of all salinated solutions of 3 proteins under study. To perform these measurements, 50 µL of each phase transferred to a fresh microcentrifuge tube and dried in Eppendorf Vaccufuge® plus and then both phases reconstituted in water. Aqueous phase diluted 2-times and H-O phase diluted 20-times before performing the conductometry measurements with the in-house made RDE microelectrode.

Table 5-4. Salt concentration measurement using the microelectrode and conductivity measurements. For ubiquitin and RNase samples, for all 5 salts, concentration of salt in the organic phase remains well below 10 mM and concentration of salt in Aqueous phase is above 100 mM, except for ammonium sulfate sample.



Ubiquitin

	Conductivity in Org (μs)	Conductivity in Aq (μs)	Concentration in Org (mM)	Concentration in Aq (mM)
Na2H2PO4	7.3	79	0.3	132
Na2SO4	13.9	151	0.5	134
MgCl2	117.3	140	11.5	141
(NH4)SO4	39.0	170	1.7	80
NaCl	45.0	103	4.7	116

Rnase

	Conductivity in Org (μs)	Conductivity in Aq (μs)	Concentration in Org (mM)	Concentration in Aq (mM)
Na2H2PO4	21.5	71	2.9	118
Na2SO4	35.0	144	1.7	126
MgCl2	52.0	136	4.0	135
(NH4)SO4	22.5	165	0.9	77
NaCl	35.0	98	3.5	109

lysozyme

	Conductivity in Org (μs)	Conductivity in Aq (μs)	Concentration in Org (mM)	Concentration in Aq (mM)
Na2H2PO4	13.9	73	1.5	121
Na2SO4	26.8	150	1.2	132
MgCl2	93.5	135	8.8	135
(NH4)SO4	15.6	169	0.6	79
NaCl	38.9	105	4.0	118

Considering that initial concentration of salts used in all samples are 100 mM, results of salt concentration measurement show that after forming the two-phase, conc of salt in the H-O phase which contains the protein, always remains below 5 mM, with one exception (Ubiquitin salinated with $MgCl_2$) and concentration of slats in aqueous phase in above 100 mM except for protein samples that are salinated with $(NH_4)_2SO_4$.

It is worthy to mention that the after formation of tow-phase, the accuracy of separation of the phases using pipette is very important. The aqueous phase should be removed to the greatest extend; since the concentration of salt is very high in the aqueous phase, a very small residue can heavily contaminate H-O phase. In addition, after removal of aqueous phase it is highly recommended to transfer the H-O phase to a fresh microcentrifuge tube because we observed that usually parts of aqueous phase solution remain on the inner surface of microcentrifuge tube and forms small beads. Processing H-O phase in this condition also causes contamination of H-O phase.

5.6 CONCLUSION

As an alternative to common methods for protein desalting, we introduced a fast and inexpensive method that desalts protein in FOAS. This system has unique abilities that are not exist in other methods. For example, this method is effective, very inexpensive, and fast, does not require expensive lab equipment and it has the capacity to be scaled-up. Unique characteristics makes these methods separate from other methods; salt in the sample extracted with the efficiency of more than 99% and proteins are extracted to the H-O phase almost salt free. We believe aggregation of HFIP, butanone and water, is the key factor in the ability of H-O phase for solubilizing proteins. The fact that 20% of H-O phase is composed of water, makes a big difference between this system and other methods that use organic solvent to precipitate proteins and desalt them. Water in the H-O phase assists solubilization of proteins and aggregation of this with organic solvents pushes the salt to the aqueous phase. 3 proteins have been desalted using tis system form 5 different salts. The results in terms of signal intensity of protein sin MS after desalting is very impressive except for two salts that have sulfate anion in the structure. Sulfate known to be a good agent for protein precipitation due to its ability for increasing the surface tension of water. We believe that salts that have sulfate as anion are interfering with protein solubilization of proteins in the water from the start of the experiment and this will lead to poor signal intensity of proteins after desalting as well.

5.7 REFERENCES

(1) Donnelly, D. P.; Rawlins, C. M.; DeHart, C. J.; Fornelli, L.; Schachner, L. F.; Lin, Z.; Lippens, J. L.; Aluri, K. C.; Sarin, R.; Chen, B.; et al. Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat Methods* **2019**, *16* (7), 587-594. DOI: 10.1038/s41592-019-0457-0.

(2) McPhie, P. Methods in Enzymology. 1971. Jr., W. E. P.; Hudgin, R. L.; Ashwell, G.; Stockert, R. J.; Morell, A. G. Methods in Enzymology. In *A membrane receptor protein for asialoglycoproteins*, 1974; Vol. 34, pp 688-691.

(3) Pohl, T. Concentration of proteins and removal of solutes. *Methods Enzymol* **1990**, *182*, 68-83. DOI: 10.1016/0076-6879(90)82009-q.

(4) Roger L. Hudgin; William E. Pricer; Gilbert Ashwell; Richard J. Stockert; Anatol G. Morell. The Isolation and Properties of a Rabbit Liver Binding Protein Specific for Asialoglycoproteins. Journal of Biological Chemistry: 1974; Vol. 249, pp 5536-5543.

(5) Wessel, D.; Flügge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem* **1984**, *138* (1), 141-143. DOI: 10.1016/0003-2697(84)90782-6.

(6) Porath, J. Molecular Sieving and Adsorption. 1968; Vol. 218.

(7) Hedlund, H. Desalting and buffer exchange of proteins using size-exclusion chromatography. *CSH Protoc* **2006**, *2006* (1). DOI: 10.1101/pdb.prot4199.

(8) Pan, Y. C.; Wideman, J.; Blacher, R.; Chang, M.; Stein, S. Use of high-performance liquid chromatography for preparing samples for microsequencing. *J Chromatogr* **1984**, *297*, 13-19. DOI: 10.1016/s0021-9673(01)89024-5.

(9) Naldrett, M. J.; Zeidler, R.; Wilson, K. E.; Kocourek, A. Concentration and desalting of peptide and protein samples with a newly developed C18 membrane in a microspin column format. *J Biomol Tech* **2005**, *16* (4), 423-428.

(10) Jenö, P.; Scherer, P. E.; Manning-Krieg, U.; Horst, M. Desalting electroeluted proteins with hydrophilic interaction chromatography. *Anal Biochem* **1993**, *215* (2), 292-298. DOI: 10.1006/abio.1993.1589.

