LEVERAGING CAPILLARY ZONE ELECTROPHORESIS-MASS SPECTROMETRY FOR MULTI-LEVEL PROTEOMICS

By

Xiaojing Shen

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Chemistry - Doctor of Philosophy

2020

ABSTRACT

LEVERAGING CAPILLARY ZONE ELECTROPHORESIS-MASS SPECTROMETRY FOR MULTI-LEVEL PROTEOMICS

By

Xiaojing Shen

Mass spectrometry (MS) coupled with online liquid-phase separation is the major tool for large-scale bottom-up proteomics (peptide-centric), top-down proteomics (proteoform-centric), and native proteomics (protein complex-centric). While liquid chromatography (LC)-MS is the dominant method for proteomics at different levels, capillary zone electrophoresis (CZE)-MS has emerged as a valuable and complementary technique, which provides high-capacity separation and highly sensitive detection of peptides, proteoforms and even protein complexes under native conditions. This work focuses on developing novel CZE-MS/MS methods for multi-level proteomics (bottom-up, top-down, and native).

In Chapter 2, a high-throughput bottom-up proteomics workflow was developed by coupling immobilized trypsin-based speedy protein digestion with fast CZE-MS/MS. Immobilized trypsin produced almost the same digestion performance as free trypsin for complex proteomes with about 50-times higher speed (15 min *vs.* 12 h). Integration of immobilized trypsin (IM)-based rapid protein cleavage and fast CZE-MS/MS enables the identification of thousands of proteins from the mouse brain proteome in only 3 h, which is significantly faster than the typical LC-MS-based bottom-up proteomics workflow (3 h *vs.* >12 h). The high-throughput workflow was expected to be useful for bottom-up proteomics of human clinical samples (*e.g.,* serum and urine).

Chapter 3 presents the first example of CZE-MS/MS with activated ion-electron capture dissociation (AI-ECD) on a high-end quadrupole-time-of-flight (Q-TOF) mass spectrometer for top-down proteomics, enabling high-resolution separation, highly sensitive detection, and extensive gas-phase backbone cleavages of proteoforms. The CZE-AI-ECD method will be useful to the top-down proteomics community for the comprehensive characterization of proteoforms in complex proteomes.

Chapter 4 and 5 focus on the development of novel CZE-MS methods for native proteomics, delineating proteins and protein complexes under native conditions. In Chapter 4, a native CZE-MS/MS platform with an Orbitrap mass spectrometer was established for native proteomics of a complex proteome (*E. coli)*, leading to the identification of 23 protein complexes in discovery mode. The work represents the first example of native proteomics via coupling online liquid-phase separation to native MS and MS/MS. The characterization of large protein complexes (up to 200 kDa) was also achieved with a new CZE-MS system on a high-end Q-TOF mass spectrometer.

In Chapter 5, a novel native capillary isoelectric focusing (cIEF)-assisted CZE-MS method is presented for the characterization of monoclonal antibodies (mAbs) with large sample loading capacity and high separation resolution. Using the method, the potential separations of different conformations of the SigmaMAb and the detection of its various glyco-proteoforms and homodimer were documented. The method separated the NISTmAb into three peaks with a microliter sample loading volume, corresponding to its different proteoforms. In addition, eight glyco-proteoforms of the NISTmAb and its homodimer were detected. The results demonstrate the potential of the native cIEF-assisted CZE-MS method for advancing the characterization of large proteins (*i.e.*, mAbs) and protein complexes under native conditions.

This thesis is dedicated to my grandparents, my parents, my husband, and my cats.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Professor Liangliang Sun. I am grateful that he accepted me to join in his group when I had to transfer to a new group in my second year of the PhD program. I had no experience on mass spectrometry and proteomics before, and he spent some time every day to teach me the basic knowledge of mass spec and guide me to the proteomics field. Those one-to-one tutoring sessions built the solid base for my future research. I am also very glad to be his first student. I still remember the time when there were only him and me in the lab, we together cleaned up the temporary lab, placed orders, open packages like open the Christmas gifts and fixed the LTQ mass spec, although it was mainly him fixing the instrument and I was holding the flashlight and trying to memorize the name of each component. I got so many unique and valuable memories that I do not have enough space to write everything here, but I will remember them for my life. I also want to thank him for always being patient and would like to listen to me and discuss with me when I met troubles in my experiments or when we have different opinions. His support got me through thousands of failures on my projects so that I did not give up on the halfway and finally arrived here. I am proud to be part of his group and I sincerely wish the Sun group will become better and better.

I would like to thank Professor Dana Spence for his guidance in my research. He is a helpful and friendly teacher and showed me how to communicate as a scientist. I have to admit that I sometimes felt pressure from him, but those pressure pushed me become a qualified PhD. I would like to thank Professor Dan Jones for his help and advice for my research, my seminar as well as my job seeking, which is helpful not only for my experiments but also for my future career development. I would like to thank Professor Sophia Lunt for her help on my experiments and my application for scholarship. I would also like to thank my former advisor Professor

Merlin Bruening. Although I only worked with him for one year, he was no doubt a good advisor. He also helped a lot during my transferring to the current group.

I would like thank my collaborators, Professor Heedeok Hong, Professor Xuefei Huang and Professor Chen Chen from Michigan State University, Professor Xiaowen Liu from IUPUI, Professor Wenjun Du from Central Michigan University, Dr. James Xia from CMP Scientific, Dr. Joseph Beckman, Dr. Valery Voinov, Blake Hakkila, and Mike Hare from e-MSion, and Dr. David Wong and John Sausen from Agilent Technologies, for their help in my research projects.

I would like to thank my group members. Rachele Lubeckyj and Eli McCool helped me better adapter the culture here. Daoyang Chen and Zhichang Yang provided valuable advice on my experiments. Qianjie Wang, Tian Xu and Qianyi Wang helped a lot on my projects. Qianjie and Tian also help me survive the last few months when I was living alone for the first time. I really thank them for their help in my work and also being great friends.

I would like to thank Shuang, Weijing, Ke, Wenjing, Yongle, Ruiqiong, Yiqing, Yijing, Xiaopeng, Yi, Jiaqi, Zhilin, Chenjia and all other friends for their accompanying in the past five and half years. I would also like to thank Professor Wulff for his wonderful wine and cheese and all the friends who shared those great parties with me.

Finally, I would like to give a special gratitude to my husband Dr. Li Zheng. He is one of the major reasons that made me decide to pursue the PhD degree at Michigan State University, which turns out to be a great decision and let me have so many great experiences and memories. He is also my biggest supporter who helps me overcome frustration and homesickness and is always the first one to celebrate my success. I am glad I have him to share my joys and sorrows in the past nine years and I look forward to the future in which we will spend the rest of our life together.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AI-ECD	Activated ion electron capture dissociation
AI-ETD	Activated ion electron transfer dissociation
CID	Collision induced dissociation
CZE	Capillary zone electrophoresis
Da	Dalton
DDA	Data-dependent acquisition
ECD	Electron capture dissociation
E. coli	*Escherichia coli*
EOF	Electroosmotic flow
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FT-ICR	Fourier-transform ion cyclotron resonance
FWHM	Full width at half maximum
HCD	High-energy collision dissociation
ID	Identification
LC	Liquid chromatography
LFQ	Label-free quantification
m/z	Mass-to-charge ratio
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
PTM	Post-translational modification
RPLC	Reversed-phase liquid chromatography
SEC	Size exclusion chromatography
TOF	Time-of-flight
UVPD	Ultraviolet photodissociation

CHAPTER 1. Introduction

1.1 Multi-level mass spectrometry-based proteomics

1.1.1 Overview of proteomics

Proteins are one of the most essential elements of biological systems and participate in virtually every process in cells or organs [1]. The proteome, which is the entire set of proteins expressed by an organism's genome, is extremely complex due to the heterogeneity derived from multiple biological processes [2] including genetic variation, alternative RNA splicing and post-translational modifications (PTMs), **Figure 1.1**. The protein products of a gene with different molecular forms due to the above modifications are designated as proteoforms [3]. Another word, "proteoform family", represents a group of proteoforms derived from the same gene [4].

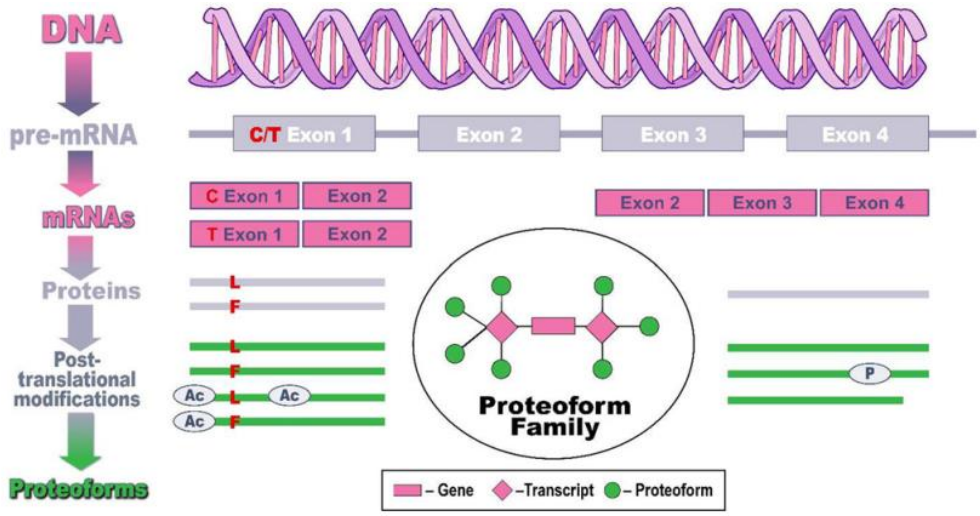

Figure 1.1. The sources of protein variation contributing to different proteoforms. The figure is reprinted with permission from reference [2].

Part of this chapter was adapted with permission from: X. Shen, Z. Yang, E. N. McCool, R. A. Lubeckyj, D. Chen, L. Sun, TrAC, Trends Anal. Chem. 120 (2019) 115644.

Proteomics is the study of proteomes including all proteoforms as well as their structures, interactions, and functions [1,5]. It is a critical technology for biological and clinical research, such as biomarker discovery, single cell analysis and disease diagnosis [6-9]. With the development of electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI), two advanced ionization techniques for mass spectrometric analysis of biological macromolecules that garnered the Chemistry Nobel Prize in 2002, mass spectrometry (MS) has become the core tool for proteomic analysis [5,10]. In a typical proteomic study, proteins are first extracted from cell or tissue samples. Then the proteins can be either digested into peptides or remain intact, followed by one or multiple dimensions of liquid-phase separations performed online or offline. Both mass spectra for precursor ions and tandem mass spectra for fragment ions are collected, and proteins are identified by database searching against a genome sequence database. Two approaches of proteomics, bottom-up and top-down, are named for the strategies with or without proteolytic digestion (**Figure 1.2**).

1.1.2 Bottom-up proteomics

Bottom-up proteomics is an indirect measurement of proteins through their digested peptides. Proteolytic digestion is usually performed after denaturation, reduction and alkylation of proteins. The serine protease trypsin is the enzyme of choice for digestion because it specifically cleaves proteins into peptides with an average size of 600-1000 Da, which is ideal for mass spectrometry analysis [11]. Also, trypsin has very high proteolytic activity, so that only a limited amount of trypsin is required for protein digestion, which minimizes the interference by autoproteolysis peptides in sample analysis. Then the resultant peptides are typically separated by liquid chromatography (LC) and analyzed by ESI-MS/MS. The protein identifications are accomplished by inferring the peptide sequence to the protein sequence [13]. The peptides can be

either uniquely annotated to a single protein or shared with multiple proteins, thus all proteins inferred from the same peptides are reported as one protein group.

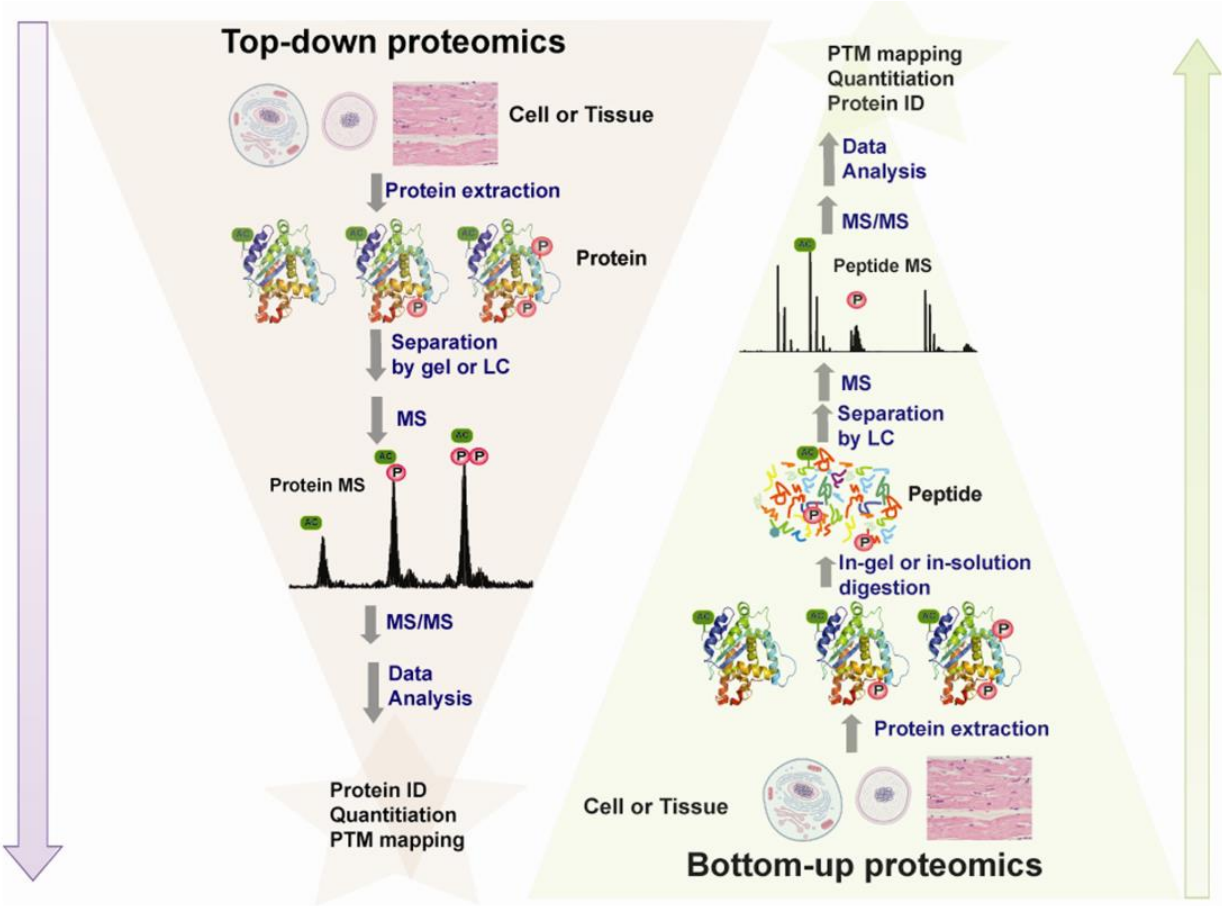

Figure 1.2. Schematic of top-down and bottom-up proteomics. The figure is reprinted with permission from reference [12].

Quantitative analysis is also commonly conducted in bottom-up proteomics to reflect the abundance changes of proteins and better address biological questions [13]. Isotope labeling and label-free quantification (LFQ) are two widely used quantification methods to compare the relative abundance of proteins and peptides between different samples. In isotope labeling, peptides are labeled with isobaric tags for relative and absolute quantification (iTRAQ) [14] or tandem mass tags (TMT) [15]. The chemical structure of TMT is shown as an example in **Figure**

1.3. It is composed of an amine-reactive group, a spacer arm (mass normalizer), and a mass reporter.

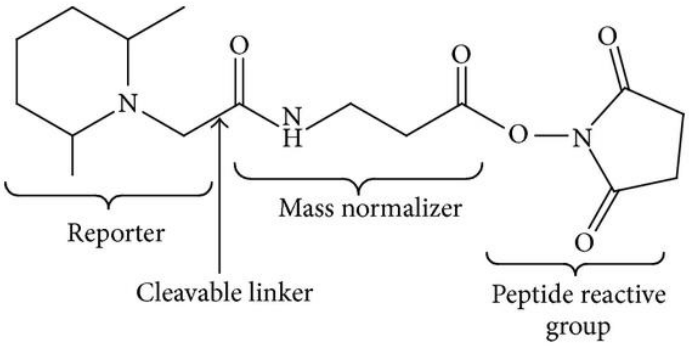

Figure 1.3. Functional regions of the TMT reagent's chemical structure, including MS/MS sites of fragmentation by HCD. The figure is reprinted with permission from reference [16].

Briefly, peptides from multiple biological samples (up to 16 samples) are labeled with different isobaric tags through the reaction of the amine-reactive group with the amine groups on peptides. Peptides with the same mass are isolated for fragmentation in tandem mass spectrometry (MS/MS), in which the linkers are cleaved, and the mass reporters are released. Then the intensity of these mass reporter ions represents the relative abundance of labeled peptides from different samples (**Figure 1.4**). LFQ as an alternative method quantifies proteins based on the signal intensity of peptides in base peak chromatograms. The area of extracted chromatographic peak is used to represent peptide abundance [17]. The relative abundance of a protein is achieved by integrating all peak area of peptides from the given protein [18]. The advantage of LFQ is it avoids the additional sample preparation steps related to the isotope labeling, but it has a higher requirement for software programs that consider alignment of peptide retention times, background noise and mass accuracy, and is compatible with specified

systems [13]. The limited reproducibility owing to ion suppression or coelution of the samples would also affect the performance of LFQ.

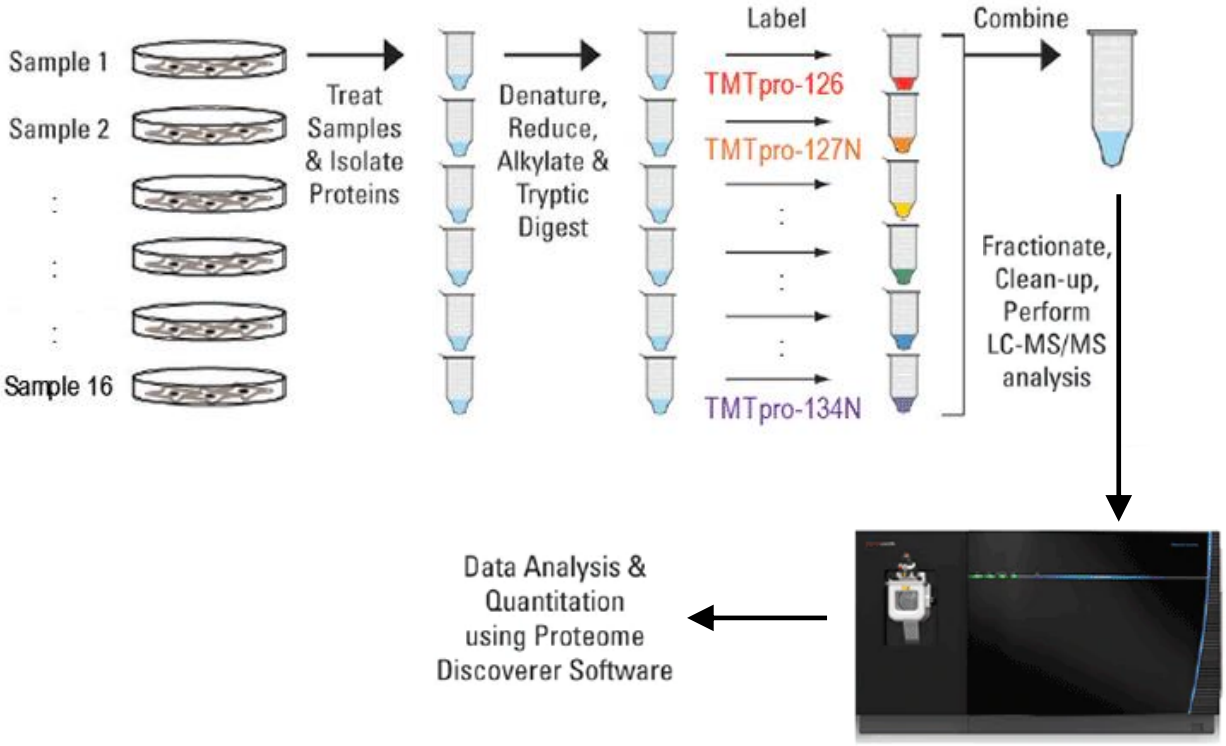

Figure 1.4. Procedure summary for MS experiments using TMT isobaric mass tagging reagents. This figure was modified from (https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-mass-spectrometry-analysis/protein-quantitation-mass-spectrometry/tandem-mass-tag-systems.html).