(11) Kadjo, A. F.; Stamos, B. N.; Shelor, C. P.; Berg, J. M.; Blount, B. C.; Dasgupta, P. K. Evaluation of Amount of Blood in Dry Blood Spots: Ring-Disk Electrode Conductometry. *Anal Chem* **2016**, *88* (12), 6531-6537. DOI: 10.1021/acs.analchem.6b01280.

(12) Wingfield, P. Protein precipitation using ammonium sulfate. *Curr Protoc Protein Sci* **2001**, *Appendix 3*, Appendix 3F. DOI: 10.1002/0471140864.psa03fs13.

CHAPTER 6

SUMMERY AND PERSPECTIVES

The need for a practical, fast, and inexpensive fractionation method, compelled us to use HFIP-induced two-phase systems in the workflow of yeast bottom-up proteomics for the purpose of improving identification of proteins. FA*i*C systems have been used to fractionate yeast proteome before enzymatic digestion; because of this fractionation, the number of identified proteins with α-helix structure and low abundance proteins significantly improved. We have demonstrated that changes in the ratio between SDC and TBAB can alter the fractionation pattern of proteins based on pI value of the proteins. This has revealed a new way of protein fractionation based on pI without ion chromatography. We suggest that surfactants like SDS to be used as a negatively charged surfactant since it has good solubilizing power for hydrophobic proteins like membrane proteins. Alternation between the ratio of cationic and anionic surfactants is another area that we strongly suggest being investigated.

Like the FA*i*C, FOAS used to fractionate yeast protein digest (peptides). this system unlike FA*i*C does not use surfactants in the composition and that makes it easier to perform since it does not require steps of protein purification that FA*i*C requires. Yeast peptides fractionated in this system and the results are very impressive in the field of enriching peptides with α-helix structures. In addition to that, because of this fractionation, the number of identified proteins with PTMs increased significantly.

Another important and significant use of FOAS is for purification of proteins form salt. In top-down proteomics such as characterization of antibodies, a quick, inexpensive, and effective method for purification of proteins form salts to increase signal intensity in ESI-MS is crucial. We demonstrated the capacity and effectiveness of this system for purification proteins before MS. We have shown that optimization of two-phase system and methods of proteins desalting can have a great impact in the efficiency of the desalting. There is a lot more room for these types of optimizations to further improve the desalting efficiency.

In the FA*i*C and FOAS projects, fractionation of proteins and peptides from secondary structure point of view investigated. To find the secondary structure of proteins for example α-helix structures, we developed a powerful tool to predict such structural information. ypssc is an extension to NetSurfP-2.0 that makes it possible to calculate of secondary structure of proteins form bottom-up proteomics data. This tool instead of calculating secondary structure of yeast peptides in NetSurfP-2.0, uses a database that is provided form NetSurP-2.0 to make a comparison between the sequences that found in the sample and the database to return the structural information about that sequence. At the end the program puts all the sequences form a protein together to provide a coherent picture about the secondary structure of that protein.

APPENDIX 2-1

DESALTING PROCEDURE:

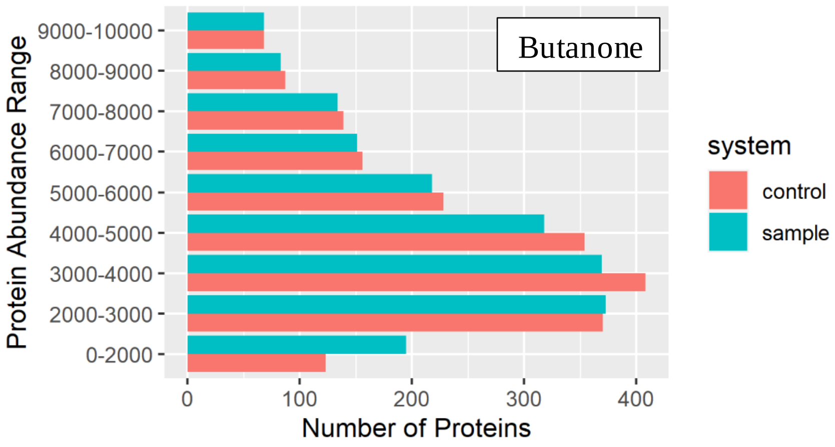
1- Precondition the C18 Sep-Pak columns with 3 mL ACN, 1 mL 0.1 TFA in 75% ACN, 1 mL 0.1 TFA in 50% ACN, 3 mL 0.1 TFA in water.

2- Centrifuge acidified samples at 8000g for 1 min, then load the samples to the columns.

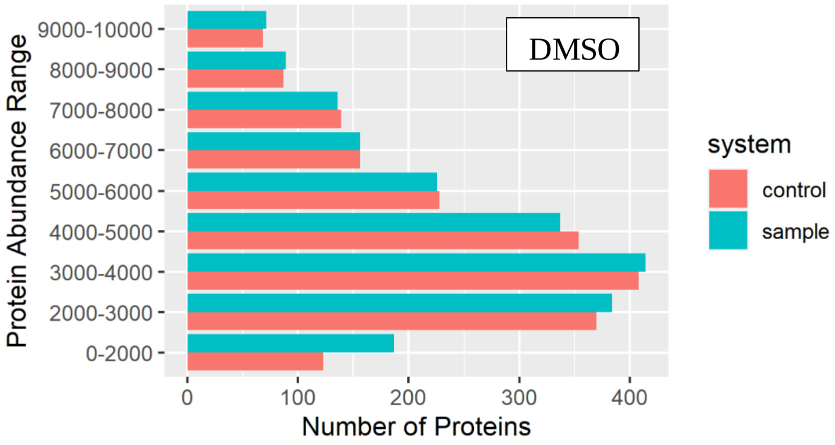
3- Wash the samples with 3 mL 0.1 TFA in water.

4- Move the Sep-Pak to a 2-ml microcentrifuge tube. Elute the sample with 0.6 ml of 0.1 % TFA in 50% ACN, followed by 0.6 ml of 0.1% TFA in 75% This step should be performed by gravity, finally push the samples with pipet.

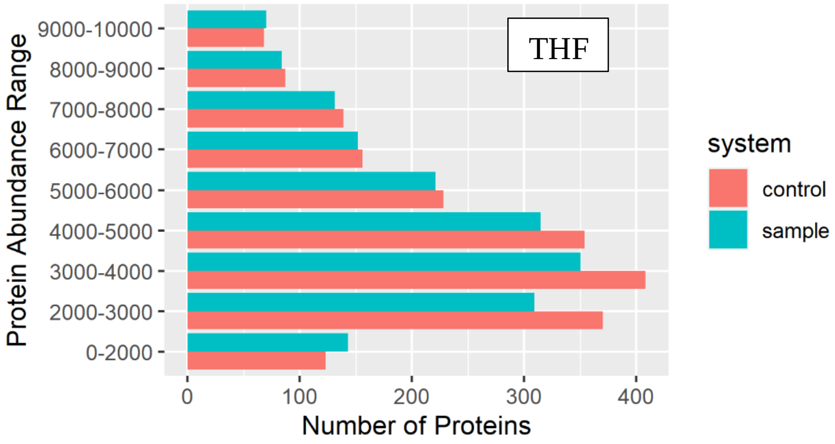
APPENDIX 2-2



Butanone
Protein Abundance Range
9000-10000
8000-9000
7000-8000
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
0-2000
0
100
200
300
400
Number of Proteins
system
control
sample

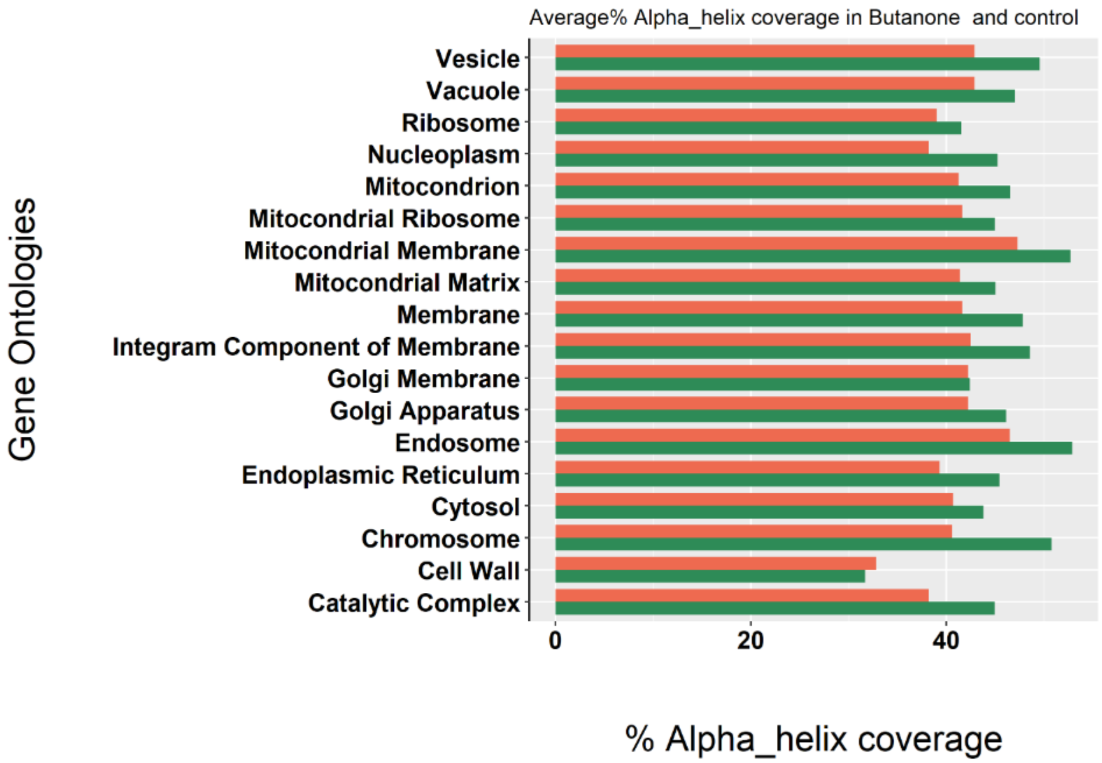


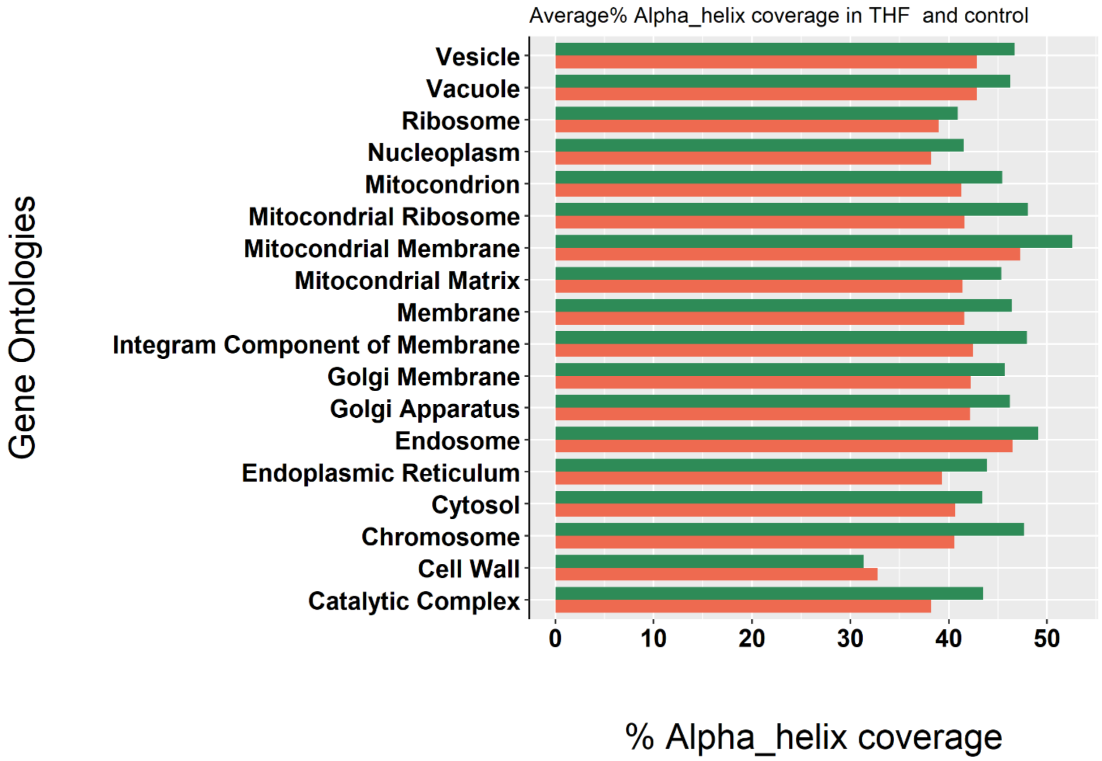
DMSO
Protein Abundance Range
9000-10000
8000-9000
7000-8000
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
0-2000
0
100
200
300
400
Number of Proteins
system
control
sample