Bottom-up proteomics is the most widely used approach for protein sequence analysis for the last two decades because of several reasons. First, separation at the peptide level is highly efficient and easy to be performed. Reversed-phase LC (RPLC) is the most employed separation mechanism in bottom-up proteomics research, which provides high-resolution separation of peptides and fully automated instrument settings. A meter-long capillary packed RPLC column can provide more than 1000 peak capacity (theoretical number of peaks that can be resolved

within a retention window) for a single-dimension separation of peptides from yeast and bacteria lysate [19]. The peak capacity of one-dimension separation is usually not enough to resolve the entire complexity of whole proteome, thus, combing multiple dimensional separation methods together could be helpful. Multi-dimensional separation of mouse brain proteome peptides can reach about 7000 peak capacity [20]. The superior separation performance significantly reduces the coelution of peptides and allows large-scale characterization of proteins from complex samples [21]. Second, the small size of peptides is more favorable for mass spectrometry. The ionization efficiency is higher for peptides compared to intact proteins. High mass resolution of different isotopic compositions is easily achieved without rigorous demand for advanced mass spectrometers. The sensitivity of peptide analysis is also higher in general. The advanced nanoRPLC-ESI-MS/MS system confidently identified more than 4000 protein groups from only 1 μg of HeLa digest [22]. Moreover, peptides can be efficiently fragmented in mass spectrometers, which greatly improves the identification accuracy and localization of post-translational modifications (PTMs). Third, bioinformatic tools for bottom-up proteomics have been maturely developed. Many database searching software programs are either commercialized along with the instruments or freely open to the public, and are reliable, stable and user-friendly [23-26].

However, several inherent drawbacks exist in bottom-up proteomics. The sample preparation is tedious in the bottom-up approach and may cause artificial modifications on proteins and peptides. Moreover, limited sequence coverage and the use of protein inference from peptides cannot accurately determine proteoforms in bottom-up proteomics as shown in **Figure 1.5**.

1.1.3 Top-down proteomics

Top-down proteomics, as an alternative approach for proteomics, characterizes proteoforms in the cell. The workflow of top-down proteomics is similar to that of bottom-up proteomics except that proteins are not digested with proteolytic enzymes and are separated and detected at the intact protein level (**Figure 1.2**). The top-down approach surpasses the bottom-up approach in higher sequence coverage and potential full protein characterization. Thus, it can reveal the information of specific proteoforms including the relationship between primary protein sequence and PTM combinations (**Figure 1.5**). A subdiscipline of top-down proteomics is native top-down proteomics (native proteomics), which is performed under near-physiological conditions. It allows the characterization of protein complexes with noncovalent interactions at a global scale and in discovery mode [27,28].

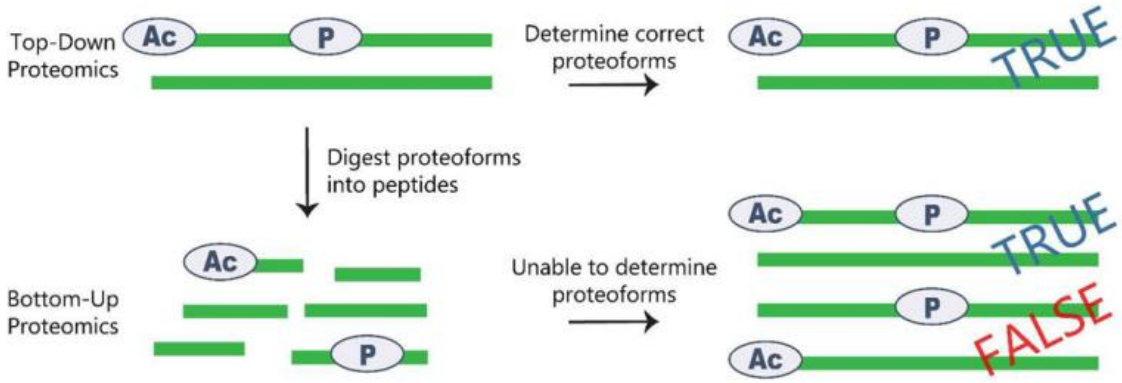

Figure 1.5. Determination of proteoforms from top-down and bottom-up proteomics. In bottom-up proteomics, proteoform information can be lost because proteoforms are digested into peptides. Ac: acetylation; P: phosphorylation. The figure is reprinted with permission from reference [2].

Despite its advantages, top-down proteomics encounters several technical challenges that preclude its wide application compared to the bottom-up approach. One challenge is the limited number of proteoform identifications. The proteome samples can be extremely complicated. It has been estimated that the human proteome contains over 1 million proteoforms with an enormous dynamic range [29]. Therefore, front-end separation is critical in the top-down workflow to reduce sample complexity and allow the detection despite high abundance proteoforms. With single-dimensional RPLC-MS/MS, around 1600 proteoforms from 563 proteoform families could be identified from *Salmonella typhimurium* [30]. However, compared to peptides, separation of proteins is less efficient due to the low solubility of many intact proteins and stronger adsorption effect of proteins on chromatography media [31]. Employing multi-dimensional liquid-phase separation before MS can greatly expand the proteome coverage. Catherman *et al.* combined subcellular fractionation, solution isoelectric focusing (sIEF), gel elution liquid fraction entrapment electrophoresis (GELFrEE), and RPLC-MS/MS for deep top-down proteomics and identified over 5000 proteoforms and 1220 proteoform families from human cell line H1299 proteome [32]. An orthogonal multidimensional separation platform developed by the Sun group that couples size exclusion chromatography (SEC) and RPLC based protein prefractionation to capillary zone electrophoresis (CZE)-MS/MS allowed nearly 6000 proteoform identifications from 850 proteoform families from the *Escherichia coli* (*E. coli*) proteome [33]. Nevertheless, the identification of proteoforms is still far away from complete proteome coverage of complex samples. Proteoforms usually have wide charge state distribution, so the signal of multiple proteoforms coeluted from the same peak could overlap on each other and interfere the generation of clear MS/MS spectra. More efforts are needed to explore new separation strategies as well as combinations of multi-dimensional separation methods.

For native proteomics, the requirement of native condition largely restricts the options of separation methods. Traditional native separation techniques include SEC and ion exchange chromatography (IEX) which have been reported previously to characterize simple samples such as standard protein aggregates [34,35]. However, their relatively low separation peak capacity limits their application on more complex samples. Alternative native separation techniques such as native GELFrEE [36] and native CZE [37] recently show the potential for large-scale native proteomics and are worth further investigation. In addition to front-end separation, high-resolution MS instrumentation with extended detection ranges is crucial for the native approach, because native protein complexes usually have much larger mass than single intact proteins. Recent advancements in mass analyzers including Fourier Transform Ion Cyclotron Resonance MS (FTICR), Orbitrap, and time-of-flight (TOF) continually expand the MS performance at large mass-to-charge ratio (*m/z*) range and facilitate the application of native top-down as well as denaturing top-down proteomics for large proteins and protein complexes.

Comprehensive characterization of proteoforms is another challenge in top-down proteomics. The lack of extensive gas-phase fragmentation of proteoforms hinders the accurate localization of PTMs. Besides the widely used vibrational energy-based dissociation, such as collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD), alternative gas-phase fragmentation methods including electron-based methods and ultraviolet photodissociation (UVPD) have been applied to increase the backbone cleavage coverage in MS/MS. More details are discussed in ensuing sections on tandem mass spectrometry (**section 1.1.5**).

1.1.4 Mass spectrometry

Mass spectrometry (MS) is a major analytical method to study proteins and can provide comprehensive information of each protein such as abundance, modifications, structure,

interactions, *etc* [38]. It measures the *m/z* of charged analytes in the gas phase. Because MS for proteins is usually coupled with liquid phase *separation*, soft ionization methods are required to transform nonvolatile protein and peptide ions in liquid phase to gaseous forms without degrading the analytes. Electrospray ionization (ESI) is the most widely used ionization technique for protein studies. In ESI, a high voltage is applied at the end of a capillary tip infused with the sample solution (**Figure 1.6**). The flow rate is about several hundred μL/min and well compatible with LC separation. Then charged droplets containing analytes are emitted from a Taylor cone formed at the capillary tip [39]. These droplets then experience rapid solvent evaporation and jet fission, and their size shrinks from couple microns to a few nanometers [40]. Eventually, gaseous ions are generated from the nanodroplets. NanoESI is a variation of ESI, in which the orifice of the emitter tip is only a few microns instead of about 100 microns in regular ESI [41]. The smaller orifice size results in smaller radii of the initial droplets, leading to enhanced ionization efficiency and sensitivity [42]. NanoESI is operated at lower flow rates (nL/min), so it can be coupled with capillary LC and CZE.

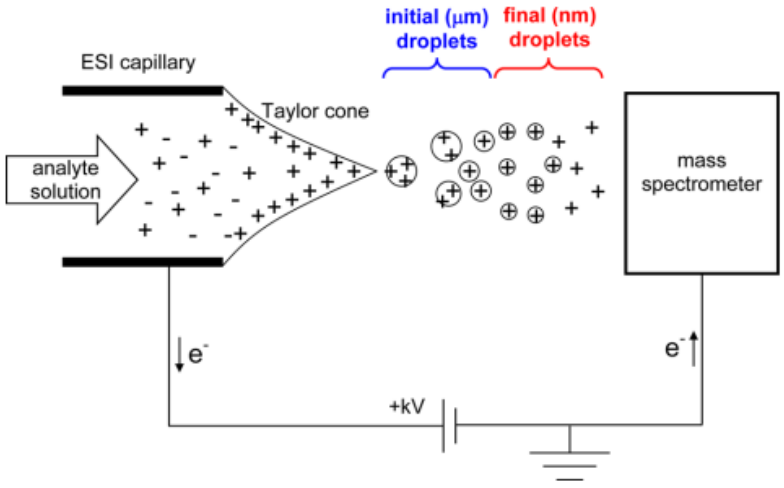

Figure 1.6. Schematic ESI process operated in positive ion mode. The figure is reprinted with permission from reference [41].

After ionization, protein and peptide ions are measured by a mass analyzer, which is the central component in a mass spectrometer. The Orbitrap mass analyzer has achieved increased popularity in the proteomics field in the last decade because of its high mass accuracy (sub-ppm level) and high mass resolution (up to 1 million FWHM at *m/z* 200 for Orbitrap Eclipse Tribrid Mass Spectrometer). It was developed by Makarov [43] and the commercialized Orbitrap-based mass spectrometer was first released in 2005 by Thermo Fisher Scientific [44]. The Orbitrap analyzer is composed of one spindle-like central electrode and two cup-shaped outer electrodes [45], as shown in **Figure 1.7**. The trajectories of ions in the Orbitrap analyzer are driven by three kinds of cyclic motions: the rotational motion around the central electrode, the radial motion between the outer and central electrodes, and the axial oscillations along the central electrode. The first two kinds of motions make the ions spiral around the central electrode and are responsible for trapping the ions in the analyzer, while the third kind of motion is used to determine the *m/z* of ions. The ion movements along the central electrode induce image currents whose frequencies depend on *m/z*, which can be Fourier-transformed and further converted to *m/z* [46]. Recent development of the Orbitrap mass spectrometer enables the detection range up to 80,000 *m/z*, which provides the potential for characterization of large biomolecules at MDa levels.

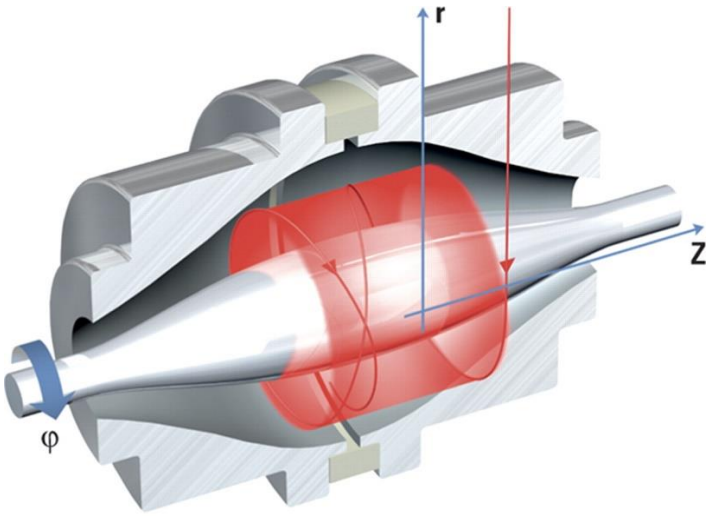

Figure 1.7. Orbitrap mass analyzer. The figure is modified with permission from reference [47].

The time-of flight mass (TOF) analyzer is another mainstream device for mass spectrometry investigations of proteins. It has fairly high mass resolution, fast data acquisition, and an unlimited mass scale in theory [48]. The first TOF-based mass spectrometer can be traced back to 1946 [49]. In the 1950s, Wiley and McLaren made key advances, leading to the commercialization of TOF-MS by Bendix [50]. Briefly, a TOF analyzer contains a flight tube and a system of grids to accelerate (typically orthogonally) ions from the ion source in an electric field [51]. All ions from the ion source gain the same kinetic energy during acceleration, but they drift in the field-free flight tube with different velocities if they have different *m/z* values. Ions with larger *m/z* values fly slower than ions with smaller *m/z* values, so that the *m/z* value of the ions can be determined as a function of time required to arrive at the detector. Reflectron TOF-MS (**Figure 1.8**) is widely employed to increase the mass resolution by correcting spatial and kinetic energy distributions of the ions during acceleration [48]. Although a TOF mass spectrometer generally has lower mass resolution than an Orbitrap mass spectrometer, it has a

faster acquisition speed and can work for large biomolecules (*e.g.* up to 30,000 *m/z* for an Agilent 6545XT Q-TOF mass spectrometer) with much lower cost.

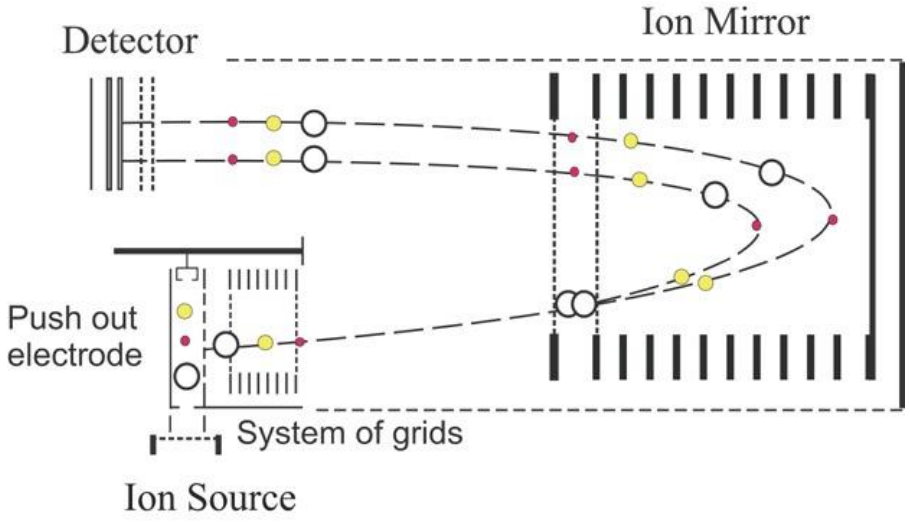

Figure 1.8. Reflectron TOF mass spectrometer with orthogonal acceleration. The figure is reprinted with permission from reference [48].

1.1.5 Tandem mass spectrometry

Proteomics studies usually employ tandem mass spectrometry (MS/MS) to achieve more comprehensive information of proteins, such as primary sequence and PTMs, in addition to the intact masses. MS/MS is comprised of two (or more) mass analyzers in the instrument [52]. The first analyzer is used to isolate protein or peptide ions with a specific *m/z*, which are called the precursor ions. The selected precursor ions are then dissociated into small pieces, termed as product (fragment) ions, typically with some kind of gas-phase fragmentation methods, followed by detection with a second mass analyzer. A variety of fragmentation methods with different fundamental mechanisms have been developed to achieve more efficient fragmentation and provide complementary information about proteins and peptides.

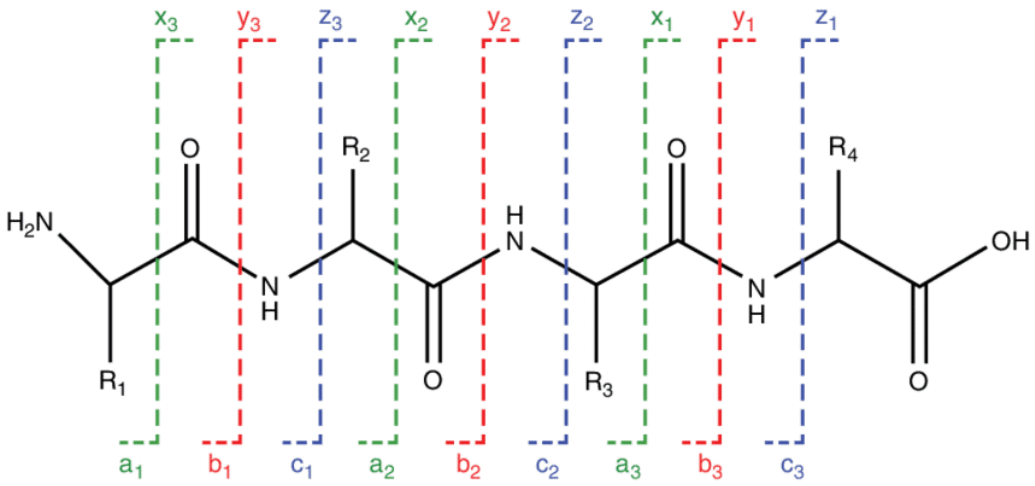

Figure 1.9. The peptide fragmentation nomenclature. The figure is reprinted with permission from reference [53].

Collision-induced dissociation (CID) is the most widely used fragmentation method for peptides and proteins in proteomics studies. In CID, ions are accelerated by an electrical potential and collide with neutral gas molecules like nitrogen or helium in an ion trap. Their kinetic energy is then converted to internal energy, which results in bond cleavage and produces b- and y- type ions depending on localization of charge [54], **Figure 1.9**. Another collision-based fragmentation method is higher energy collisional dissociation (HCD). It has the same mechanism as CID, but the dissociation occurs in the multipole collision cell [55], thus it has no low-mass cutoff and can be applied with isotope labeling quantitation [56]. However, CID and HCD have preferential cleavage of the most labile bonds, limiting the sequence coverage and labile PTM localization of proteoforms [57].

Electron-based activation methods like electron capture dissociation (ECD) and electron transfer dissociation (ETD) are alternatives to CID and HCD for protein and peptide fragmentation. They both generate cation radicals of the analytes that will undergo breakage of bonds quickly and produce c- and z-type fragment ions, as shown in **Figure 1.9**. In ECD, a

multiply-charged analyte cation reacts with a free electron, leading to an excited cation radical [58]. While in ETD, the unstable cation radical is generated by transferring an electron from an anion radical to the analyte cation [59], **Figure 1.10**.

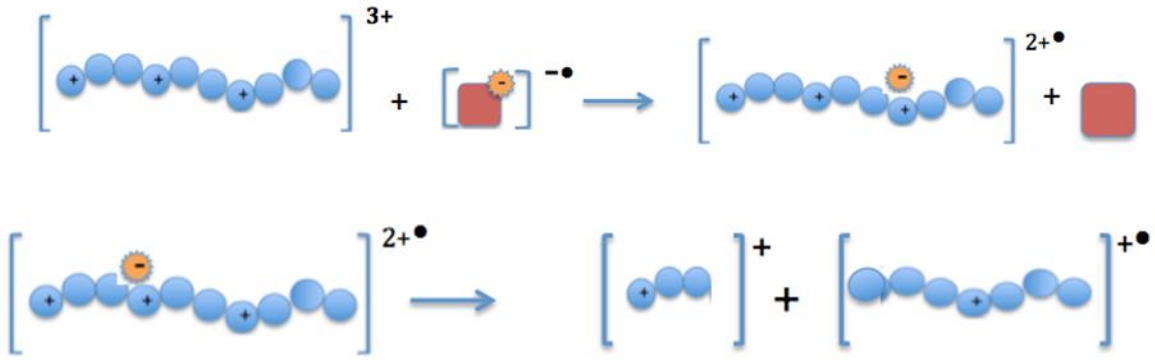

Figure 1.10. Fragmentation mechanism of ETD. The figure is reprinted from (https://en.wikipedia.org/wiki/Electron-transfer_dissociation#cite_note-13).

Fragmentation by ECD and ETD is a non-ergodic process [60], which means the bond cleavage happens rapidly at the sites where cation radicals were formed originally before the energy is redistributed through the molecules. Therefore, they can provide more extensive fragmentation of the peptide chain as well as retain labile modifications. For example, Kelleher *et al.* showed ECD in a Fourier-transform ion cyclotron resonance (FT-ICR) mass spectrometer localized γ-CO_2 moiety and SO_3 modifications, which were ejected from peptides by CID [61]. Molina *et al.* demonstrated that ETD could identify 60% more phosphopeptides than CID in an ion trap mass spectrometer [62]. Because of their nature being non-ergodic processes, ECD and ETD have become powerful tools to increase sequence coverage of intact proteoforms in top-down proteomics. However, the conversion efficiency from precursor ions to fragment ions in ECD is relatively low (typically <33% [63]), so ECD is commonly employed for target protein fragmentation through direct infusion MS, in which high sample concentration can be used and a

large number of spectra can be combined. With ECD, about 93% sequence coverage can be achieved for carbonic anhydrase II (29 kDa) [64]. While ETD has been applied for large-scale top-down proteomics and enabled 3,000-5,000 proteoform identifications from human proteome by 4D-separation-MS/MS [32,65], one limitation of ETD is nondissociative electron transfer dissociation (ETnoD), which hinders the extensive fragmentation of ETD for intact proteins. ETnoD is a process in which protein backbone is cleaved, but product ions are held together by noncovalent interaction, thus no fragment ions are produced. Activated ion ETD (AI-ETD) is developed as an improved ETD technique to minimize ETnoD [66,67]. Supplemental energy provided by infrared photoactivation concurrent with ETD can disrupt the noncovalent binding and increase ETD efficiency while incurring no additional time costs to the MS/MS scan event. Our group recently performed large-scale top-down proteomics using SEC-CZE-MS/MS with AI-ETD on an Orbitrap Fusion Lumos mass spectrometer and identified 3028 proteoforms and 387 proteoform families from *E. coli* cells [68].

Ultraviolet photodissociation (UVPD) is another gas-phase fragmentation technique and has also been well recognized for enhancing proteoform fragmentation [69-71]. Because of the absorption of high energy UV photons (typically 193 nm or 213 nm), the protein ions are activated to the electronic excitation states, and bond cleavage subsequently occurs over the entire amino acid sequence of the protein [31]. Therefore, UVPD can produce a variety of fragment ions (a-, b-, c-, x-, y-, z-type ions), and obtain better protein fragmentation coverage compared to CID and HCD. Cleland *et al.* demonstrated that although HCD could lead to a higher number of proteoform identifications from Human cell lysate, 193 nm UVPD provided higher average sequence coverage of proteoforms and more confident localization of PTMs [71].

1.2 Capillary zone electrophoresis-mass spectrometry

1.2.1 Capillary zone electrophoresis (CZE)

The dominant separation technique for proteomics research is RPLC, which separates analytes based on their hydrophobicity [72]. However, a comprehensive and high-throughput analysis cannot be achieved by this single separation method due to the peak capacity limit. Moreover, drawbacks have emerged when RPLC is applied to intact protein separation [73]. For example, the adsorption of proteins to the stationary phase and the protein conformational heterogeneity result in peak broadening and poor peak capacity during separation.

CZE is a complementary separation technique, which separates proteins and peptides based on their electrophoretic mobility that relates to analytes' sizes and charges [74]. Usually, CZE employs fused silica open tubular capillary and the separation is performed under an electric field. The typical inner diameter (i.d.) of the fused silica capillary used for CZE is in the range of 10-75 μm; the typical length of the capillary ranges from 20 to 100 cm. As shown in **Figure 1.11**, the analyte apparent mobility (μ_{APP}) is contributed by two parts: the electrophoretic mobility (μ_{EP}) and the electroosmotic mobility (μ_{EO}):

$$\mu_{APP} = \mu_{EP} + \mu_{EO} \qquad \text{(Equation 1.1)}$$

The μ_{EP} of analyte ions are determined by Debye-Hückel-Henry theory as

$$\mu_{EP} = \frac{ze}{6\pi\eta r} \qquad \text{(Equation 1.2)}$$

where z is the ion's net charge, e is the elementary charge, η is the viscosity of the background electrolyte (BGE), and r is the ion's radius. As a result, the cations migrate towards the cathode

and the anions migrate towards the anode. Neutral analytes have no charge, therefore, they have no electrophoretic mobility in an electric field..

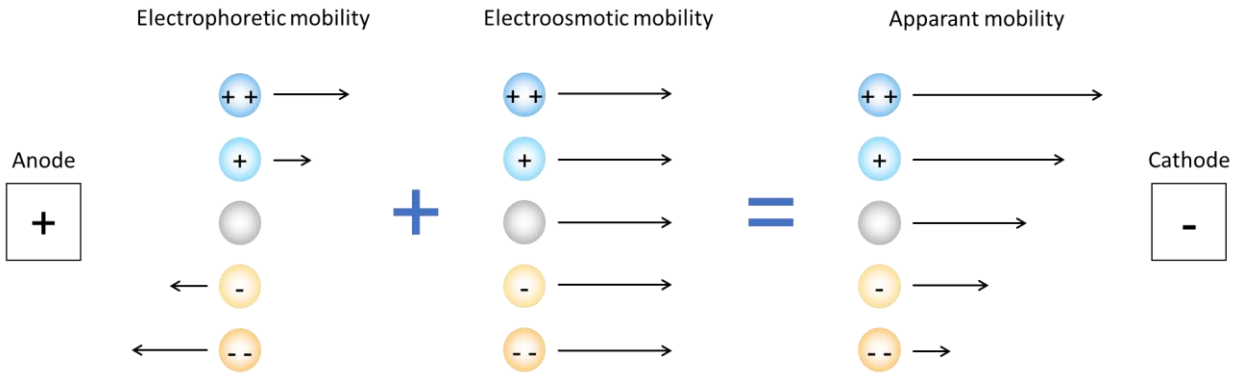

Figure 1.11. The mechanism of CE separations.

An analyte's μ_{EO} is caused by the electroosmotic flow (EOF) in the capillary. Because the inner wall of the fused silica capillary is covered with silanol groups, which are negatively charged at pH 3 and higher, it attracts cations from BGE and forms the electrical double layer at the capillary-solution interface. When high voltage is applied across the capillary, those cations carry the solution in the capillary and together move towards the cathode. This migration is termed EOF and μ_{EO} is defined as

$$\mu_{EO} = \frac{\varepsilon\zeta}{4\pi\eta} E \qquad \text{(Equation 1.3)}$$

where ε is the dielectric constant of the BGE, η is the viscosity of the BGE, E is the electric field, and ζ is the zeta potential.

CZE has an almost orthogonal separation mechanism to RPLC and can improve the proteome coverage in proteomics studies [75]. It has also shown better performance for low sample volume and the separation can be accurately predicted due to the simple separation mechanism.

Moreover, CZE-MS has several valuable features for top-down proteomics. First, CZE can reach a highly efficient separation of large biomolecules like proteoforms. As shown in Equation 1.4,

$$N = \frac{\mu V}{2D} \quad \text{(Equation 1.4)}$$

the number of theoretical plates from CZE (N) only relates to the electrophoretic mobility of analytes (μ), the voltage applied across the capillary (V), and the analytes' diffusion coefficient (D). Large biomolecules like proteoforms usually have low diffusion coefficients in solution, leading to high separation efficiency in CZE. Our most recent data showed that CZE can achieve up to one million theoretical plates for the separation of certain proteoforms [76]. Second, CZE-MS has extremely high sensitivity for top-down characterization of proteins. In 1996, the McLafferty group achieved the detection of attomole amounts of intact proteins (less than 1 pg in mass) using CZE-MS [77]. The Yates group has reported that CZE-MS approached comparable signal-to-noise ratios (S/N) to the widely used nanoRPLC-MS for characterization of intact proteins with 100-fold less sample consumption [78]. Third, CZE has the capability for high-resolution separation of protein complexes under native conditions [79]. This feature is unique to CZE and makes CZE-MS valuable for native top-down proteomics that aims to characterize endogenous protein complexes in the cell at a proteome-scale and in discovery mode [80,81].

Wide application of CZE-MS for large-scale top-down proteomics has been impeded by multiple factors. First, the stability and sensitivity of the CE-MS interface have been major obstacles. Second, the separation window and sample loading capacity of CZE has been at least 10-fold narrower and 100-fold lower than that of RPLC-MS, respectively, which hampered the adoption of CZE-MS in deep and large-scale proteomics.

1.2.2 CE-ESI-MS interfaces

CE-MS requires an interface that can complete the electrical circuit for CE separation and provide voltage for ESI. There are two major categories of CE-MS interface: sheath-flow interface and sheath-less interface.

Using sheath liquid for electrical contact is simple and convenient as it is decoupled from separation conditions. The sheath liquid can also help to modify the eluent composition for better ESI performance. In 1988, the Smith group reported the pioneering development of a coaxial sheath-flow CE-MS interface [82]. As shown in **Figure 1.12**, the separation capillary is inserted in two coaxial tubes [83]. The sheath liquid is filled in the gap between the capillary and the inner tube and merges with CE effluent at the exit of the capillary to realize the electrical contact. The gap between the coaxial tubes is filled with the sheath gas to assist the ESI process. However, the sheath liquid flow rate (1-10 μL/min) is much higher than the flow rate in CZE (20-100 nL/min), which would cause significant sample dilution and lower the sensitivity of detection.

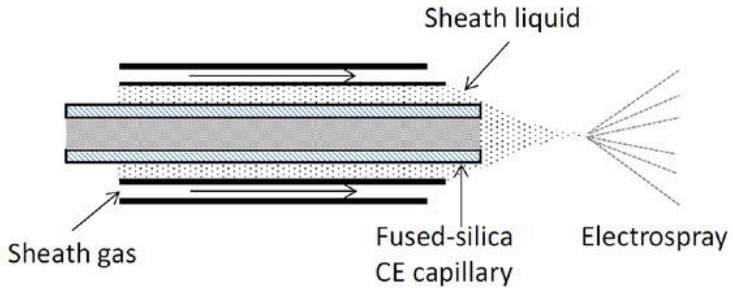

Figure 1.12. Schematic of the coaxial sheath-liquid interface. The figure is reprinted with permission from reference [83].

The sensitivity of sheath-flow interfaces could be improved by reducing the flow rate of sheath liquid. The Chen group developed a junction-at-the-tip type CE-MS interface in 2010

[84]. The design allows the sheath liquid solution driven by pressure flow through at a much lower flow rate compared to the coaxial sheath-flow CE-MS interface (nL/min vs. µL/min), leading to significantly higher sensitivity and a stable spray [84].

The Dovichi group reported the electro-kinetically pumped sheath flow interface in 2010 [85] and improved it further in 2013 and 2015 [86,87]. **Figure 1.13** shows diagrams of the basic interface and its different generations. High potential applied in the sheath buffer reservoir produces electroosmotic flow (EOF) in the glass emitter, which pumps sheath liquid at nL/min flow rates through the emitter for ESI, leading to extremely high sensitivity [86]. Larger ESI emitter orifice and shorter distance between the capillary end and emitter orifice improve the robustness and sensitivity of the CE-MS interface. The improved electro-kinetically pumped sheath flow interface has been commercialized by CMP Scientific (http://www.cmpscientific.com).

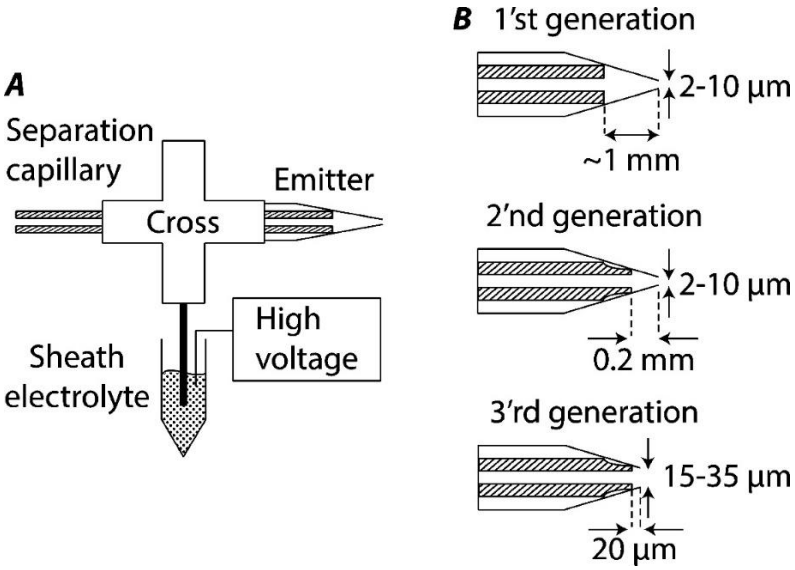

Figure 1.13. Diagrams of the basic design of the electrokinetically pumped sheath flow CE-MS interface (A) and its three different generations (B). The figure is reprinted with permission from reference [87].

Another strategy to increase the sensitivity is not employing sheath liquid in the CE-MS interface. The Moini group developed a sheathless CE-MS interface using a porous capillary end as the ESI emitter in 2007 [88]. The major benefit of the sheathless interface is the elimination of sample dilution by sheath liquid, thus leading to high sensitivity. The sheathless interface has been commercialized by Sciex and is used in the CESI 8000 and 8000 plus systems.

1.2.3 Capillary coating

CZE typically employs a regular fused silica capillary for separation and the inner wall of the capillary is covered with silanol groups, which causes an EOF as mentioned in **section 1.2.1** and will push the analytes out of the capillary for detection quickly. Therefore, CE separation is typically fast with a separation window in a range of 1-30 min [86,89,90]. This feature makes CZE-MS attractive for high throughput analysis of relatively simple samples. However, only about 5 MS/MS spectra can be generated per second in a typical MS/MS approach. The narrow separation window of CZE limits the number of MS/MS spectra that can be acquired during one run, leading to unsatisfying performance of CZE-MS for large-scale proteomics. Thus, boosting the separation window of CZE is crucial. The inner wall properties of fused silica capillaries could have a big impact on the separation window.

Several kinds of neutral and hydrophilic coatings, *e.g.*, linear polyacrylamide (LPA) and hydroxypropyl cellulose (HPC), have been utilized to cover the capillary inner wall and eliminate the EOF in CZE, leading to wider separation windows. The neutral coatings could also suppress protein adsorption on the capillary inner wall. The LPA coating is the most widely used neutral coating for CE-MS-based proteomic studies [75]. The preparation of the LPA coating has been reviewed recently by the Dovichi group [75]. CZE-MS system with an LPA-coated capillary could produce a 90-min separation window and a high peak capacity of nearly 300 for

top-down proteomics of an *E. coli* sample [91]. The separation window of the CZE-MS system is significantly wider than that of typical CZE-MS systems with uncoated capillaries [86,89,90].

Cationic coatings have also been used to coat the inner wall of the capillary for top-down proteomics [92-95]. In this case, the capillary inner wall has rich positive charges that reduce the protein adsorption on the capillary inner wall because the proteins are also positively charged in an acidic BGE. Upon applying a negative potential across the capillary, EOF towards the ESI tip will be generated. Proteins will migrate to the inlet of the capillary, but meanwhile will be pushed to the outlet of the capillary by the EOF [93-95]. Therefore, the migration rate of proteins can be abated, resulting in a wider separation window. However, the improvement of separation window using capillaries with cationic coatings is modest because of the strong EOF inside of the capillary.

1.2.4 Sample loading capacity of CZE

CZE employs an open-tubular capillary for separation without stationary phase, meaning that the analytes cannot be trapped at the front end of the separation capillary like in RPLC, which results in a low loading capacity of CZE. The typical sample loading volume is less than 1% of the total capillary volume to obtain high separation efficiency. For a 1-meter-long capillary with a 50-μm i.d., the total capillary volume is about 2 μL, and the sample loading volume needs to be only 20 nL or lower. Sample loading volumes in CZE are orders of magnitude lower than that in RPLC, which is challenging for the detection of low abundance proteoforms in a complex sample. The use of online sample preconcentration/stacking methods could help to improve the sample loading capacity.

Several preconcentration methods have been applied in CZE-MS-based proteomics, including field-amplified sample stacking (FASS), transient isotachophoresis (tITP), and dynamic pH junction. FASS is a simple technique for sample stacking. It is based on the idea that sample ions experience a dramatic decrease in velocity when migrating through a low-conductivity sample plug into a high-conductivity BGE zone and are stacked at the boundary between the sample and BGE zones. The addition of organic solvents, *e.g.*, acetonitrile (ACN), in the sample buffer for lowering the conductivity of the sample zone is an efficient way to perform FASS [96,97]. In the top-down proteomics study of a *Mycobacterium marinum* secretome, Zhao *et al.* employed a 70% (v/v) acetic acid to dissolve the sample, which had much lower conductivity than the BGE that was 0.25% (v/v) formic acid, to realize FASS and increased the sample loading volume to 120-nL (12% of the total capillary volume), leading to the identification of 22 proteoform families and 58 proteoforms from the secretome sample [98].

TITP requires the presence of a leading electrolyte (LE) and a terminating electrolyte (TE), whose electrophoretic mobility is higher and lower than the sample ions. At the beginning of a CZE separation, a plug of sample dissolved in a leading electrolyte (LE) and a plug of a terminating electrolyte (TE) are sequentially introduced into the capillary. After a voltage is applied, sample ions between LE and TE are arranged in the order of their mobility and are concentrated to achieve the same migrating velocity towards the outlet of the capillary. Larsson *et al.* employed cITP for peptide analysis with CZE and enlarged the sample injection volume to up to 900 nL, which was about 45% of the total capillary volume [99]. Li *et al.* were able to boost the sample loading volume of CZE to approximately 15% of the total capillary volume with tITP and identified 65 proteins in top-down proteomics of a *Pseudomonas aeruginosa* PA01 lysate using CZE-MS [94].

Dynamic pH junction is also a widely utilized stacking technique in top-down CZE-MS studies. A simplified diagram of the dynamic pH junction method with a neutrally coated capillary is shown in **Figure 1.14**. The sample is usually dissolved in a basic buffer (*i.e.*, ammonium bicarbonate, pH 8) and is injected into the separation capillary filled with an acidic BGE (*i.e.*, 5%(v/v) acetic acid, pH 2.4). Both ends of the capillary are then immersed in the BGE vials and two pH boundaries are formed in the capillary. The analytes in the basic sample zone mostly have negative charges. Upon applying a positive potential at the injection end of the capillary for separation, the protons titrate the basic sample zone gradually, and the pH boundary I starts to move towards the pH boundary II. Meanwhile, the negatively charged analytes migrate towards the moving pH boundary I and are concentrated there. Once the moving pH boundary I meets with the static pH boundary II, the analytes undergo a normal CZE separation. The dynamic pH junction method was invented by the Chen group in 2000 [100] and is a highly efficient method for online concentration of analytes, enabling the focusing of at least 95% of analytes injected into the capillary [101]. The Sun group systematically optimized the conditions of the dynamic pH junction-based CZE-MS in 2017 and achieved a microliter scale sample loading volume for both bottom-up and top-down proteomics, leading to nearly 600 proteoform identifications in a single run [91,102].

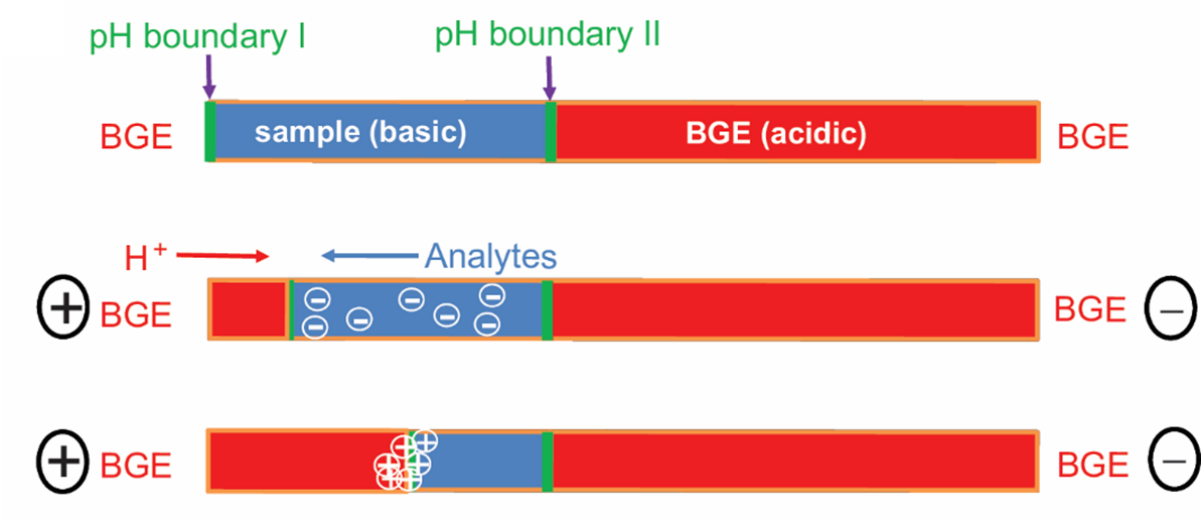

Figure 1.14. A simplified diagram of the dynamic pH junction method with a neutrally coated capillary.

1.3 Summary

This chapter introduced mass spectrometry-based proteomics for protein studies. Bottom-up proteomics, top-down proteomics and native proteomics are powerful tools for delineation of proteins, proteoforms, and protein complexes in complex proteomes with their own advantages and disadvantages. Combining the different approaches, termed multi-level proteomics, will be a future research direction for more comprehensive proteomic studies. Advances in the development of ESI, liquid phase separation, MS instrumentation, gas-phase fragmentation for MS/MS, and bioinformatic tools enable deep and large-scale analysis of peptides and proteoforms from complex samples with increasing proteome coverage. CZE as an alternative separation technique to RPLC has shown great potential for bottom-up and top-down proteomics. Numerous efforts have been made for CZE separation as well as CE-MS interfaces, capillary coating and online sample preconcentration methods, which significantly improve the

separation capacity, detection sensitivity and analysis stability of CZE-MS and render CZE-MS as a useful and complementary technique for the proteomics community. The subsequent chapters in this dissertation will describe four projects on proteomics by CZE-ESI-MS/MS with an electro-kinetically pumped sheath flow interface.