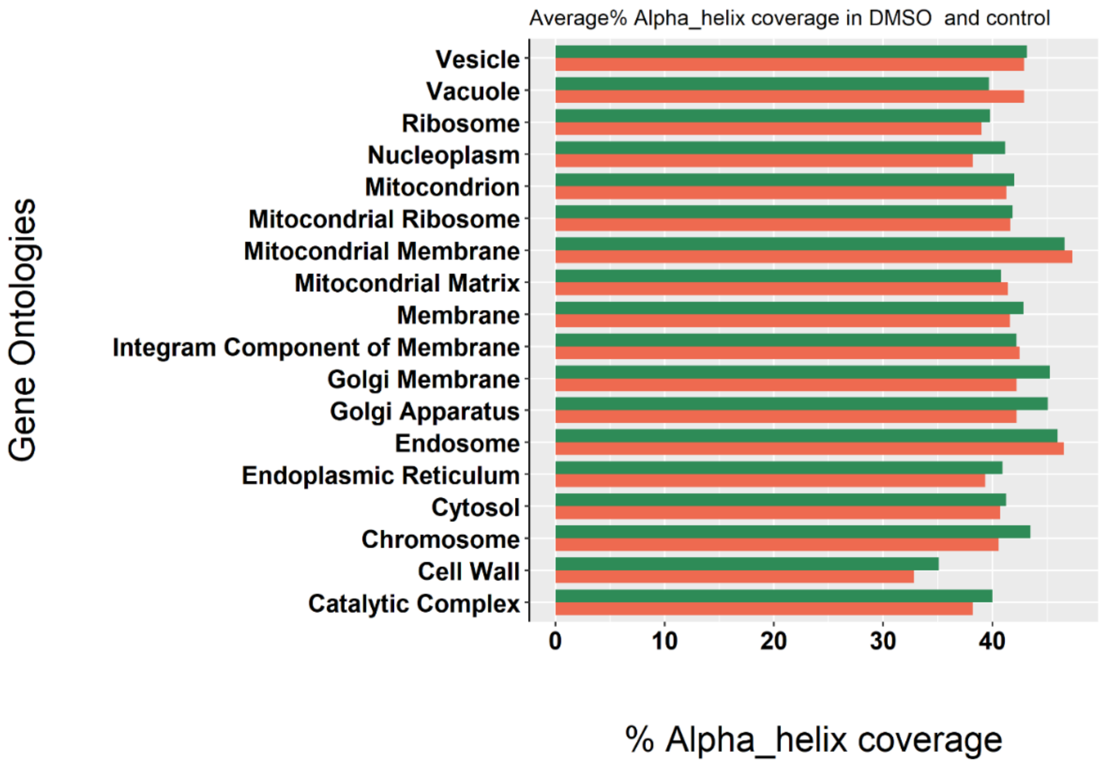


THF
Protein Abundance Range
9000-10000
8000-9000
7000-8000
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
0-2000
0
100
200
300
400
Number of Proteins
system
control
sample

APPENDIX 2-3

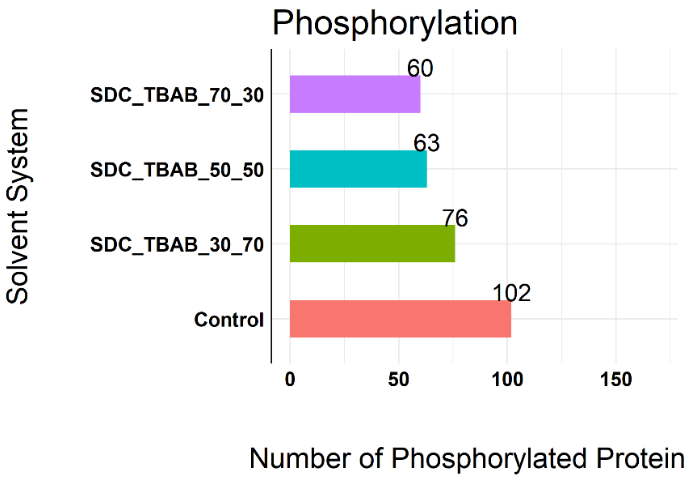

Average% Alpha_helix coverage in Butanone and control
Gene Ontologies
Vesicle
Vacuole
Ribosome
Nucleoplasm
Mitocondrion
Mitocondrial Ribosome
Mitocondrial Membrane
Mitocondrial Matrix
Membrane
Integram Component of Membrane
Golgi Membrane
Golgi Apparatus
Endosome
Endoplasmic Reticulum
Cytosol
Chromosome
Cell Wall
Catalytic Complex
0
20
40
% Alpha_helix coverage

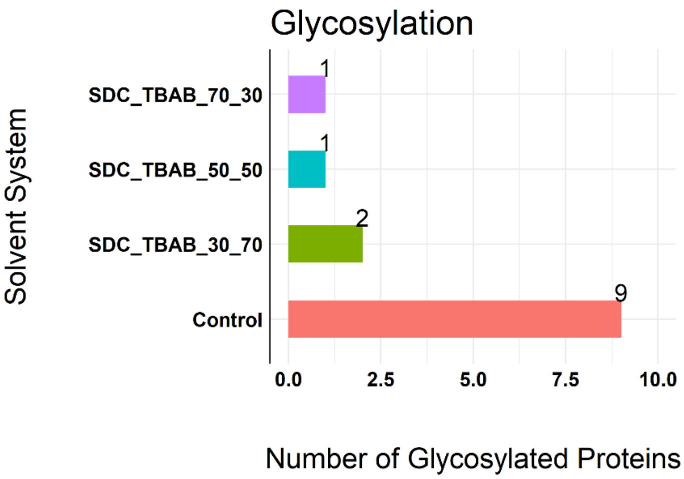

Average% Alpha_helix coverage in THF and control
Gene Ontologies
Vesicle
Vacuole
Ribosome
Nucleoplasm
Mitocondrion
Mitocondrial Ribosome
Mitocondrial Membrane
Mitocondrial Matrix
Membrane
Integram Component of Membrane
Golgi Membrane
Golgi Apparatus
Endosome
Endoplasmic Reticulum
Cytosol
Chromosome
Cell Wall
Catalytic Complex
0
10
20
30
40
50
% Alpha_helix coverage

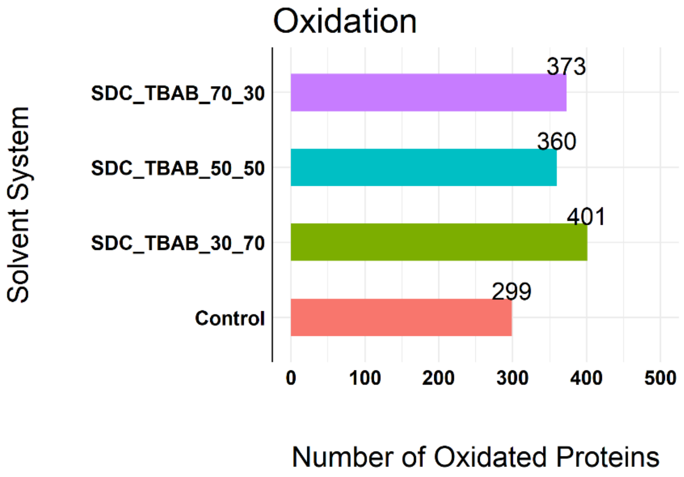

Average% Alpha_helix coverage in DMSO and control
Gene Ontologies
Vesicle
Vacuole
Ribosome
Nucleoplasm
Mitocondrion
Mitocondrial Ribosome
Mitocondrial Membrane
Mitocondrial Matrix
Membrane
Integram Component of Membrane
Golgi Membrane
Golgi Apparatus
Endosome
Endoplasmic Reticulum
Cytosol
Chromosome
Cell Wall
Catalytic Complex
0
10
20
30
40
% Alpha_helix coverage

APPENDIX 3-1

post translational analysis for phosphorylation and glycosylation, does not show any improvement for any of the samples. the only PTM that is showing improvement is oxidation which could occur during sample handling, and it is not super reliable.







APPENDIX 3-2

SOLUTIONS

- 100 mM Trisma buffer, pH= 8.5: add 422 mg Tris HCl and 872 mg Tris base in a 100 mL volumetric flask and bring the volume to 100 mL by adding water.
- 0.5 M NaCl: add 1461 mg NaCl in a 50 mL volumetric flask and bring the volume to 50 mL by adding water.
- 50 mM ABC buffer, pH= 7.8: add 395 mg ammonium bicarbonate in a 100 mL volumetric flask and bring the volume to 100 mL by adding water.
- 100 mM ABC buffer, pH= 7.8: add 395 mg ammonium bicarbonate in a 50 mL volumetric flask and bring the volume to 50 mL by adding water.
- UTT solution: 5 M urea and 2 M thiourea in 100 mM tris buffer (pH=8.5):
  Add 15.015 g urea and 7.612 g thiourea in a 50 mL volumetric flask and bring the volume to 50 mL by adding 100 mM tris buffer (pH=8.5)
- 450 mM TBAB: add 7.253 g tetrabutylammonium bromide in a 50 mL volumetric flask and bring the volume to 50 mL by adding water.

FASP PROTOCOL FOR COACERVATION WITH 8% HFIP AT 50 MM TBAB:

A) COACERVATION

1- Take 400 µg protein (for example, if the concentration of cell lysate is about 8 mg/ml, 50 µl cell lysate is approximately equal to 400 µg)

2- In a 1.6 ml vial, add the following (the total Vol is 1 mL):

  - 400 µg proteins
  - 111 µL 450 mM TBAB
  - DI water: 1000 µL – 111 µL (Vol. TBAB Sol.) – 80 µL (Vol. of HFIP) – Vol. of cell lysate (µL)
  - 80 µL HFIP

3- Centrifuge at 10,000g for 15 min and separate two phases.

Note: measure the protein concentration in each phase before loading the sample to the filters. You will need the amount of protein in each phase when you want to add trypsin with the ratio of 1:25.

Note: For protein concentration measurement, because HFIP has interference with Bradford Assay, evaporate the HFIP prior to the protein measurement (measure the protein concentration in step B-3 for the coacervate phase and at the end of step C-1 for the aqueous phase).

B) FOR THE COACERVATE PHASE:

1- Dry the coacervate with nitrogen gas for about 1 min to get rid of HFIP (not completely dried, adjust the flow of nitrogen to prevent drip).

2- Condition the filter by adding 500 µL UTT solution, then centrifuge at 14,000g for 5 min, 1/3 of the UTT should pass the filter. Again, centrifuge at 14,000g until a thin layer of UTT remains in the filter.

3- Add 450 µL 70% IPA to the coacervate, then add 76 mg thiourea to dissolve it. Vortex 30 sec, then sonicate 5 min at room temperature (not in ice, because it does not dissolve at low temperatures), and load the dissolved coacervate to the to the pre-conditioned FASP filters. Take a small amount of this solution for protein concentration measurement.

4- Centrifuge at 14,000g for 40 min (If necessary, centrifuge again at 14,000g, until the volume reaches to about 20 µL)

5- Add 200 µL 70% IPA, centrifuge at 14,000g for 40 min (or more than 40 min until about 20 µL of sample remains in the filter)

Note: each time you add a solution, mix it up and down by pipet: if you want to add 200 µL, first add 100 µL, mix it with pipet, and then wash the same pipet with another 100 µL solution.