REFERENCES

REFERENCES

[1] M. Tyers, M. Mann, Nature 422 (2003) 193-197.

[2] L. V. Schaffer, R. J. Millikin, R. M. Miller, L. C. Anderson, R. T. Fellers, Y. Ge, N. L. Kelleher, R. D. LeDuc, X. Liu, S. H. Payne, L. Sun, P. M. Thomas, T. Tucholski, Z. Wang, S. Wu, Z. Wu, D. Yu, M. R. Shortreed, L. M. Smith, Proteomics 19 (2019) 1800361.

[3] L. M. Smith, N. L. Kelleher, Nat. Methods 10 (2013) 186-187.

[4] M. R. Shortreed, B. L. Frey, M. Scalf, R. A. Knoener, A. J. Cesnik, L. M. Smith, J. Proteome Res. 15 (2016) 1213-1221.

[5] R. Aebersold, M. Mann, Nature 422 (2003) 198-207.

[6] L. Chang, P. Graham, J. Hao, J. Bucci, D. Malouf, D. Gillatt, Y. Li, Cancer Lett. 369 (2015) 289–297.

[7] A. Nazeri, H. Ganjgahi, T. Roostaei, T. Nichols, M. Zarei, NeuroImage 102 (2014) 657-665.

[8] C. Lombard-Banek, S. A. Moody, P. Nemes, Angew. Chem. Int. Ed. Engl. 55 (2016) 2454-2458.

[9] L. Sun, K. M. Dubiak, E. H. Peuchen, Z. Zhang, G. Zhu, P. W. Huber, N. J. Dovichi, Anal. Chem. 88 (2016) 6653-6657.

[10] J. Cox, M. Mann, Cell 130 (2007) 395-398.

[11] Ü. A. Laskay, A. A. Lobas, K. Srzentić, M. V. Gorshkov, Y. O. Tsybin, J. Proteome Res. 12 (2013) 5558-5569.

[12] Z. R. Gregorich, Y. -H. Chang, Y. Ge, Pflügers Arch. 466 (2014) 1199-1209.

[13] Y. Zhang, B. R. Fonslow, B. Shan, M. -C. Baek, J. R. Yates, Chem. Rev. 113 (2013) 2343-2394.

[14] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, D. J. Pappin, Mol. Cell. Proteom. 3 (2004) 1154-1169.

[15] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, C. Hamon, Anal. Chem. 75 (2003) 1895-1904.

[16] S. Abdul-Ghani, K. J. Heesom, G. D. Angelini, M. -S. Suleiman, BioMed Res. Int. 2014 (2014) 1-11.

[17] D. Chelius, P. V. Bondarenko, J. Proteome Res. 1 (2002) 317-323.

[18] P. V. Bondarenko, D. Chelius, T. A. Shaler, Anal. Chem. 74 (2002) 4741-4749.

[19] F. Xie, R. D. Smith, Y. Shen, J. Chromatogr. A 1261 (2012) 78-90.

[20] D. Chen, X. Shen, L. Sun, Anal. Chim. Acta 1012 (2018) 1-9.

[21] J. Han, L. Ye, L. Xu, Z. Zhou, F. Gao, Z. Xiao, Q. Wang, B. Zhang, Anal. Chim. Acta 852 (2014) 267-273.

[22] C. D. Kelstrup, C. Young, R. Lavallee, M. L. Nielsen, J. V. Olsen, J. Proteome Res. 11 (2012) 3487-3497.

[23] D. C. Anderson, W. Li, D. G. Payan, W. S. Noble, J. Proteome Res. 2 (2003) 137-146.

[24] A. Keller, J. Eng, N. Zhang, X. Li, R. Aebersold, Mol. Syst. Biol. 1 (2005).

[25] J. Cox, M. Mann, Nat. Biotechnol. 26 (2008) 1367-1372.

[26] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, M. J. MacCoss, Bioinformatics 26 (2010) 966-968.

[27] B. Chen, K. A. Brown, Z. Lin, Y. Ge, Anal. Chem. 90 (2017) 110-127.

[28] F. Lermyte, Y. O. Tsybin, P. B. O'Connor, J. A. Loo, J. Am. Soc. Mass Spectrom. 30 (2019).

[29] R. Aebersold, J. N. Agar, I. J. Amster, M. S. Baker, C. R. Bertozzi, E. S. Boja, C. E. Costello, B. F. Cravatt, C. Fenselau, B. A. Garcia, Y. Ge, J. Gunawardena, R. C. Hendrickson, P. J. Hergenrother, C. G. Huber, A. R. Ivanov, O. N. Jensen, M. C. Jewett, N. L. Kelleher, L. L. Kiessling, N. J. Krogan, M. R. Larsen, J. A. Loo, R. R. Ogorzalek Loo, E. Lundberg, M. J. MacCoss, P. Mallick, V. K. Mootha, M. Mrksich, T. W. Muir, S. M. Patrie, J. J. Pesavento, S. J. Pitteri, H. Rodriguez, A. Saghatelian, W. Sandoval, H. Schlüter, S. Sechi, S. A. Slavoff, L. M. Smith, M. P. Snyder, P. M. Thomas, M. Uhlén, J. E. Van Eyk, M. Vidal, D. R. Walt, F. M. White, E. R. Williams, T. Wohlschlager, V. H. Wysocki, N. A. Yates, N. L. Young, B. Zhang, Nat. Chem. Biol. 14 (2018) 206-214.

[30] C. Ansong, S. Wu, D. Meng, X. Liu, H. M. Brewer, B. L. Deatherage Kaiser, E. S. Nakayasu, J. R. Cort, P. Pevzner, R. D. Smith, F. Heffron, J. N. Adkins, L. Pasa-Tolic, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 10153-10158.

[31] T. K. Toby, L. Fornelli, N. L. Kelleher, Annual Review of Anal. Chem. 9 (2016) 499-519.

[32] A. D. Catherman, K. R. Durbin, D. R. Ahlf, B. P. Early, R. T. Fellers, J. C. Tran, P. M. Thomas, N. L. Kelleher, Mol. Cell. Proteom. 12 (2013) 3465-3473.

[33] E. N. McCool, R. A. Lubeckyj, X. Shen, D. Chen, Q. Kou, X. Liu, L. Sun, Anal. Chem. 90 (2018) 5529-5533.

[34] K. Muneeruddin, M. Nazzaro, I. A. Kaltashov, Anal. Chem. 87 (2015) 10138-10145.

[35] K. Muneeruddin, J. J. Thomas, P. A. Salinas, I. A. Kaltashov, Anal. Chem. 86 (2014) 10692-10699.

[36] O. S. Skinner, L. H. F. Do Vale, A. D. Catherman, P. C. Havugimana, M. V. de Sousa, P. D. Compton, N. L. Kelleher, Anal. Chem. 87 (2015) 3032-3038.

[37] A. M. Belov, R. Viner, M. R. Santos, D. M. Horn, M. Bern, B. L. Karger, A. R. Ivanov, J. Am. Soc. Mass Spectrom. 28 (2017) 2614-2634.

[38] S. P. Gygi, R. Aebersold, Curr. Opin. Chem. Biol. 4 (2000) 489-494.

[39] X. Wu, R. D. Oleschuk, N. M. Cann, Analyst 137 (2012) 4150.

[40] P. Kebarle, U. H. Verkerk, Mass Spectrom. Rev. 28 (2009) 898-917.

[41] L. Konermann, E. Ahadi, A. D. Rodriguez, S. Vahidi, Anal. Chem. 85 (2012) 2-9.

[42] M. Wilm, Mol. Cell. Proteom. 10 (2011) M111. 009407.

[43] A. Makarov, Anal. Chem. 72 (2000) 1156-1162.

[44] S. Eliuk, A. Makarov, Annual Review of Anal. Chem. 8 (2015) 61-80.

[45] E. S. Hecht, M. Scigelova, S. Eliuk, A. Makarov, Encyclopedia of Anal. Chem. (2019) 1-40.

[46] R. H. Perry, R. G. Cooks, R. J. Noll, Mass Spectrom. Rev. 27 (2008) 661-699.

[47] M. Scigelova, M. Hornshaw, A. Giannakopulos, A. Makarov, Mol. Cell. Proteom. 10 (2011) M111. 009431.

[48] A. Radionova, I. Filippov, P. J. Derrick, Mass Spectrom. Rev. 35 (2015) 738-757.

[49] M. M. Wolff, W. E. Stephens, Rev. Sci. Instrum. 24 (1953) 616-617.

[50] W. C. Wiley, Science 124 (1956) 817-820.

[51] A. M. Haag, Mass Analyzers and Mass Spectrometers, Springer, Cham, 2016.

[52] R. D. Mittal, Indian J. Clin. Biochem. 30 (2015) 121-123.

[53] Z. Hao, Q. Hong, F. Zhang, S. -L. Wu, P. Bennett, Current Methods for the Characterization of Posttranslational Modifications in Therapeutic Proteins Using Orbitrap Mass Spectrometry, John Wiley & Sons, Inc., 2017.

[54] R. E. March, Encyclopedia of Spectroscopy and Spectrometry (2017) 330-337.

[55] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, M. Mann, Nat. Methods 4 (2007) 709-712.

[56] G. C. McAlister, D. H. Phanstiel, J. Brumbaugh, M. S. Westphall, J. J. Coon, Mol. Cell. Proteom. 10 (2011) O111. 009456.

[57] Y. Huang, J. M. Triscari, G. C. Tseng, L. Pasa-Tolic, M. S. Lipton, R. D. Smith, V. H. Wysocki, Anal. Chem. 77 (2005) 5800-5813.

[58] R. A. Zubarev, N. L. Kelleher, F. W. McLafferty, J. Am. Chem. Soc. 120 (1998) 3265-3266.

[59] M. -S. Kim, A. Pandey, Proteomics 12 (2012) 530-542.

[60] Q. Zhang, A. Frolov, N. Tang, R. Hoffmann, T. van de Goor, T. O. Metz, R. D. Smith, Rapid Commun. Mass Spectrom. 21 (2007) 661-666.

[61] N. L. Kelleher, R. A. Zubarev, K. Bush, B. Furie, B. C. Furie, F. W. McLafferty, C. T. Walsh, Anal. Chem. 71 (1999) 4250-4253.

[62] H. Molina, D. M. Horn, N. Tang, S. Mathivanan, A. Pandey, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 2199-2204.

[63] M. A. McFarland, M. J. Chalmers, J. P. Quinn, C. L. Hendrickson, A. G. Marshall, J. Am. Soc. Mass Spectrom. 16 (2016).

[64] J. B. Shaw, N. Malhan, Y. V. Vasil'ev, N. I. Lopez, A. Makarov, J. S. Beckman, V. G. Voinov, Anal. Chem. 90 (2018) 10819-10827.

[65] J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman, K. R. Durbin, J. D. Tipton, A. Vellaichamy, J. F. Kellie, M. Li, C. Wu, S. M. M. Sweet, B. P. Early, N. Siuti, R. D. LeDuc, P. D. Compton, P. M. Thomas, N. L. Kelleher, Nature 480 (2011) 254-258.

[66] N. M. Riley, M. S. Westphall, J. J. Coon, Anal. Chem. 87 (2015) 7109-7116.

[67] M. J. P. Rush, N. M. Riley, M. S. Westphall, J. J. Coon, Anal. Chem. 90 (2018) 8946-8953.

[68] E. N. McCool, J. M. Lodge, A. R. Basharat, X. Liu, J. J. Coon, L. Sun, J. Am. Soc. Mass Spectrom. 30 (2019) 2470-2479.

[69] X. Dang, N. L. Young, Proteomics 14 (2014) 1128-1129.

[70] J. B. Shaw, W. Li, D. D. Holden, Y. Zhang, J. Griep-Raming, R. T. Fellers, B. P. Early, P. M. Thomas, N. L. Kelleher, J. S. Brodbelt, J. Am. Chem. Soc. 135 (2013) 12646-12651.

[71] T. P. Cleland, C. J. DeHart, R. T. Fellers, A. J. VanNispen, J. B. Greer, R. D. LeDuc, W. R. Parker, P. M. Thomas, N. L. Kelleher, J. S. Brodbelt, J. Proteome Res. 16 (2017) 2072-2079.

[72] E. Shishkova, A. S. Hebert, J. J. Coon, Cell Syst. 3 (2016) 321-324.

[73] A. L. Capriotti, C. Cavaliere, P. Foglia, R. Samperi, A. Laganà, J. Chromatogr. A 1218 (2011) 8760-8776.

[74] J. Jorgenson, K. Lukacs, Science 222 (1983) 266-272.

[75] Z. Zhang, Y. Qu, N. J. Dovichi, TrAC Trends in Anal. Chem. 108 (2018) 23-37.

[76] R. A. Lubeckyj, A. R. Basharat, X. Shen, X. Liu, L. Sun, J. Am. Soc. Mass Spectrom. 30 (2019) 1435-1445.

[77] G. A. Valaskovic, N. L. Kelleher, F. W. McLafferty, Science 273 (1996) 1199-1202.

[78] X. Han, Y. Wang, A. Aslanian, B. Fonslow, B. Graczyk, T. N. Davis, J. R. Yates, J. Proteome Res. 13 (2014) 6078-6086.

[79] A. Nguyen, M. Moini, Anal. Chem. 80 (2008) 7169-7173.

[80] O. S. Skinner, N. A. Haverland, L. Fornelli, R. D. Melani, L. H. F. Do Vale, H. S. Seckler, P. F. Doubleday, L. F. Schachner, K. Srzentić, N. L. Kelleher, P. D. Compton, Nat. Chem. Biol. 14 (2017) 36-41.

[81] H. Li, H. H. Nguyen, R. R. Ogorzalek Loo, I. D. G. Campuzano, J. A. Loo, Nat. Chem. 10 (2018) 139-148.

[82] R. D. Smith, C. J. Barinaga, H. R. Udseth, Anal. Chem. 60 (1988) 1948-1952.

[83] H. Wu, K. Tang, Reviews in Anal. Chem. 39 (2020) 45-55.

[84] E. J. Maxwell, X. Zhong, H. Zhang, N. van Zeijl, D. D. Chen, Electrophoresis 31 (2010) 1130-1137.

[85] R. Wojcik, O. O. Dada, M. Sadilek, N. J. Dovichi, Rapid Commun. Mass Spectrom. 24 (2010) 2554-2560.

[86] L. Sun, G. Zhu, Y. Zhao, X. Yan, S. Mou, N. J. Dovichi, Angew. Chem. Int. Ed. Engl. 52 (2013) 13661-13664.

[87] L. Sun, G. Zhu, Z. Zhang, S. Mou, N. J. Dovichi, J. Proteome Res. 14 (2015) 2312-2321.

[88] M. Moini, Anal. Chem. 79 (2007) 4241-4246.

[89] S. B. Choi, M. Zamarbide, M. C. Manzini, P. Nemes, J. Am. Soc. Mass Spectrom. 28 (2017) 597-607.

[90] M. Moini, B. Martinez, Rapid Commun. Mass Spectrom. 28 (2014) 305-310.

[91] R. A. Lubeckyj, E. N. McCool, X. Shen, Q. Kou, X. Liu, L. Sun, Anal. Chem. 89 (2017) 12059-12067.

[92] R. Haselberg, C. K. Ratnayake, G. J. de Jong, G. W. Somsen, J. Chromatogr. A 1217 (2010) 7605-7611.

[93] X. Han, Y. Wang, A. Aslanian, M. Bern, M. Lavallee-Adam, J. R. Yates 3rd, Anal. Chem. 86 (2014) 11006-11012.

[94] Y. Li, P. D. Compton, J. C. Tran, I. Ntai, N. L. Kelleher, Proteomics 14 (2014) 1158-1164.

[95] D. R. Bush, L. Zang, A. M. Belov, A. R. Ivanov, B. L. Karger, Anal. Chem. 88 (2016) 1138-1146.

[96] L. Sun, M. D. Knierman, G. Zhu, N. J. Dovichi, Anal. Chem. 85 (2013) 5989-5995.

[97] Y. Zhao, L. Sun, M. D. Knierman, N. J. Dovichi, Talanta 148 (2016) 529-533.

[98] Y. Zhao, L. Sun, M. M. Champion, M. D. Knierman, N. J. Dovichi, Anal. Chem. 86 (2014) 4873-4878.

[99] M. Larsson, E. S. M. Lutz, Electrophoresis 21 (2000) 2859-2865.

[100] P. Britz-McKibbin, D. D. Y. Chen, Anal. Chem. 72 (2000) 1242-1252.

[101] L. Wang, D. MacDonald, X. Huang, D. D. Chen, Electrophoresis 37 (2016) 1143-1150.

[102] D. Chen, X. Shen, L. Sun, Analyst 142 (2017) 2118-2127.

CHAPTER 2. Systematic Evaluation of Immobilized Trypsin Based Fast Protein Digestion for Deep and High-Throughput Bottom-Up Proteomics

2.1 Introduction

Comprehensive characterization of complex proteomes using bottom-up proteomics has been achieved in only a couple of hours using modern RPLC-ESI-MS/MS [1]. However, at least 12 h is typically required to prepare the sample for RPLC-MS/MS analysis, which limits the overall throughput of bottom-up proteomics. The most time-consuming step during the sample preparation is the digestion of proteins using free trypsin (FT), which typically requires 12 h for complete digestion. Immobilized trypsin (IM) has been well recognized for speeding protein digestion [2,3]. IM can accomplish protein digestion in minutes due to the much higher concentration of trypsin compared with FT [2,3]. Moreover, the immobilization of trypsin greatly reduces the autodigestion of trypsin molecules.

Various solid matrixes have been used to immobilize trypsin, *e.g.*, beads [4-10], monolithic materials [11-16], and membranes [17-19]. Sun *et al.* reported 2,100 protein IDs from MCF7 cell lysate using 20-min IM digestion, and the number of protein IDs was comparable with that using 12-h FT digestion [5]. They also observed a significant loss of basic peptides using the IM digestion compared with FT digestion, most likely due to the negatively charged solid matrix (carboxyl groups functionalized magnetic beads) used for trypsin immobilization [5]. Fan *et al.* also observed a significant effect of the IM matrixes on the identified protein and peptide pools

---

Part of this chapter was adapted with permission from X. Shen, L. Sun, Proteomics 18 (2018) 1700432.

from complex proteome samples digested with IM [6]. Those data indicate that the surface chemistry of the solid matrix of IM can influence the tryptic digestion process.

IM has also been coupled to CZE-MS/MS [20-23] or RPLC-MS/MS [23-31] for online protein digestion, peptide separation, and identification. However, IM has not played a significant role in routine deep bottom-up proteomics studies. Some questions involving IM activity need to be answered to facilitate its wide application.

First, how does the solid matrix of IM influence its preference for protein cleavage in comparison to FT? Quantitative proteomics has been employed to reveal the preference of protein cleavage catalyzed by FT [32,33]. Using dimethyl labeling based quantitative proteomics [34,35], Ye *et al.* observed that the cleavage sites surrounded by neutral amino acids could be cleaved quickly, while sites surrounded by negatively charged amino acids (aspartic and glutamic acids) were cleaved much more slowly [32,33]. Šlechtová *et al.* also reached a similar conclusion about the cleavage preference of peptides catalyzed by FT using synthetic peptides as trypsin substrates [36]. To our best knowledge, the preference of protein cleavage catalyzed by IM and the effect of the solid matrixes of IM on the cleavage preference have not been studied using quantitative proteomics.

Second, how well can IM perform digestion of complex proteomes for deep proteomics compared with FT? Only a few reports in the literature have applied the IM based fast protein digestion for large-scale proteomics, resulting in 1000-3000 protein IDs from mammalian cell lines or tissues [5,6,28], and fewer than 1000 protein IDs from yeast cell lysate [24,26]. The routine deep bottom-up proteomics studies using FT digestion have approached over 8000 protein IDs from mammalian cell lines or tissues [37-40]. Deep proteomics datasets using IM

based fast protein digestion are required to demonstrate the capability of IM for deep proteomics and to confirm that IM can speed protein digestion without bias.

In this work, experiments were performed to provide answers to those two questions. I prepared amine and carboxyl functionalized magnetic beads-based IM (IM-N and IM-C), which represented a nearly neutral and negatively charged solid matrix surface at the trypsin digestion pH (pH 8). The preference of protein cleavage catalyzed by FT and two types of IM were investigated using label-free quantitative proteomics. Furthermore, both qualitative and quantitative analysis were conducted for the mouse brain proteome samples digested by FT (12 h) and IM-N (15 min) using both 1D- and 2D-LC-ESI-MS/MS. The FT and IM-N results were compared in terms of the identified protein and peptide pools that contained nearly 9,000 proteins and over 100,000 peptides. Finally, a high-throughput bottom-up workflow was developed using IM-N based rapid protein digestion and fast CZE-MS/MS analysis.

2.2 Experimental

2.2.1 Materials and reagents

Bovine pancreas TPCK-treated trypsin, 3-(Trimethoxysilyl)propyl methacrylate, ammonium persulfate, glycine, ammonium bicarbonate (NH_4HCO_3), dithiothreitol (DTT), iodoacetamide (IAA), formic acid (FA), acetic acid (AA), glutaraldehyde, sodium cyanoborohydride ($NaCNBH_3$) and ethanolamine were purchased from Sigma-Aldrich (St. Louis, MO). Methanol, hydrofluoric acid (HF), LC/MS grade water and acetonitrile (ACN) were purchased from Fisher Scientific (Pittsburgh, PA). Acrylamide, *N*-(3-dimethylaminopropyl)-*N′*-ethylcarbodiimide hydrochloride (EDC) and benzamidine were purchased from Acros Organics (NJ, USA). N-

hydroxysulfosuccinimide sodium salt (sulfo-NHS), urea and 4-morpholineethanesulfonic acid monohydrate (MES) were purchased from Alfa Aesar (Tewksbury, MA).

Bare fused silica capillaries (50 μm i.d., 360 μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ). Carboxyl functionalized magnetic microspheres (BioMag®Plus carboxyl) and amine functionalized magnetic microspheres (BioMag®Plus Amine) were purchased from Bangs Laboratories, Inc. (Fishers, IN). C18 spin columns were purchased from Pierce Biotechnology (Rockford, IL).

2.2.2 Preparation of LPA-coated separation capillary for CZE

The inner wall of the separation capillary was coated with linear polyacrylamide (LPA) based on prior protocols [41,42]. A bare fused silica capillary (50 μm i.d., 360 μm o.d.) was successively flushed with 1 M hydrochloric acid, water, 1 M sodium hydroxide, water, and methanol, followed by treatment with 3-(trimethoxysilyl) propyl methacrylate to introduce carbon-carbon double bonds on the inner wall of the capillary. The treated capillary was filled with degassed acrylamide solution in water containing ammonium persulfate, followed by incubation in a 50 ˚C water bath for 35 to 40 min with both ends sealed by silica rubber. After that, the capillary was flushed with water to remove the unreacted reagents. Then one end of the LPA-coated capillary was etched with HF based on prior protocol [43] to reduce its outer diameter to around 70 μm.

2.2.3 Preparation of magnetic beads-based IM

The detailed procedures for trypsin immobilization on amine and carboxyl functionalized magnetic beads have been reported [4]. Briefly, for IM-C, carboxyl groups on the magnetic microspheres were first activated with 50 mg/mL sulfo-NHS and EDC solution. Then trypsin

was immobilized on the surface of the microspheres via the reaction between amine groups on trypsin and succinimide groups on the bead surface. The remaining succinimide groups on the magnetic beads were blocked with 100 mM glycine solution. The IM-C magnetic beads were stored in 20 mM ammonium bicarbonate (pH 8.0) at 4 ^{o}C with a final concentration of 5 mg/mL. The amount of trypsin bound to the magnetic beads was about 70 µg/mg magnetic beads-based on the estimation with the bicinchoninic acid (BCA) method [4].

For IM-N, amine groups on the magnetic microspheres were first activated with a 5% glutaraldehyde solution. Trypsin was then immobilized on the surface of the microspheres in the presence of $NaCNBH_3$. Ethanolamine solution (0.4 M, pH 8.4) was used to block the unreacted aldehyde group on the microspheres after the cyanoborohydride reduction. The IM-N magnetic beads were stored in 20 mM ammonium bicarbonate (pH 8.0) at 4 ^{o}C with a final concentration of 5 mg/mL. The amount of trypsin bound to the magnetic beads was estimated by measuring the trypsin concentration in the solution before and after immobilization with the BCA method. Based on the trypsin concentration difference, the amount of trypsin immobilized on the magnetic beads was calculated. The amount of trypsin immobilized on the magnetic beads was about 70 µg/mg magnetic beads.

2.2.4 Sample preparation

Two samples were prepared for the experiment, *E. coli* and mouse brain proteome. *E. coli* (strain K-12 substrain MG1655) cells were kindly provided by Professor Heedeok Hong's group in the Department of Chemistry, Michigan State University. The *E. coli* cells were lysed in a lysis buffer containing 8 M urea, 100 mM Tris-HCl (pH 8.0) and protease inhibitors, followed by sonication in a Branson Sonifier 250 (VWR Scientific, Batavia, IL) on ice for 10 min. After centrifugation (18,000 x g for 10 min), the supernatant was collected and the protein

concentration was determined by BCA assay. The leftover protein extracts were stored at -80 ˚C before use.

The mouse brain tissue from a 6-month old male mouse (strain BL-6, wild type) was kindly provided by Professor Chen Chen's group in the Department of Animal Science, Michigan State University The whole protocol related to the mouse samples was performed following guidelines defined by the Institutional Animal Care and Use Committee of Michigan State University. The mouse brain tissue was cut into small pieces, washed with PBS to remove the blood, and suspended in 8 M urea and 100 mM NH_4HCO_3 with complete protease inhibitor, followed by homogenization with a Homogenizer 150 (Fisher Scientific) on ice and sonication with a Branson Sonifier 250 on ice for 10 min. After centrifugation (10,000 x g for 10 min), the supernatant was collected, and the protein concentration was measured with BCA assay. The supernatant was then aliquoted equally into 1.7 mL Eppendorf tubes. Each tube contained about 500 µg of protein, which were purified by acetone precipitation. The protein pellet was air dried in the chemical hood for several minutes and stored at -20 ^{o}C.

In the study for digestion performance, both samples were denatured in 8 M urea and 100 mM ammonium bicarbonate (pH 8.0) at 37 ^{o}C for 30 min, followed by reduction and alkylation with DTT and IAA. The resulting protein solution was diluted with 100 mM ammonium bicarbonate by a factor of five to produce a 1 mg/mL protein solution for experiments.

The *E. coli* proteome sample was digested by three methods. For FT digestion, 20 µL of the protein sample (20 µg of proteins) were digested at 37 ^{o}C with trypsin-to-protein mass ratio as 1:100 for 30 s, 5 min, 1 h and 14 h, respectively. After digestion, the digests were acidified immediately by adding 5 µL of 20% (v/v) FA to terminate the tryptic reaction. For digestion using IM-N and IM-C, 20 µL of the protein sample (20 µg of proteins) was added into a tube

containing IM and the proteins were digested under trypsin-to-protein mass ratio as 1:100 (w/w) as well. The mass of magnetic beads used for digestion was calculated based on the immobilization capacity of trypsin on those beads (~70 μg trypsin/mg magnetic beads). The mixture of sample and IM was vortexed for several seconds and then transferred to a 37 ^{o}C water bath. Digestion was performed with occasional vortexing at a trypsin-to-protein mass ratio of 1:100 for 30 s, 5 min, 1 h and 14 h, respectively. After digestion, the digests were acidified immediately by adding 5 μL of 20% (v/v) FA to terminate the tryptic reaction and the magnetic beads were separated from the solution by a magnet. All samples from the three digestion methods were desalted using C18 spin columns, lyophilized and dissolved in 20 μL of 2% (v/v) AA for triplicate CZE-ESI-MS/MS analysis.

The mouse brain proteome was digested by two methods. For conventional FT digestion, three aliquots of the mouse brain protein sample (20 μL of protein solution in each aliquot) were digested in parallel at 37 ^{o}C with trypsin-to-protein mass ratio of 1:30 for 12 h. 20% (v/v) FA was applied to terminate the tryptic reaction. For digestion using IM-N, three aliquots of the mouse brain protein sample (20 μL of protein solution in each aliquot) were also digested in parallel. Each sample was added into a tube containing IM-N. The mixture was vortexed for dispersion and incubated at 37 ^{o}C for 15 min for protein digestion with vortexing every 5 min. The final concentration of trypsin in the solution during IM-N digestion was 0.5 mg/mL. After digestion, 20% (v/v) FA solution was applied to terminate the tryptic reaction and the magnetic beads were separated from the solution by a magnet. All of the six samples generated by two digestion methods were desalted by C18 spin columns, lyophilized and dissolved in 20 μL of 2% (v/v) ACN, 0.1% (v/v) FA prior to nanoLC-ESI-MS/MS analysis in duplicate.

Another two aliquots of the mouse brain proteome sample (300 μg of proteins in each aliquot) were also digested by FT and IM-N, respectively. The procedures were the same as those described in the previous paragraph. After acidification with FA, the 300 μg of mouse brain proteome digests from FT and IM-N were directly fractionated by high-pH RPLC, followed by low-pH RPLC-MS/MS analysis.

For the experiment of high-throughput bottom-up proteomics, the mouse brain proteome sample (30 μg) was dissolved in a buffer containing 8 M urea, 6.6 mM DTT, 100 mM ammonium bicarbonate (pH 8.0) and was kept at 37 ^{o}C for 30 min for denaturation and reduction, followed by alkylation with IAA (16.5 mM) for 10 min at room temperature in dark. The sample was then diluted with 100 mM ammonium bicarbonate by a factor of four to reach 1 mg/mL protein concentration. The protein sample (30 μg) was added into a tube containing IM-N. The mixture was vortexed for dispersion and incubated at 37 ^{o}C for 15 min for protein digestion with vortexing every 5 min. The final concentration of trypsin in the solution during IM-N digestion was 0.5 mg/mL. After digestion, 20% (v/v) FA solution was applied to terminate the tryptic reaction and the magnetic beads were separated from the solution by a magnet. The digest was desalted by a C18 spin column and lyophilized with a vacuum concentrator in 1 h. The sample preparation described above was repeated for three times and got three peptide samples. The peptide samples were dissolved in 30 μL of 20 mM ammonium bicarbonate (pH 8.0) and analyzed by CZE-ESI-MS/MS in triplicate. Each CZE-MS/MS run took 30 min.

2.2.5 High-pH RPLC fractionation of mouse brain proteome digests

An Agilent Infinity II HPLC system and a C18 RP column (Zorbax 300Extend-C18, 2.1 mm i.d. × 150 mm length，3.5 μm particles, Agilent Technologies) were used for peptide separation.

Buffer A (5 mM NH_4HCO_3, pH 9) and buffer B (5 mM NH_4HCO_3 containing 80% ACN, pH 9) were used as mobile phase to generate gradient for separation. The flow rate was 0.3 mL/min. The peptide samples were loaded onto the RPLC column for 5 min at 2% B. Then the peptides were separated by gradient elution: 2 min from 2% B to 10% B, 60 min from 10% B to 50% B, and 2 min from 50% B to 100% B. The mobile phase was kept at 100% B for 10 min, followed by column equilibration with 2% B for 10 min.

Fractions were collected at a rate of one fraction/min from 7 min to 67 min. In total 60 fractions were collected from each sample. Fraction number "N" and fraction number "N+30" were combined, thus leading to 30 fractions totally. The fractions were lyophilized and dissolved in 10 µL of 2% (v/v) ACN, 0.1% (v/v) FA for nanoLC-ESI-MS/MS analysis.

2.2.6 CZE-ESI-MS/MS and nanoLC-ESI-MS/MS

An ECE-001 capillary electrophoresis autosampler (CMP Scientific, Brooklyn, NY), a commercialized electrokinetically pumped sheath flow interface (CMP Scientific) [44,45] and a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) were used for CZE-ESI-MS/MS. One power supply integrated with the autosampler was used for CZE separation. Another power supply from CMP Scientific (Brooklyn, NY) was employed for electrospray. The orifice of the electrospray emitter was around 30 µm. The distance from the electrospray emitter orifice to the mass spectrometer entrance was around 2 mm.

For the experiment of *E. coli*, the total length of the LPA-coated capillary for CZE separation was 70 cm. The background electrolyte (BGE) of CZE was 5% (v/v) AA in water and the sheath buffer was 0.2% (v/v) FA containing 10% (v/v) methanol. The sample was injected into the separation capillary by air pressure (3 psi, 8 s). The voltage applied at the injection end was 25

kV and the voltage for electrospray was around 2 kV. The separation time for each run was 60 min, including 10 min flushing of BGE at 10 psi at the end of the separation.

For the experiment of high-throughput bottom-up proteomics of the mouse brain proteome, the total length of the LPA-coated capillary for CZE separation was 60 cm. The sample was injected for 6 s with 5 psi air pressure. The voltage applied at the injection end was 20 kV. The BGE of CZE was 5% (v/v) AA in water and the sheath buffer was 0.2% (v/v) FA containing 10% (v/v) methanol. The separation time for each run was 30 min, including 5 min flushing of BGE at 10 psi at the end of the separation.

An EASY-nLC™ 1200 System (Thermo Fisher Scientific) was used for the separation of the digests. Mobile phases for gradient separation were Buffer A (0.1% FA in 2% ACN) and buffer B (0.1% FA in 80% ACN). A commercial C18 reversed-phase column (Acclaim™ PepMap™, 75 μm i.d. × 50 cm, particle size 2 μm, pore size 100Å) was used as a separation column and a commercial C18 reversed-phase column (Acclaim™ PepMap™, 75 μm i.d. × 2 cm, particle size 3 μm, pore size 100Å) was used as a trap column. For mouse brain sample without fractionation, a 2-μL sample was loaded on the trap column by 100% A at a flow rate of 20 μL/min, followed by a 150-min gradient separation at a flow rate of 200 nL/min: 120 min from 8% B to 50% B, 2 min to 100% B, and maintained for 28 min. Each sample was analyzed in duplicates. For fractionated mouse brain sample, a 3-μL sample from each fraction was loaded by 100% A at a flow rate of 20 μL/min, followed by a 90-min gradient separation at a flow rate of 200 nL/min: 70 min from 8% B to 40% B, 2 min to 100% B, and maintained for 18 min. The column was equilibrated by 5 μL of 100% buffer A prior to the next sample analysis.

The parameters of Q Exactive-HF mass spectrometer (Thermo Fisher Scientific) were as follows. The ion transfer tube temperature was 320 ˚C and the S-Lens RF level was 55.0. Full

MS scans were acquired in the Orbitrap mass analyzer over the *m/z* 300–1500 range with resolution as 60,000 at *m/z* 200 and AGC target value of 3.00E+06. The twenty most intense peaks with charge state from 2 to 6 were isolated in the quadrupole with the isolation window of 1.5 *m/z*. The normalized collision energy was set as 28% for precursor fragmentation in the high energy collisional dissociation (HCD) cell. The tandem mass spectra were acquired in the Orbitrap mass analyzer with resolution of 15,000 at *m/z* 200 and AGC target value of 1.00E+05. The ion selection threshold intensity was 1.0E+05, and the maximum times of accumulating ions per scan event were 50 ms for full MS scans and 25 ms for tandem mass spectra. Peptide match and exclude isotopes were set on and dynamic exclusion was set to 30 s.

For the high-throughput bottom-up proteomics experiment with the mouse brain proteome, the parameters of the Q Exactive-HF mass spectrometer were the same as above except for those mentioned below. Full MS scans were acquired over the *m/z* 300-1800 range. The twenty most intense peaks with charge state higher than 1 were isolated in the quadrupole with the isolation window of 2 *m/z*. The tandem mass spectra were acquired with resolution of 30,000 (at *m/z* 200). The ion selection threshold intensity was 5.0E+04, and the maximum ion injection time per scan event was 50 ms for both full MS and MS/MS scans.

2.2.7 Data analysis

Raw MS files were analyzed by MaxQuant [46] version 1.3.0.5 software. MS/MS spectra were searched using the Andromeda search engine [47]. The UniProt *Escherichia coli* (strain K12) database containing forward and reverse sequences and common contaminants were used for *E. coli* data analysis. The UniProt *Mus musculus* database containing forward and reverse sequences and common contaminants was used for mouse brain data analysis. MaxQuant analysis included a first search peptide mass tolerance of 20 ppm, main search peptide mass

tolerance of 6 ppm, and fragment mass tolerance of 20 ppm. The search included full tryptic digestion, cysteine carbamidomethylation as a fixed modification and methionine oxidation, N-terminal acetylation and deamidation (NQ) as variable modifications. The minimum peptide length was set to seven amino acids. The false discovery rate (FDR) for both peptide and protein identifications was set to 0.01. The "match between runs" function was turned on with the time window as 0.7 min. For analysis of the mouse brain proteome samples digested by IM-N and FT, the label-free quantification (LFQ) function integrated in MaxQuant was enabled [48].

The RAW files from 2D-LC-MS/MS analysis of the mouse brain proteome digests prepared using IM-N were also analyzed using Proteome Discoverer 2.1 software (Thermo Fisher Scientific) with Sequest HT database search engine against *Mus musculus* databases downloaded from UniProt (http://www.uniprot.org/). The reversed database search was also performed to evaluate the FDR. The MS/MS spectra were firstly filtered with the top 12 peaks in the mass window of 100 Da. The database searching parameters included two maximum missed cleavage sites for fully specific tryptic digestion, precursor mass tolerance as 20 ppm and fragment mass tolerance as 0.05 Da. The dynamic modification was oxidation (M) and acetyl (N-Terminus), and the static modification was carbamidomethyl (C). Peptide identifications were filtered with peptide confidence value as high, corresponding to less than 1% FDR on peptide level. The results from Proteome Discoverer 2.1 software were only used for the prediction of the number of transmembrane domains (TMDs) of proteins with the TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) algorithm.

Perseus software [49] (version 1.6.0.7) was used for further analysis of the MaxQuant results. Peptide isoelectric point (pI) and grand average of hydropathy (GRAVY) values were calculated by Compute pI/Mw tool (http://web.expasy.org/compute pi/), and GRAVY CALCULATOR

(http://www.gravy-calculator.de). Biological process, cellular components and molecular functions were assigned based on Gene Ontology using DAVID Bioinformatics Resources 6.8 [50,51].

2.3 Results and discussion

To investigate how the solid matrix of IM influences its preference for protein cleavage comparing with FT, two types of IM were prepared based on reference 4, as shown in **Figure 2.1A**. IM-C was prepared with carboxyl group functionalized magnetic beads, and the remaining succinimide groups on the bead surface were blocked with glycine. Therefore, the solid matrix surface of IM-C is negatively charged at pH 8 due to the immobilized glycines' carboxyl groups. IM-N was prepared with amine group functionalized magnetic beads, and the remaining aldehyde groups were blocked with ethanolamine, resulting in the nearly neutral solid matrix at pH 8. As shown in **Figure 2.1B**, experiments 1 and 2 were performed to understand the protein cleavage catalyzed by IM. To investigate how well IM can perform for digestion of complex proteomes for deep proteomics compared with FT, a mouse brain proteome sample was digested with both IM-N and FT, followed by 2D-LC-MS/MS analysis, experiment 3. Finally, IM-N based rapid protein digestion was coupled to fast CZE-MS/MS for high-throughput characterization of the mouse brain proteome (experiment 4).

2.3.1 Investigation of protein cleavage preference catalyzed by IM

E.coli proteome samples (20 µg of proteins) were digested by FT, IM-C and IM-N for four different periods (30 s, 5 min, 1 h and 14 h), followed by single-shot CZE-MS/MS analysis in triplicate, **Figure 2.1B**. The same trypsin-to-protein mass ratio (1:100) was employed for all

digestion methods. Therefore, accurate investigation could be performed to answer the question how immobilization of trypsin affects the preference of trypsin-catalyzed cleavage.