6- Add another 200 µL 70% IPA, centrifuge at 14,000g for 40 min (or more than 40 min until about 20 µL of sample remains in the filter)

7- Add 200 µL UTT solution (UTT solution is 5 M urea, 2 M thiourea, in 0.1 M tris buffer, pH= 8.5), using a pipet break the precipitate until you see a uniform liquid, centrifuge at 14,000g for 40 min

Meanwhile doing B, do the part C

C) FOR THE AQUEOUS:

1- Put the Aqueous phase in concentrator to evaporate HFIP for about 1.5 hours, until the volume reaches to about 500 µL. At this point, majority of HFIP is evaporated (BP: 58 °C) and you can measure protein concentration.

2- Condition the filter by adding 500 µL UTT solution, then centrifuge at 14,000g for 5 min, 1/3 of the UTT should pass the filter. Again, centrifuge at 14,000g until a thin layer of UTT remains in the filter.

3- Load the concentrated aqueous to FASP filter and centrifuge at 14,000g for 40 min (or more than 40 min until about 20 µL of sample remains in the filter)

D) FOR BOTH AQUEOUS AND COACERVATE:

1- In a 1.6 mL vial, add 39 mg DDT to 1mL of UTT buffer to make 250 mM DTT in UTT buffer.

2- Add 20 µL 250mM DTT to each sample and bring the Vol. to 200 µL to bring the concentration of DTT to 25 mM with UTT solution (for example if the thin layer is 20 µL, add 20 µL 250 mM DTT and then add 160 µL UTT solution). Using a pipet mix it, vortex for 30 sec at 600 rpm, incubate at 37 ℃ for 45 min.

3- Cool the sample to room temperature.

4- Centrifuge at 14,000g for 40 min (or more than 40 min until about 20 µL of sample remains in the filter)

5- Make a stock solution of 250 mM IAA in UTT buffer (add 44 mg IAA and 1 mL UTT buffer) in darkness. Then dilute it to 54 mM (mix 216 µL of 250 mM with 784 µL UTT buffer to make a 54 mM IAA in UTT buffer).

6- Add 200 µL IAA 54 mM, then wash the same pipet with 50 µL IAA 54 mM (concentration of IAA must be 50 mM, so if the final volume is 270 µL, and the concentration would be 50 mM). vortex for 30 sec at 600 rmp and incubate at dark for 45 min. IAA should be made and added in a dark room.

7- Centrifuge at 14,000 g for 40 min.

8- Add 200 µL UTT solution and centrifuge 14,000 g for 40 min.

9- Add 200 µL ABC (50 mM) and centrifuge 14,000 g for 40 min.

10- Add 200 µL ABC (50 mM) and centrifuge 14,000 g for 40 min.

11- Add 150 µL ABC 100 mM (because trypsin as acidic and we want to bring the pH to 7). Then, add trypsin with the ratio of 1:25

(in the case of our study, for the aqueous: (250/25)*2= 20 µL trypsin; and for coacervate: (150/25)*2= 12 µL trypsin. These values may vary in other studies because the cell lysate is different.)

12- Check the pH to be around 7. Then seal the vials with parafilm. Shake at 600 rpm for 1 min. Incubate in wet chamber at 37 °C for 16 hrs.

13- Transfer the filters to new collection tubes.

14- centrifuge at 14,000 g for 40 min.

15- Add 200 µL 0.5 M NaCl, and centrifuge at 14,000 g for 40 min in the collection tube.

16- Add 200 µL 50 mM NaCl, invert the filter, and centrifuge at 1000 g for 2 min in the collection tube.

17- Acidify the sample with TFA to bring the pH below 2 (about 2-3 µL TFA is enough, DO NOT over-acidify)

18- Desalt the samples.

CONTROL:

1- Condition the filter by adding 500 µL UTT solution, then centrifuge at 14,000g for 5 min, 1/3 of the UTT buffer should pass the filter. Again, centrifuge at 14,000g until a thin layer of UTT remains in the filter.

2- Dissolve the protein in 5M urea and 2 M thiourea, sonicate for 5 minutes and load the sample to the to the FASP filters.

3- Centrifuge at 14,000g for 40 min (If necessary, centrifuge again at 14,000g, until the volume reaches to about 20 µL)

4- If you see cell debris, add 200 µL 70% IPA, centrifuge at 14,000g for 40 min, otherwise, go to step 6.

(or more than 40 min until about 20 µL of sample remains in the filter) (each time you add a solution, mix it up and down by pipet: if you want to add 200 µL, first add

100 µL, mix it with pipet, and then wash the same pipet with another 100 µL solution)

5- If you see cell debris, add another 200 µL 70% IPA, centrifuge at 14,000g for 40 min (or more than 40 min until about 20 µL of sample remains in the filter)

6- Add 200 µL UTT solution (UTT solution is 5 M urea, 2 M thiourea, in 0.1 M tris buffer, pH= 8.5), using a pipet break the precipitate until you see a uniform liquid, centrifuge at 14,000g for 40 min

In a 1.6 mL vial, add 39 mg DDT to 1mL of UTT buffer to make 250 mM DTT in UTT buffer.

7- Add 20 µL 250mM DTT to each sample and bring the Vol. to 200 µL to bring the concentration of DTT to 25 mM with UTT solution (for example if the thin layer is 20 µL, add 20 µL 250 mM DTT and then add 160 µL UTT solution). Using a pipet mix it, vortex for 30 sec at 600 rpm, incubate at 37 ℃ for 45 min.

8- Cool the sample to room temperature.

9- Centrifuge at 14,000g for 40 min (or more than 40 min until about 20 µL of sample remains in the filter)

10- Make a stock solution of 250 mM IAA in UTT buffer (add 44 mg IAA and 1 mL UTT buffer) in darkness. Then dilute it to 54 mM. (mix 216 µL of 250 mM with 784 µL UTT buffer to make a 54 mM IAA in UTT buffer)

11- Add 200 µL IAA 54 mM, then wash the same pipet with 50 µL IAA 54 mM (concentration of IAA must be 50 mM, so if the final volume is 270 µL, and the concentration would be 50 mM). vortex for 30 sec at 600 rmp and incubate at dark for 45 min. IAA should be made and added in a dark room.

12- Centrifuge at 14,000 g for 40 min.

13- Add 200 µL UTT buffer solution and centrifuge 14,000 g for 40 min.

14- Add 200 µL ABC (50 mM) and centrifuge 14,000 g for 40 min.

15- Add 200 µL ABC (50 mM) and centrifuge 14,000 g for 40 min.