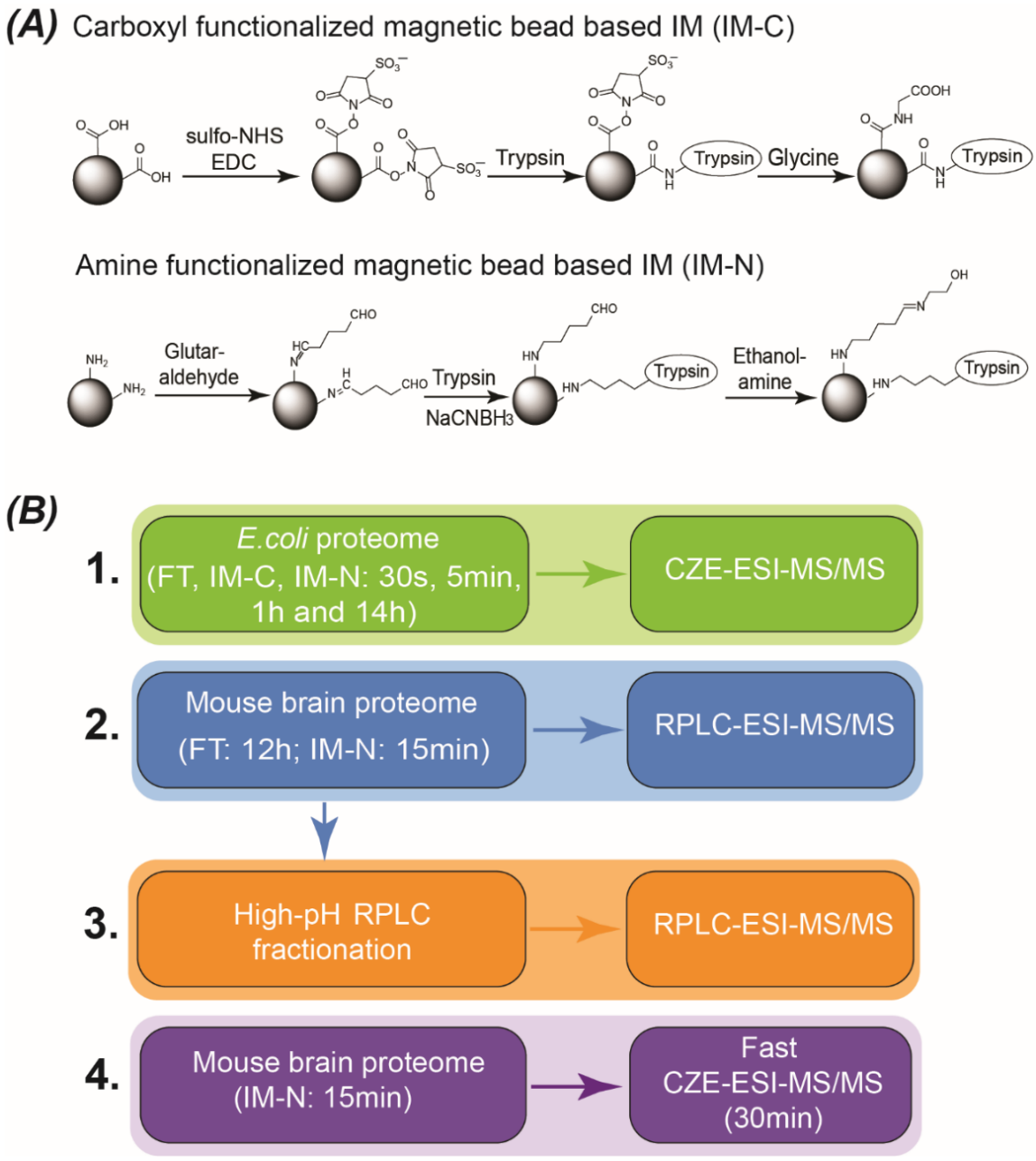

Figure 2.1. (A) Synthesis of carboxyl functionalized magnetic bead-based IM (IM-C) and amine functionalized magnetic bead-based IM (IM-N). (B) Experimental design of the work.

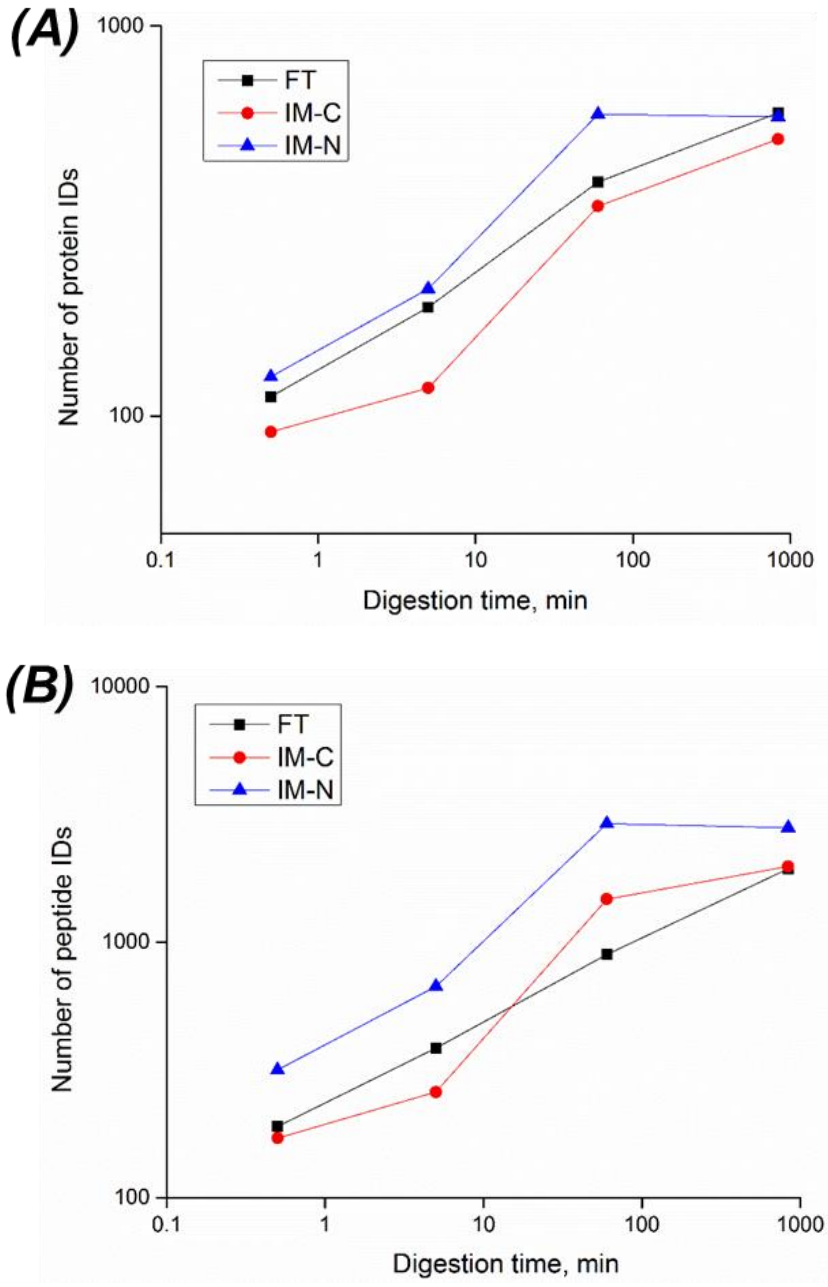

Figure 2.2. Log-log plots of (A) number of protein identifications (IDs) vs. digestion time (Top) and (B) number of peptide IDs vs. digestion time (Bottom) for the three digestion methods, FT, IM-C, and IM-N. The *E. coli* cell lysate was used for these experiments. The number of IDs was from the combined results of triplicate CZE-MS/MS runs.

IM-N had better digestion performance than IM-C based on the numbers of protein and peptide IDs in each digestion time as well as the missed cleavage distributions, **Figures 2.2 and 2.3**. One reason is that IM-N has a longer spacer arm between trypsin and solid matrix than IM-C (**Figure 2.1A**). The longer spacer arm can avoid steric hindrance and allow the trypsin on beads to stretch and catch substrates more easily [52], thus leading to faster and more complete digestion. Another possible reason is that the negatively charged proteins in the sample at pH 8 have difficulty approaching the IM-C surface due to electrostatic repulsion. IM-C tended to identify more basic proteins compared with IM-N in short digestion periods (30 s, 5 min, and 1 h), **Figure 2.4**.

IM-N showed better digestion performance than FT in each digestion period concerning the numbers of peptide/protein IDs and the missed cleavage distributions, **Figures 2.2 and 2.3**. The number of protein IDs from 1-h IM-N digestion was almost the same as that from 14-h FT digestion, suggesting that IM-N could digest proteins faster than FT. Because the same trypsin-to-protein mass ratio (1:100) was used for all of the experiments, the difference between IM-N and FT in digestion performance is most likely due to the immobilization of trypsin. In FT digestion, trypsin is consumed by auto-digestion. One tryptic peptide was clearly detected from the *E. coli* sample that was digested by FT for only 30 s, indicating that the auto-digestion of trypsin happened very fast. However, no significant signals of trypsin peptides was observe in the *E.coli* samples that were digested by IM-N for 30 s or 5 min, suggesting that immobilization of trypsin in IM-N greatly reduced the auto-digestion of trypsin. Therefore, IM-N can achieve better digestion performance than FT. FT and IM-N have no significant difference in the cumulative distribution of the pI of identified proteins, **Figure 2.4**, which is due to the nearly neutral solid matrix surface of IM-N.

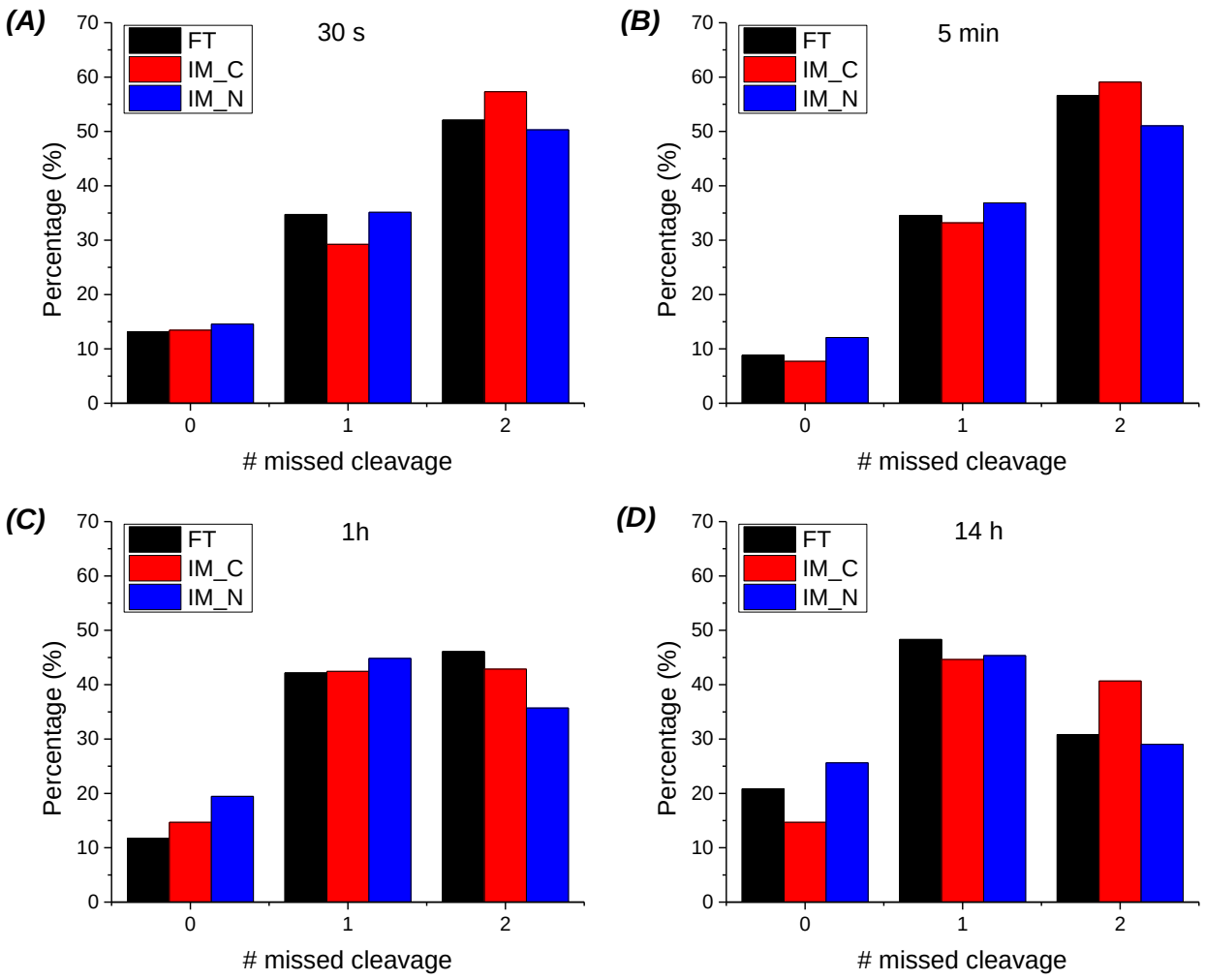

Figure 2.3. The number of missed cleavages on the peptides from IM-C, IM-N and FT digestion of the *E. coli* proteome across four different digestion periods (30 s, 5 min, 1 h and 14 h).

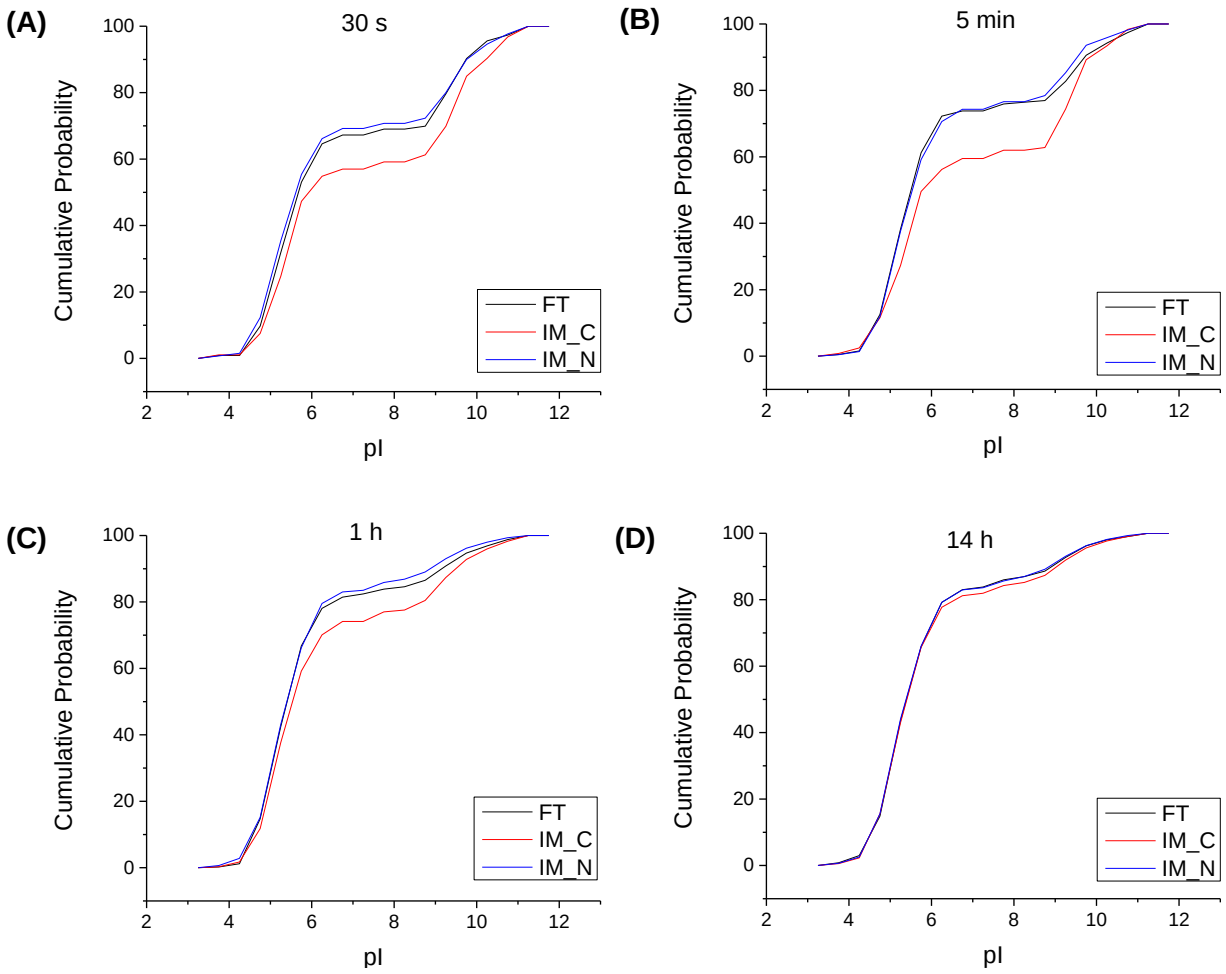

Figure 2.4. Cumulative distribution of the pI of identified proteins from FT, IM-C and IM-N digestion of the *E. coli* proteome in different digestion periods (30 s, 5 min, 1 h and 14 h).

Then the microenvironment surrounding the cleavage sites (K/R) was investigated to better our understanding of the cleavage preference catalyzed by FT, IM-N, and IM-C. The identified peptides from the four digestion periods were classified into two types, early-generated peptides, and late-generated peptides, based on the intensity change of peptides as a function of digestion time, **Figure 2.5**. Early-generated peptides were defined as peptides that appeared in the first digestion period (30 s) and had no continuous increase in intensity in longer digestion periods (**Figure 2.5, panels A-C**). The early-generated peptides contained cleavage sites (K/R) that were

cleaved quickly. Late-generated peptides were defined as peptides that had an at least 5-times continuous increase in intensity as digestion time increased (**Figure 2.5, panels D-F**). Those peptides contained cleavage sites (K/R) that were cleaved slowly.

The microenvironment surrounding the cleavage sites (K/R) for FT, IM-C, and IM-N were compared based on those early-generated peptides and late-generated peptides, **Figure 2.6**. Compared with the early-generated peptides, the cleavage sites for the late-generated peptides tended to be surrounded by more acidic amino acids (D/E). The results from FT, IM-C and IM-N digestion agreed reasonably with each other, suggesting that the immobilization of trypsin on the solid matrixes studied here did not significantly influence the cleavage preference of trypsin molecules. For the first time, the protein cleavage preference catalyzed by the IM was investigated using a complex proteome sample and quantitative proteomics.

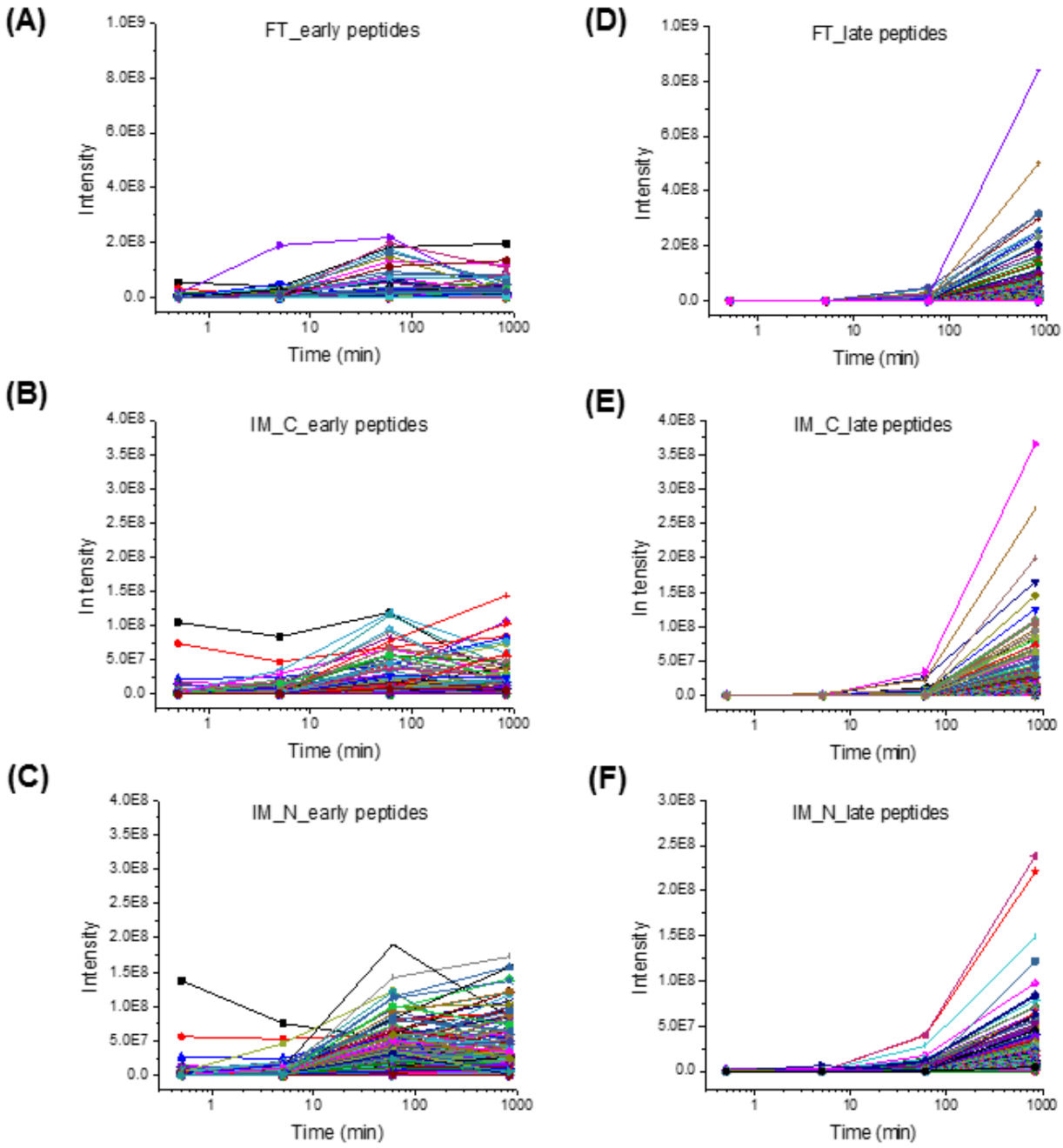

Figure 2.5. Intensity trend of early peptides and late peptides generated by FT, IM-C and IM-N digestion of the *E. coli* proteome. Each color represents a peptide. About 190, 170, and 310 early-generated peptides were determined for FT, IM-C, and IM-N, respectively; about 1380, 1280 and 760 late-generated peptides were determined for FT, IM-C, and IM-N, respectively.