16- Add 150 µL ABC 100 mM (because trypsin as acidic and we want to bring the pH to 7). Then, add trypsin with the ratio of 1:25

(for Aq: (250/25)*2= 20 µL trypsin; and for coacervate: (150/25)*2= 12 µL trypsin)

17- Check the pH to be around 7. Then seal the vials with parafilm. Shake at 600 rpm for 1 min. Incubate in wet chamber at 37 °C for 16 hrs.

18- Transfer the filters to new collection tubes.

19- centrifuge at 14,000 g for 40 min.

20- Add 200 µL 0.5 M NaCl, and centrifuge at 14,000 g for 40 min in the collection tube.

21- Add 200 µL 50 mM NaCl, invert the filter, and centrifuge at 1000 g for 2 min in the collection tube.

22- Acidify the sample with TFA to bring the pH below 2 (about 5 µL TFA is enough, DO NOT over-acidify)

23- Desalt the samples.

DESALTING PROCEDURE:

5- Precondition the C18 Sep-Pak columns with 3 mL ACN, 1 mL 0.1 TFA in 75% ACN, 1 mL 0.1 TFA in 50% ACN, 3 mL 0.1 TFA in water.

6- Centrifuge acidified samples at 8000g for 1 min, then load the samples to the columns.

7- Wash the samples with 3 mL 0.1 TFA in water.

8- Move the Sep-Pak to a 2-ml microcentrifuge tube. Elute the sample with 0.6 ml of 0.1 % TFA in 50% ACN, followed by 0.6 ml of 0.1% TFA in 75% This step should be performed by gravity, finally push the samples with pipet.