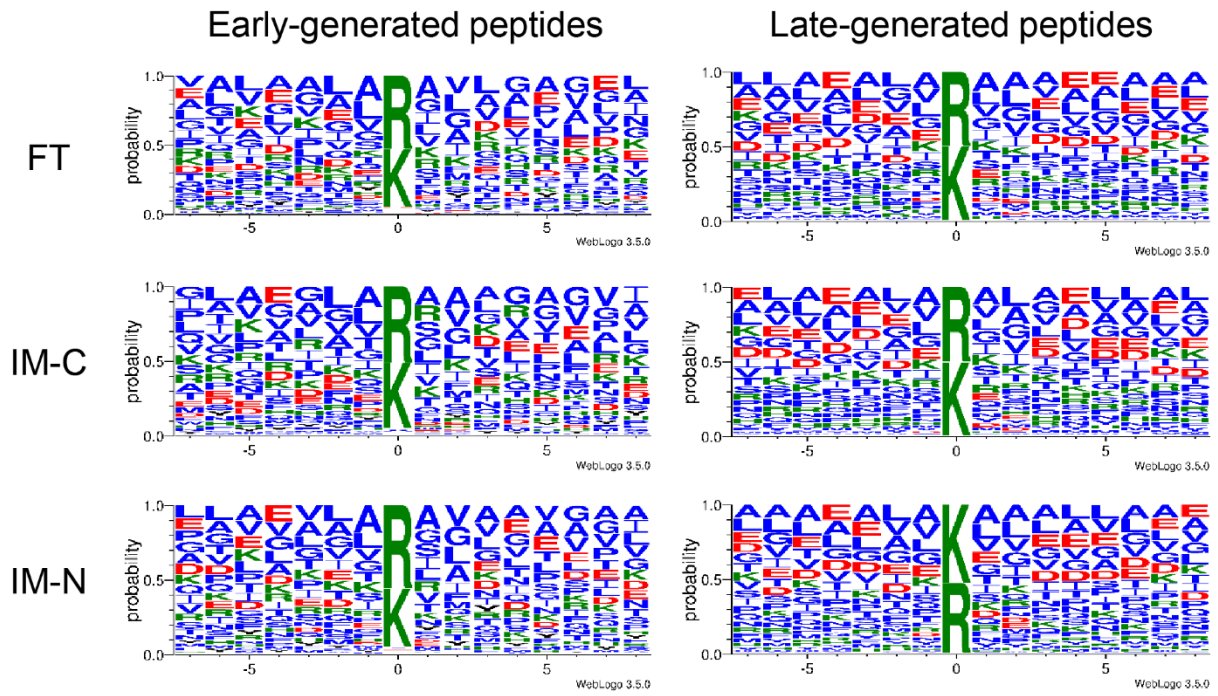

Figure 2.6. Sequence logos of the cleavage sites for early- and late-generated peptides from FT, IM-C, and IM-N digestion of the *E. coli* proteome. WebLogo software (http://weblogo.threeplusone.com/) was used to generate the sequence logos. For the x-axis, "0" represents the cleavage site; -7 to -1 represent the left amino acids; 1 to 8 represent the right amino acids. The y-axis represents the probability.

2.3.2 Reproducibility of IM-N for fast digestion of a mouse brain proteome sample

The reproducibility of IM-N for fast digestion of a mouse brain proteome sample was further investigated. Three protein samples were prepared in parallel as replicates using IM-N and FT. For IM-N digestion, each protein sample (20 µg of proteins) was mixed with IM-N for digestion at 37 °C for 15 min. The trypsin concentration in the solution during IM-N digestion was 0.5 mg/mL, and the trypsin-to-protein mass ratio was 1:2. For FT digestion, each sample (20 µg of

proteins) was digested with FT at 37 ^{o}C for 12 h under trypsin-to-protein mass ratio of 1:30. All of the six proteome digests were analyzed by RPLC-ESI-MS/MS in duplicate.

Table 2.1. Protein group and peptide identifications (# protein groups/ # peptides) and overlap between duplicated LC-MS runs (protein overlap/peptide overlap) from mouse brain proteome samples prepared by IM-N and FT in triplicate.

	1st run	2nd run	Combined	Overlap of two runs (%)
FT_1	2553/ 14799	2449/ 12877	2721/ 20003	84/ 64
FT_2	2535/ 14374	2528/ 13833	2721/ 18983	86/ 65
FT_3	2470/ 13345	2429/ 12970	2658/ 16812	84/ 72
IM_N_1	2536/ 15166	2538/ 14928	2672/ 18414	90/ 81
IM_N_2	2469/ 14260	2447/ 14031	2657/ 17733	85/ 72
IM_N_3	2487/ 14627	2446/ 14249	2709/ 18772	82/ 67

Table 2.2. Overlaps of protein group and peptide identifications (protein overlap (%)/ peptide overlap (%)) between IM-N and FT digestion from the mouse brain proteome samples.

	FT_1	FT_2	FT_3
IM_N_1	88/ 77	88/ 75	86/ 68
IM_N_2	88/ 71	87/ 70	86/ 67
IM_N_3	88/ 76	88/ 75	86/ 68

Tables 2.1 and **2.2** show the summary of protein and peptide IDs from FT and IM-N digestion. IM-N and FT digestion generated comparable numbers of protein and peptide IDs from duplicate LC-MS runs (2679±27 *vs.* 2700±36 proteins; 18306±528 *vs.*18599±1630 peptides), **Table 2.1**. Both FT and IM-N digestion were reproducible regarding the numbers of protein and peptide IDs. In addition, IM-N and FT digestion yielded the same pools of proteins and peptides. As shown in **Table 2.2**, the overlaps of protein and peptide IDs between FT and

IM-N digestion are comparable with that from duplicate LC-MS analysis of one sample. The identified peptides from FT and IM-N digestion have almost the same cumulative distributions of the pI and grand average of hydropathy (GRAVY) values as well as the same distributions of the number of missed cleavages, **Figure 2.7**.

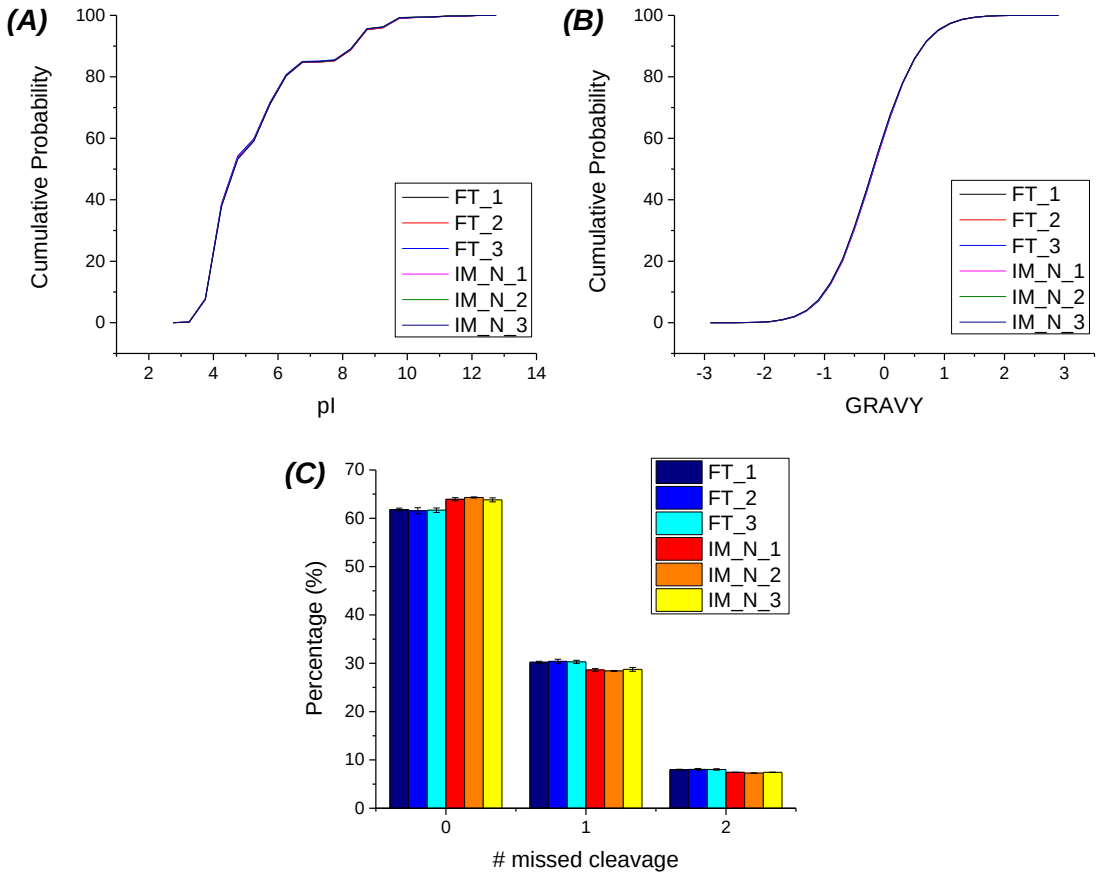

Figure 2.7. The properties of the identified peptides from the mouse brain proteome using FT and IM-N digestion. The cumulative distributions of the pI (A) and GRAVY values (B) of identified peptides; the distributions of the number of missed cleavages on the peptides (C).

The FT and IM-N digestion was further quantitatively evaluated based on the label-free quantification (LFQ) protein intensity from MaxQuant database search [46, 48]. The Perseus software was used for data analysis [49]. Good correlations of LFQ intensity were observed

among triplicate sample preparations using FT and IM-N ($r \geq 0.998$), **Figure 2.8**, indicating that both digestion methods were quantitatively reproducible. The LFQ intensity of proteins from FT and IM-N digestion also agreed well ($r \geq 0.989$), **Figure 2.8**. As shown in **Figure 2.9,** almost no proteins have significantly different LFQ intensity between replicate preparations using FT or IM-N digestion **(panels A and B)**. Only 90 out of 1488 quantified protein groups show significantly different LFQ intensity between IM-N and FT digestion, **Figure 2.9 (panel C)**. Those results further indicated that IM-N could yield very similar digestion performance to FT for complex proteomes.

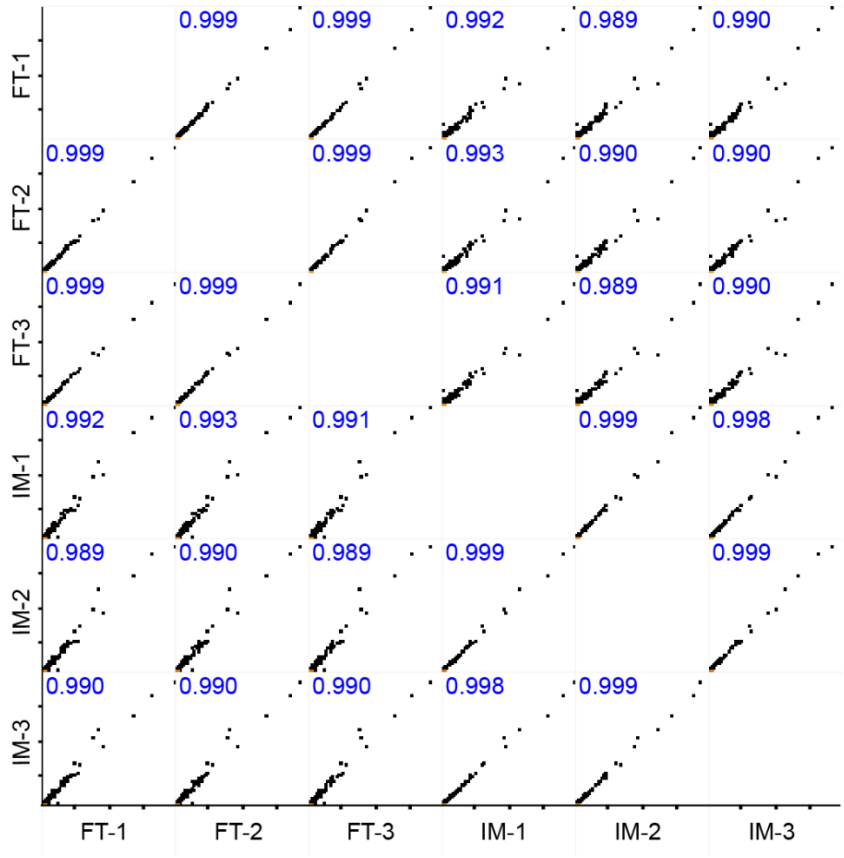

Figure 2.8. Multi-scatter correlations of protein LFQ intensity from triplicate preparations of mouse brain proteome with IM-N and FT digestion. Pearson correlation (r) values were labeled. Perseus software (version 1.6.0.7) was used to generate the correlations [49].

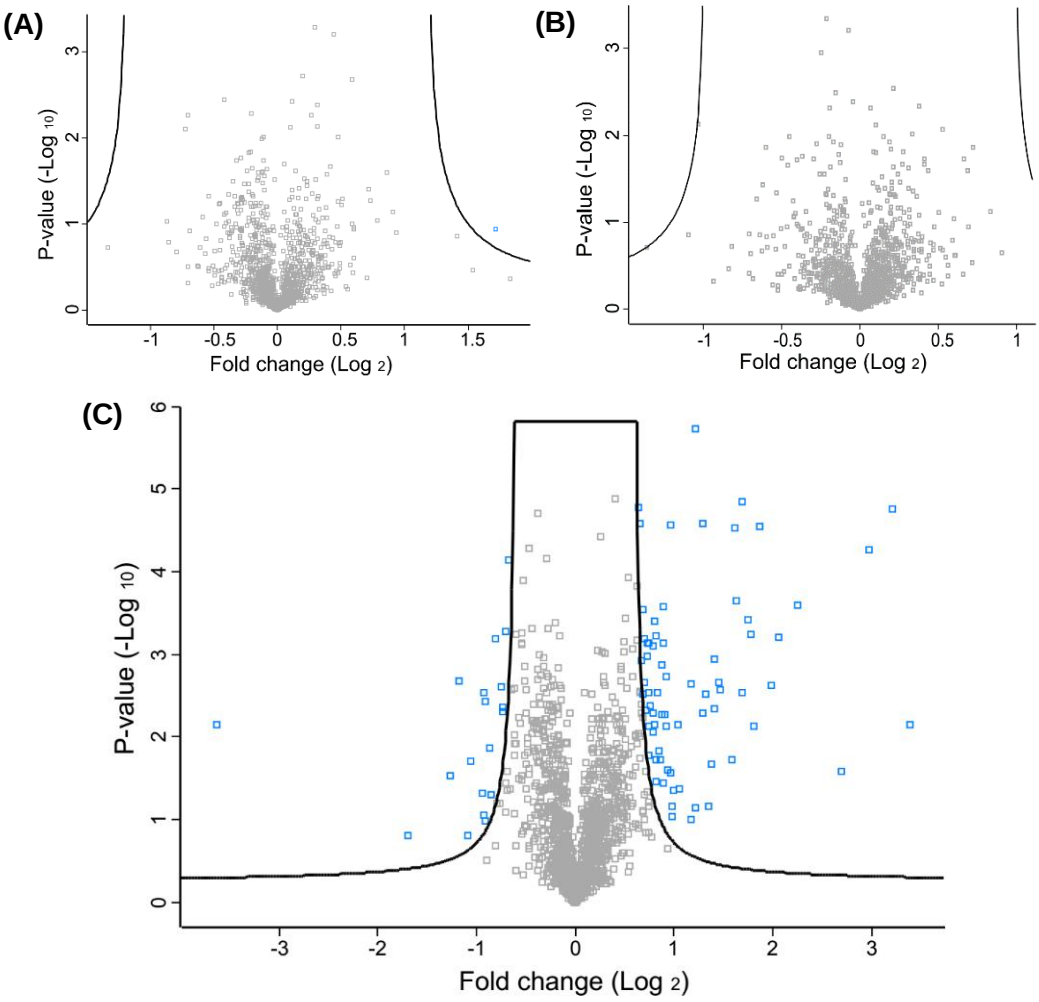

Figure 2.9. Volcano plots of the fold change (Log_2) of protein LFQ intensity (x-axis) and the P-value ($-Log_{10}$) of quantified proteins (y-axis). Comparison of protein LFQ intensity from two preparations of the mouse brain proteome sample with FT digestion (A), from two preparations with IM-N digestion (B), and from IM-N and FT digestion (C) were performed. Each spot in the figures represents a quantified protein group. Perseus software [10] (version 1.6.0.7) was used to generate the volcano plots with the following parameters: the FDR value as 0.05 and the s0 value as 1. The protein groups having significantly different protein LFQ intensity between the two conditions were marked in blue color.

2.3.3 IM-N based fast protein digestion for deep bottom-up proteomics

In order to determine how well IM can perform for deep bottom-up proteomics compared to FT, we employed two-dimensional LC-ESI-MS/MS to analyze the mouse brain proteome digests (300 µg) from the IM-N digestion (15 min) and FT digestion (12 h).

FT and IM-N digestion produced similar numbers of protein group and peptide IDs, 8716 *vs.* 8733 proteins and 96377 *vs.* 103662 peptides. This is the largest proteomic dataset using IM based fast protein digestion reported to date. Recently, Sharma *et al* reported that nearly 13000 transcripts were detected using RNA sequencing from mouse brain with common filtering criteria [53]. Our work using IM-N covered nearly 70% of the mouse brain proteome, which clearly suggests that deep proteome coverage can be approached using IM-N based fast protein digestion. More importantly, FT and IM-N approached the same pool of proteins, which was demonstrated by the 93% protein-level overlap. Although the peptide-level overlap is relatively lower compared with the peptide overlaps in **Table 2.1** and **Table 2.2** (60% *vs.* 64%-81%), we did not observe significant differences of identified peptides from IM-N and FT in the peptide pI, peptide GRAVY and the number of the missed cleavages on peptides, **Figure 2.10** (**panels A-C**).

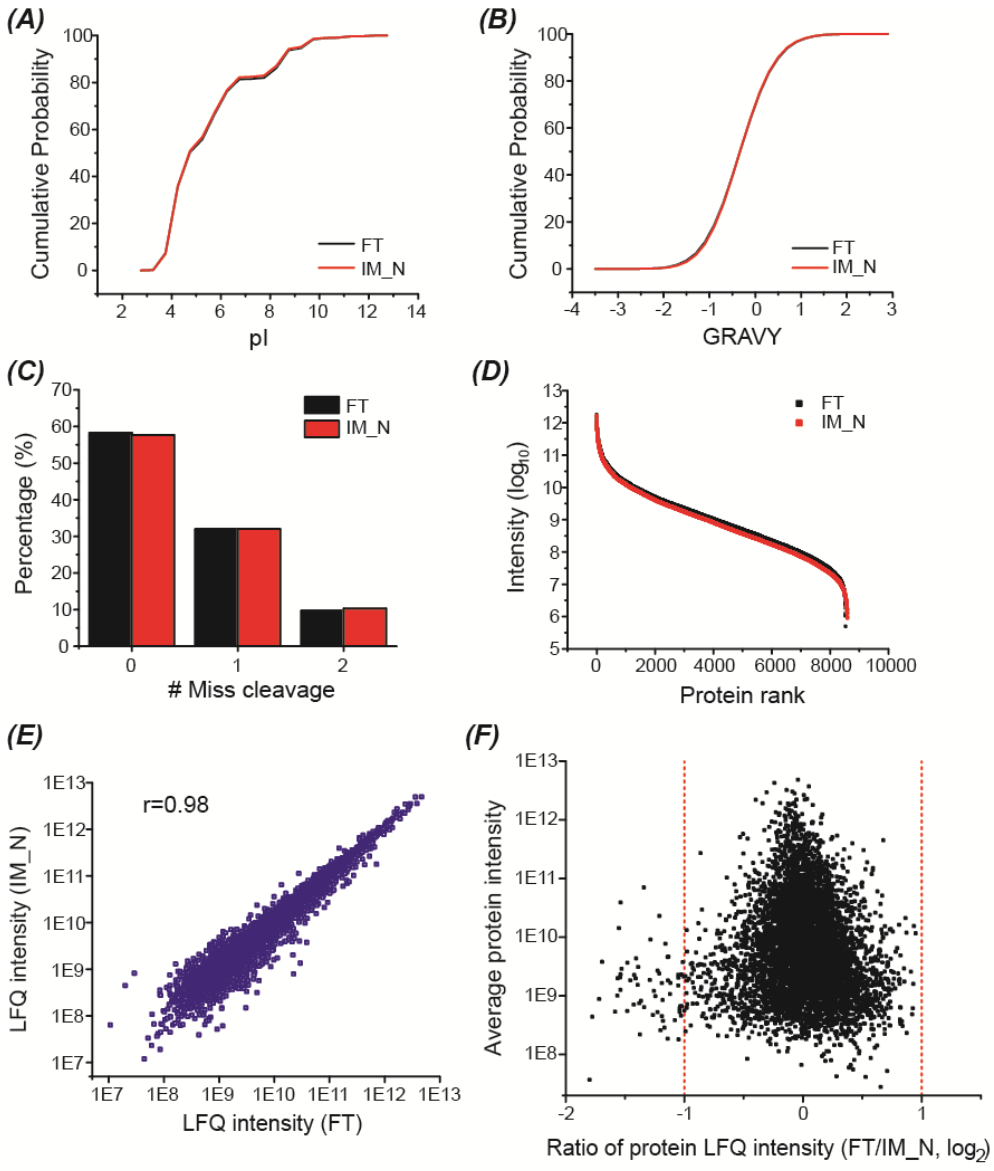

Figure 2.10. Data analysis of identified peptides and protein groups from the mouse brain proteome sample digested by FT and IM-N after analyzed by 2D-LC-MS/MS. (A) Cumulative distribution of the peptide pIs. (B) Cumulative distribution of the GRAVY values of peptides. (C) Distributions of the missed cleavages on peptides. (D) Dynamic range of observed proteome using FT and IM-N. (E) Log-log correlation of protein LFQ intensity between FT and IM-N. (F) Comparison of the protein LFQ intensity from FT and IM-N. The average LFQ intensity of each protein from FT and IM-N vs. the ratio of protein LFQ intensity between FT and IM-N (Log2).