APPENDIX 4-1

```
library(tidyverse)
library(readxl)
library(stringr)
library(eulerr)
library(ggplot2)
setwd('C:/Users/tasharofis/Documents)
df<-read.csv("proteinGroups.csv")
df1<-read_excel("SGD GRAVY pI MW Database.xlsx")
df2<-read_excel("Abundance database-R.xlsx")
GOs<-read_excel("GOs.xlsx")
df<-filter(df, !grepl(';',Majority.protein.IDs ))
df<-filter(df, !grepl('\\+',Only.identified.by.site ))
df<-filter(df, !grepl('\\+',Reverse ))
df<-filter(df, !grepl('\\+',Potential.contaminant ))
# keep the Majority.protein.IDs, iBAQs, and that is it ####
df[c(1,3:151,179:191)]<-NULL
df<-filter(df, iBAQ>0)
df4<-left_join(df1,df2,by="SGD")
df5<-left_join(df,df4,by="Majority.protein.IDs")
#reading the control file separately####
df3<-read.csv('ProteinGroups_control.csv')
df3<-filter(df3, !grepl(';',Majority.protein.IDs ))
df3<-filter(df3, !grepl('\\+',Only.identified.by.site ))
```

```
df3<-filter(df3, !grepl('\\+',Reverse ))

df3<-filter(df3, !grepl('\\+',Potential.contaminant ))

names(df)

# keep the Majority.protein.IDs, iBAQs, and that is it ####

df3[c(1,3:86,101:113)]<-NULL

df3<-filter(df3, iBAQ>0)

df6<-left_join(df3,df4,by="Majority.protein.IDs")

names(df6)

setwd('C:/Users/tasharofis/Documents/Mehdi/Mehdi_protein analysis_program and feed/Protein analysis_multiple controls/results')

################### controls ####

controls<-select(df6,c(1,10:12))

control1<-filter(controls, iBAQ.CT.W.FASP_1>0)

proteinListControl1<-select(control1, 1)

write.csv(proteinListControl1, file="Protein list of control 1.csv",row.names = FALSE)

control2<-filter(controls, iBAQ.CT.W.FASP_2>0)

proteinListControl2<-select(control2, 1)

write.csv(proteinListControl2, file="Protein list of control 2.csv",row.names = FALSE)

control3<-filter(controls, iBAQ.CT.W.FASP_3>0)

proteinListControl3<-select(control3, 1)

write.csv(proteinListControl3, file="Protein list of control 3.csv",row.names = FALSE)

#df7<-df6[which(rowSums(df6) >1),]

controls_2of3<-controls[rowSums(controls==0) <=1,]

write.csv(controls_2of3, file="2 of 3 in controls.csv",row.names = FALSE)

########## c ontrols euler diagram ###
```

```
eulercontrol<-euler(c("Control 1"=proteinListControl1,
                     "Control 2"=proteinListControl2,
                     "Control 3"=proteinListControl3))
tiff("euler for controls.tiff",units="px", width=700, height=700, res=300)
#pdf('euler for controls.pdf')
plot(eulercontrol,
     edges=FALSE,
     quantities = list(type=c("counts","percent"),
                      fontsize=8),
     fills= list(fill=c("yellow","red","lightblue"),alpha=1),
     #labels = list(labels = c("CT 1", "CT 2", "CT 3"),
                  #fontsize = 24),
     legend=list(labels=c("Control 1","Control 2","Control 3"),
                fontsize=9,
                side="bottom"))
dev.off()
########## controls low abundance proteins ###
controls_2of3<-left_join(controls_2of3,df4,by="Majority.protein.IDs" )
abd_0_2000<-nrow(controls_2of3%>% filter(abd>=0 & abd<=2000))
abd_2000_3000<-nrow(controls_2of3%>% filter(abd>=2000 & abd<=3000))
abd_3000_4000<-nrow(controls_2of3%>% filter(abd>=3000 & abd<=4000))
abd_4000_5000<-nrow(controls_2of3%>% filter(abd>=4000 & abd<=5000))
abd_5000_6000<-nrow(controls_2of3%>% filter(abd>=5000 & abd<=6000))
abd_6000_7000<-nrow(controls_2of3%>% filter(abd>=6000 & abd<=7000))
```

```
abd_7000_8000<-nrow(controls_2of3%>% filter(abd>=7000 & abd<=8000))
abd_8000_9000<-nrow(controls_2of3%>% filter(abd>=8000 & abd<=9000))
abd_9000_10000<-nrow(controls_2of3%>% filter(abd>=9000 & abd<=10000))
################### Sample ####
names(df5)
x<-11
for (i in 1:8){
  spl_names<-
c('DTAB_90mM','TBAB_1M','TBAB_90mM','TBAB_HFMIP_1M','TBAB_HFMIP','TBAB_TF
E_1M',"TEAB_1M","TEAB_90mM")
  x<-x+2
  y<-x+1
spl<-select(df5,c(1,x:y,29:39))
spl_A<-filter(spl, spl[,2]>0)
write.csv(spl_A, file=paste("sample Aq",i,".csv"),row.names = FALSE)
spl_O<-filter(spl, spl[,3]>0)
write.csv(spl_O, file=paste("sample Org",i,".csv"),row.names = FALSE)
spl_total<-union(spl_A,spl_O)
write.csv(spl_total, file=paste("sample Aq & Org",i,".csv"),row.names = FALSE)
########## Sample euler diagram ###
a<-spl_A[,1]
a<-select(spl_A,1)
b<-select(spl_O,1)
spl_euler<-euler(c("a"=a,"b"=b))
```

```
tiff(paste("sample Aq-Org euler",spl_names[i],".tiff"), units="px", width=700, height=700,
res=300)

#pdf('euler for controls.pdf')

plot<-plot(spl_euler,

    edges=FALSE,

    quantities = list(type=c("counts","percent"),

                      fontsize=8),

    fills= list(fill=c("lightblue","yellow"),alpha=1),

    #labels = list(labels = c("CT 1", "CT 2", "CT 3"),

                   #fontsize = 24),

    legend=list(labels=c("Aqueous Phase","Organic Phase"),

              cex=0.6,

              side="bottom"),

    main=list(label=paste(spl_names[i]),cex=0.7)

    )

print(plot)

dev.off()

########## ST158 low abundance proteins bar chart ###

spl_0_2_K<-nrow(spl_total%>% filter(abd>=0 & abd<=2000))

spl_2_3_K<-nrow(spl_total%>% filter(abd>=2000 & abd<=3000))

spl_3_4_K<-nrow(spl_total%>% filter(abd>=3000 & abd<=4000))

spl_4_5_K<-nrow(spl_total%>% filter(abd>=4000 & abd<=5000))

spl_5_6_K<-nrow(spl_total%>% filter(abd>=5000 & abd<=6000))

spl_6_7_K<-nrow(spl_total%>% filter(abd>=6000 & abd<=7000))

spl_7_8_K<-nrow(spl_total%>% filter(abd>=7000 & abd<=8000))
```

```
spl_8_9_K<-nrow(spl_total%>% filter(abd>=8000 & abd<=9000))
spl_9_10_K<-nrow(spl_total%>% filter(abd>=9000 & abd<=10000))
control_abd<-data.frame(
  range=c("0-2000","2000-3000","3000-4000","4000-5000","5000-6000",
          "6000-7000","7000-8000","8000-9000","9000-10000"),
  value=c(abd_0_2000,abd_2000_3000,abd_3000_4000,abd_4000_5000,
          abd_5000_6000,abd_6000_7000,abd_7000_8000,abd_8000_9000,abd_9000_10000),
  system="control"
)
spl_abd<-data.frame(
  range=c("0-2000","2000-3000","3000-4000","4000-5000","5000-6000",
          "6000-7000","7000-8000","8000-9000","9000-10000"),
  value=c(spl_0_2_K,spl_2_3_K,spl_3_4_K,
          spl_4_5_K,spl_5_6_K,spl_6_7_K,
          spl_7_8_K,spl_8_9_K,spl_9_10_K),
  system='sample'
)
data_full_<-full_join(control_abd,spl_abd,by=c('range','value','system'))
tiff(paste("sample abd bar",spl_names[i],".tiff"), units="px", width=1500, height=800, res=300)
bar_plot<-ggplot(data_full_,aes(x=range,y=value,fill=system))+
  geom_bar(stat="identity",position="dodge")+
  coord_flip()+
  xlab('Protein Abundance Range')+
  ylab('Number of Proteins')
```

```
print(bar_plot)
dev.off()
########## ST158 low abundance proteins bar chart in percentage ###
diff<-(spl_abd$value-control_abd$value)/control_abd$value*100
perc<-data.frame(
  range=c("0-2000","2000-3000","3000-4000","4000-5000","5000-6000",
          "6000-7000","7000-8000","8000-9000","9000-10000"),
  diff)
tiff(paste("sample abd bar prc",spl_names[i],".tiff"), units="px", width=1500, height=800,
res=300)
bar_plot_prc<-ggplot(perc,aes(x=range,y=diff))+
  geom_bar(stat="identity",position="dodge",fill="darkGreen")+
  coord_flip()+
  xlab('Protein Abundance Range')+
  ylab('% Improvement')
print(bar_plot_prc)
dev.off()
########## Sample euler diagram for GO ####
GO_names<-c('Catalytic Complex','Cell Wall','Chromosome','Cytosol','Endoplasmic Reticulum',
            'Endosome','Golgi Apparatus','Golgi Membrane','Integram Component of Membrane',
            'Membrane','Mitocondrial Matrix','Mitocondrial Membrane','Mitocondrial Ribosome',
            'Mitocondrion','Nucleoplasm','Ribosome','Vacuole','Vesicle')
j<-1
a<--1
for(j in 1:18){
  a<-a+2

  GO<-select(GOs,a)

  names(GO)<-"Majority.protein.IDs"

  d<-select(spl_A,1)

  e<-select(spl_O,1)

  spl_A_GO<-intersect(d,GO)

  spl_O_GO<-intersect(e,GO)

  spl_euler<-euler(c("a"=spl_A_GO,"b"=spl_O_GO))

  tiff(paste("euler_GO_",spl_names[i],"_",GO_names[j],".tiff"), units="px", width=700,
height=700, res=300)

  #pdf('euler for controls.pdf')

  plot<-plot(spl_euler,

            edges=FALSE,

            quantities = list(type=c("counts","percent"),

                              fontsize=8),

            fills= list(fill=c("lightblue","yellow"),alpha=1),

            #labels = list(labels = c("CT 1", "CT 2", "CT 3"),

            #fontsize = 24),

            legend=list(labels=c("Aqueous Phase","Organic Phase"),

                        cex=0.6,

                        side="bottom"),

            main=list(label=paste(spl_names[i],'_',GO_names[j]),cex=0.7)

  )

  print(plot)

  dev.off()
```

```
}

}
```