We further analyzed the dynamic range of the observed mouse brain proteomes from IM-N and FT digestion and compared the LFQ intensity of 6099 proteins quantified from IM-N and FT digestion, **Figure 2.10** (**panels D-F**). IM-N and FT digestion both yielded close to 6.5 orders of magnitude proteome dynamic range, **Figure 2.10D**. The protein LFQ intensity from FT and IM-N agreed well across the complete dynamic range of the observed proteome (r=0.98), **Figure 2.10E**. The data suggest that IM-N has no bias in the digestion of low abundant proteins compared to FT. Around 99% of the quantified proteins had less than 2-fold differences in LFQ intensity between FT and IM-N, **Figure 2.10F**. The results clearly indicate that IM-N (15 min) can perform as well as FT (12 h) for the digestion of complex proteomes qualitatively and quantitatively.

Compared to FT, IM-N had no bias in the digestion of proteins that were involved in various biological processes, were located in different components of cells, and had diverse functions, **Figure 2.11** (**panels A-C**). As shown in **Table 2.3**, around 50% or higher of the proteins related to the selected biological processes, cellular components, and molecular functions in the UniProt mouse database were covered by the proteome dataset from IM-N digestion. The result further indicates that IM-N based fast protein digestion can approach very deep proteome coverage.

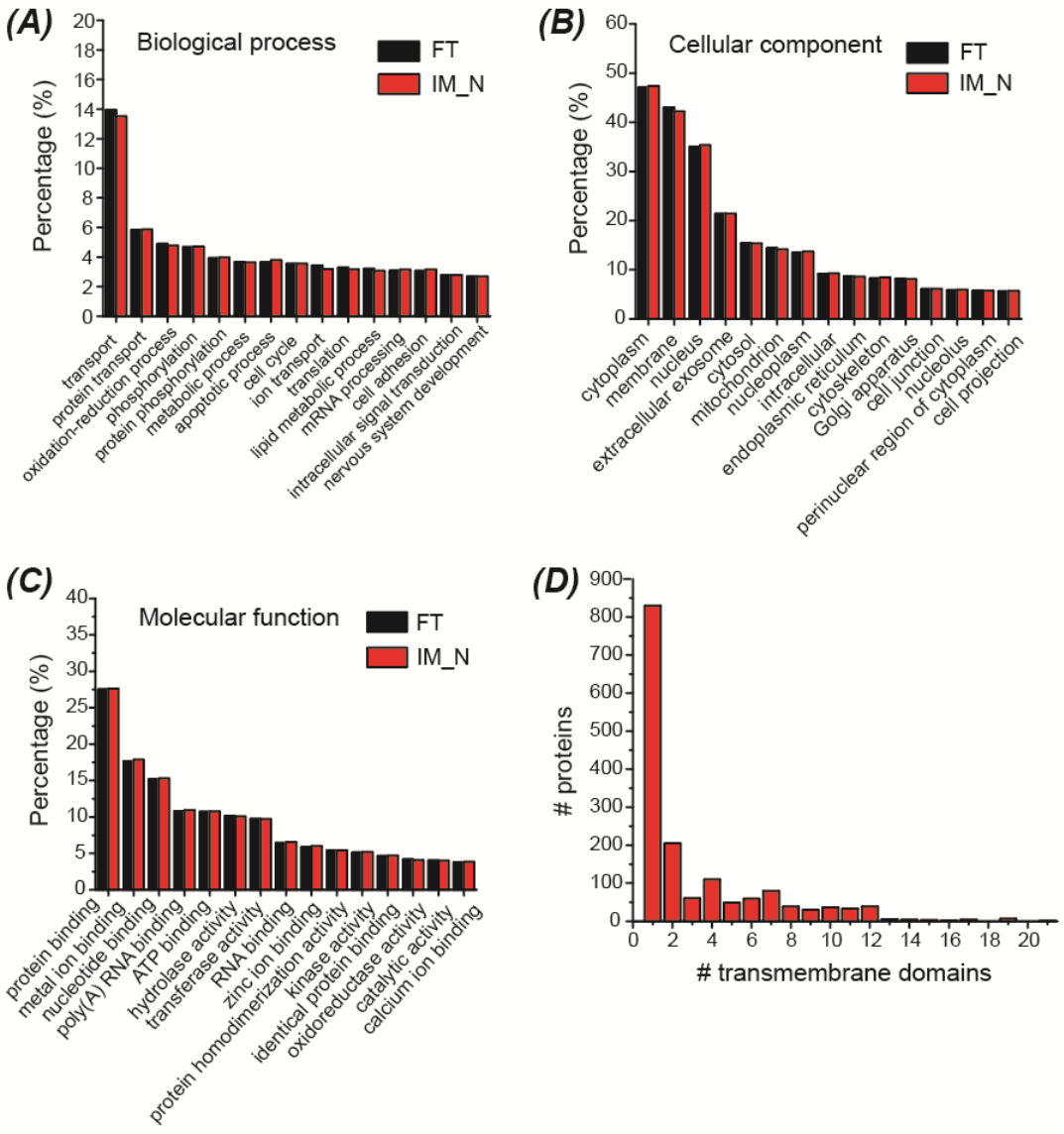

Figure 2.11. GO analysis of the identified proteins from the mouse brain proteome sample (A-C). The distribution of the transmembrane domains of identified proteins using IM-N digestion (D). DAVID Bioinformatics Resources 6.8 was used for the GO analysis. TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) algorithm was used for the prediction of the number of transmembrane domains based on the protein sequences.

Table 2.3. Summary of selected GO information of all proteins in UniProt *Mus musculus* database and the identified proteins from the mouse brain proteome sample digested by IM-N using 2D-LC-MS/MS.

Category	Term	# proteins from UniProt database (A)	# proteins from IM-N (B)	B/A (%)
Biological process	Ion transport	583	261	45
	Cell cycle	611	290	47
	Nervous system develop	377	220	58
Cellular component	Membrane	6951	3447	50
	Nucleus	5849	2889	49
	Mitochondrion	1684	1157	69
Molecular function	Zinc ion binding	1068	492	46
	Kinase activity	674	423	63

We identified 3447 membrane proteins using IM-N digestion, which is 50% of all the annotated membrane proteins in the UniProt *Mus musculus* database, **Table 2.3**. The data represents the first example of fast IM digestion for deep membrane proteomics. 1549 identified proteins have at least one transmembrane domain (TMD). The number of predicted TMDs on those proteins ranged from 1 to 21, **Figure 2.11**. Chen *et al.* identified 1897 membrane proteins from a rat brain lysate using FT digestion and 2D-LC-MS/MS [54]. Wiśniewski *et al.* identified 2700 membrane proteins from mouse hippocampus using FT digestion and 2D-LC-MS/MS [55]. Very recently, Zhao *et al.* performed deep membrane proteomics of HeLa cells using FT digestion and 2D-LC-MS/MS [38]. They identified 3785 membrane proteins from HeLa cells, representing the largest membrane protein dataset from human cell lines. Overall, the results here support that IM-N could perform as well as FT for digestion of hydrophobic membrane proteins.

2.3.4 Coupling IM-N based protein digestion to CZE-MS/MS for high-throughput bottom-up proteomics

We developed a high-throughput bottom-up proteomics workflow encompassing the protein sample pretreatment (denaturation, reduction and alkylation) in 40 min, protein digestion with IM-N in 15 min, desalting and lyophilization of the peptides in 1 h, peptide analysis with CZE-MS/MS in 30 min, and data analysis for protein ID in 30 min, **Figure 2.12A**. In total, this workflow only required ~3 h. This fast workflow enabled the identification of over 1000 proteins and 6000 peptides from the mouse brain proteome in only 3 h and with good qualitative and quantitative reproducibility, **Figure 2.12B** and **Figure 2.13**.

Much effort has been made to improve the throughput of proteomic sample preparation for bottom-up proteomics in order to facilitate fundamental research and clinical diagnostics [56-57]. However, hours of tryptic digestion were required in those studies. We believe the high-throughput bottom-up proteomics workflow comprising IM-N based rapid protein digestion and fast CZE-MS/MS analysis will benefit many clinical applications.

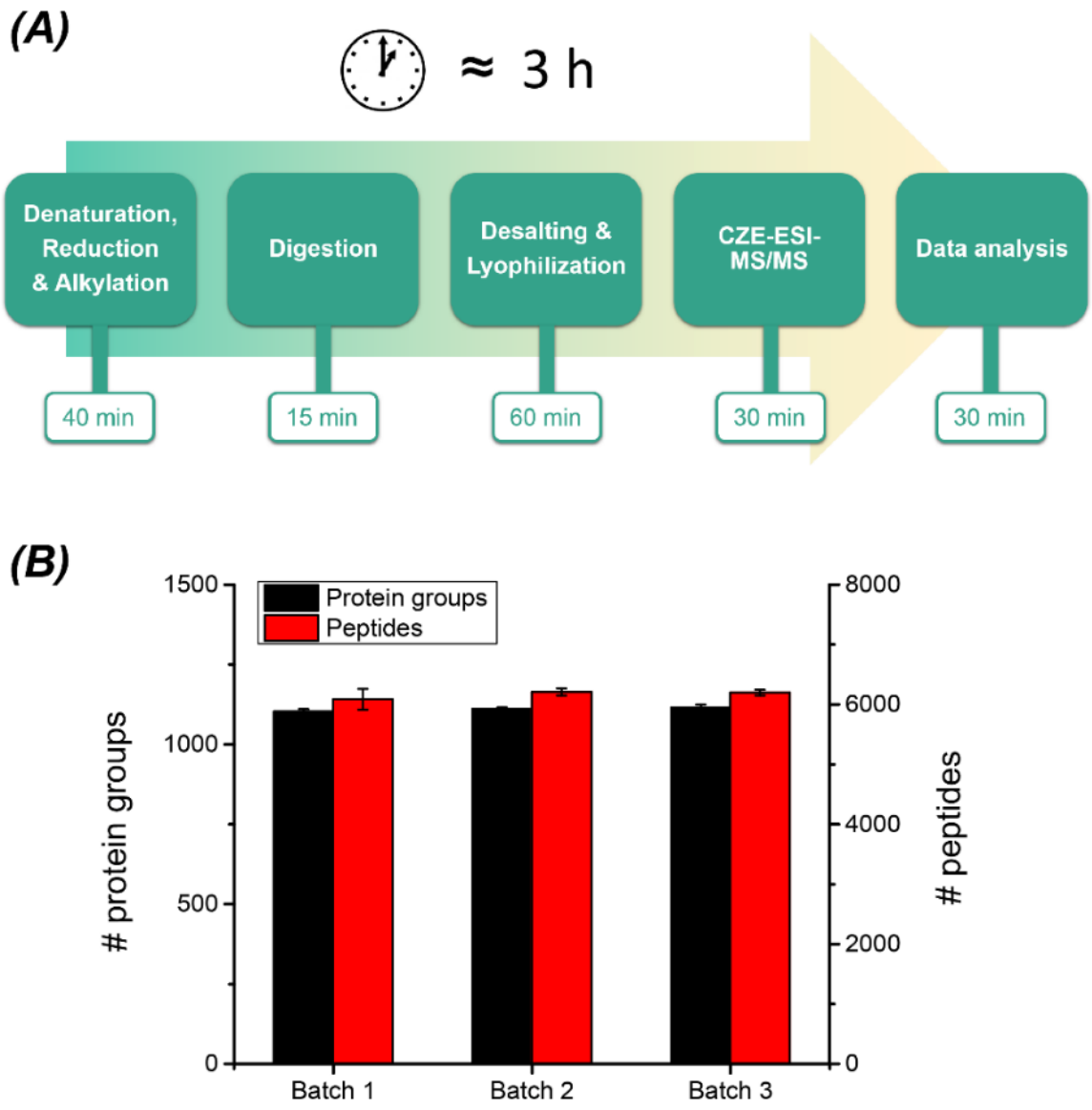

Figure 2.12. The high-throughput bottom-up proteomics workflow using IM-N for rapid protein digestion and CZE-MS/MS for fast sample analysis (A). The number of protein and peptide IDs from the mouse brain proteome using the workflow (B). Three samples were prepared and analyzed by the workflow as three batches. Each sample was analyzed by the CZE-MS/MS in triplicate. The error bars represent the standard deviations of the number of protein and peptide IDs from the triplicate CZE-MS/MS analyses.

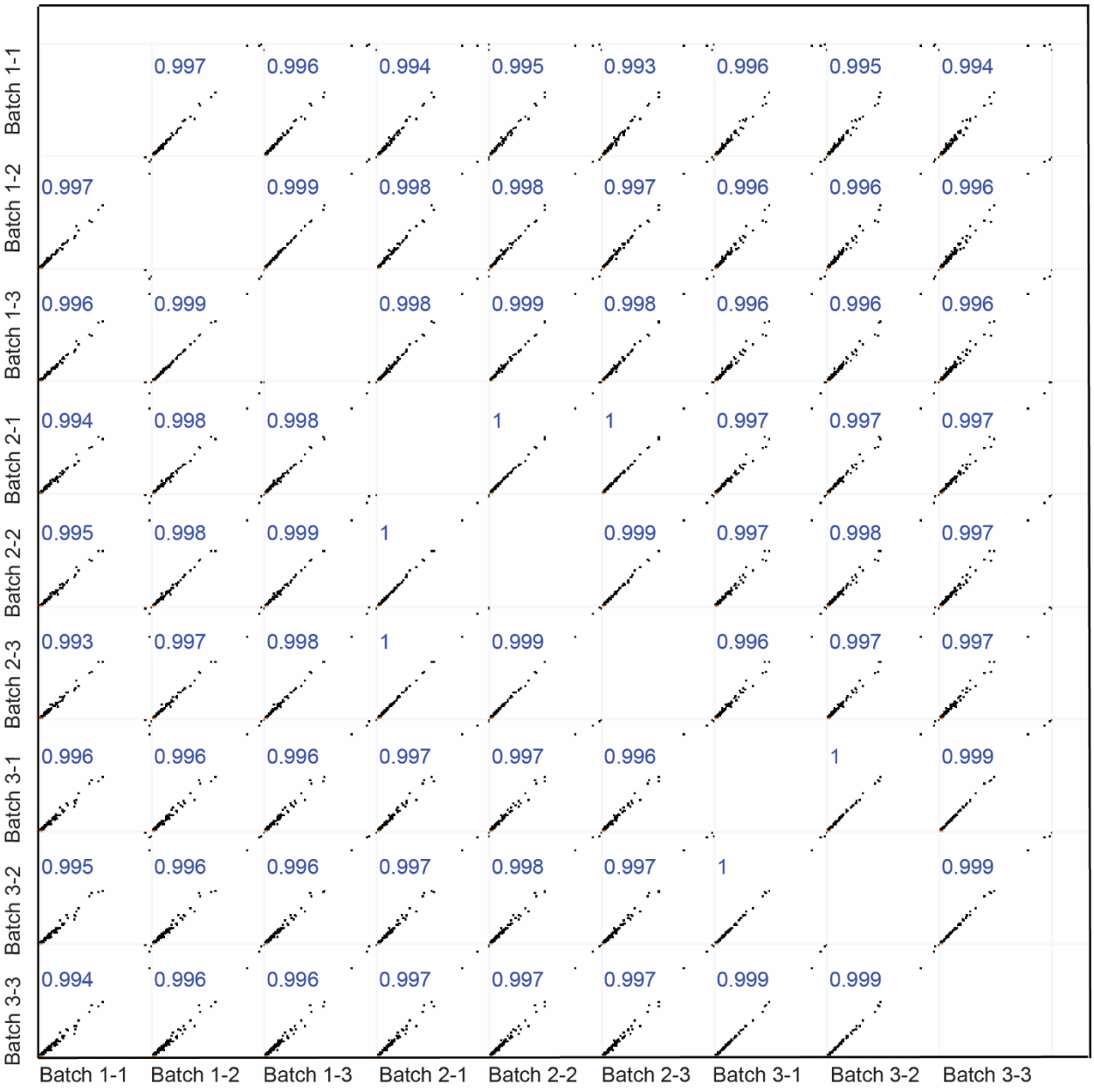

Figure 2.13. Multi-scatter correlations of protein LFQ intensity from the CZE-MS/MS analyses of three batches of the mouse brain proteome digests. Three mouse brain samples were prepared and analyzed by the high-throughput bottom-up proteomics workflow as three batches. Each sample was analyzed by the CZE-MS/MS in triplicate. For example, batch 1-1, 1-2 and 1-3 represent the triplicate CZE-MS/MS analysis of batch 1. Pearson correlation (r) values were labeled. Perseus software (version 1.6.0.7) was used to generate the correlations [10].

2.4 Conclusion

We provided clear answers to two important questions of IM. First, the surface property of the immobilized trypsin microreactors could change cleavage preference compared with FT. Second, IM-N (15-min digestion) can perform as well as FT (12-h digestion) for deep bottom-up proteomics of complex proteomes. Compared to FT, IM-N did not introduce any bias in the digestion of proteins that were involved in various biological processes, were located in different components of cells, had diverse functions, and were expressed in varying abundance. We developed a high-throughput bottom-up proteomics workflow that coupled IM-N based rapid protein digestion to fast CZE-MS/MS analysis. The workflow enabled the characterization of complex proteomes in only 3 h. In our next step, we would further increase the throughput of the workflow. For example, an alternative reducing reagent, tris(2-carboxyethyl)phosphine (TCEP), can be used with alkylating reagents simultaneously, which could simplify the sample pretreatment process and shorten the sample pretreatment time [58]. We will apply the fast workflow for clinical studies such as blood and urine tests.

2.5 Acknowledgment

We thank Prof. Heedeok Hong's group in the Department of Chemistry, Michigan State University for kindly providing the Escherichia coli cells for our experiments. We thank Prof. Chen Chen's group in the Department of Animal Science, Michigan State University for kindly providing the mouse brain for our research. We thank the support from the National Institute of General Medical Sciences, National Institutes of Health (NIH), through Grant R01GM125991 and the support from Michigan State University.

REFERENCES

REFERENCES

[1] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, J. J. Coon, Mol. Cell. Proteomics 13 (2014) 339-347.

[2] J. Ma, L. Zhang, Z. Liang, Y. Shan, Y. Zhang, Trac-Trend Anal. Chem. 30 (2011) 691-702.

[3] L. Switzar, M. Giera, W. M. Niessen, J. Proteome Res. 12 (2013) 1067-1077.

[4] L. Sun, Y. Li, P. Yang, G. Zhu, N. J. Dovichi, J. Chromatogr. A 1220 (2012) 68-74.

[5] L. Sun, G. Zhu, X. Yan, S. Mou, N. J. Dovichi, J. Chromatogr. A 1337 (2014) 40-47.

[6] C. Fan, Z. Shi, Y. Pan, Z. Song, W. Zhang, X. Zhao, F. Tian, B. Peng, W. Qin, Y. Cai, X. Qian, Anal. Chem. 86 (2014) 1452-1458.

[7] W. Qin, Z. Song, C. Fan, W. Zhang, Y. Cai, Y. Zhang, X. Qian, Anal. Chem. 84 (2012) 3138-3144.

[8] G. Vale, H. M. Santos, R. J. Carreira, L. Fonseca, M. Miró, V. Cerdà, M. Reboiro-Jato, J. L. Capelo, Proteomics 11 (2011) 3866-3876.

[9] S. Lin, G. Yao, D. Qi, Y. Li, C. Deng, P. Yang, X. Zhang, Anal. Chem. 80 (2008) 3655-3665.

[10] Y. Deng, C. Deng, D. Qi, C. Liu, J. Liu, X. Zhang, D. Zhao, Adv. Mater. 21 (2009) 1377-1382.

[11] A. K. Palm, M. V. Novotny, Rapid Commun. Mass Spectrom. 18 (2004) 1374-1382.

[12] J. Ma, Z. Liang, X. Qiao, Q. Deng, D. Tao, L. Zhang, Y. Zhang, Anal. Chem. 80 (2008) 2949-2956.

[13] J. Duan, L. Sun, Z. Liang, J. Zhang, H. Wang, L. Zhang, W. Zhang, Y. Zhang, J. Chromatogr. A 1106 (2006) 165-174.

[14] M. T. Dulay, Q. J. Baca, R. N. Zare, Anal. Chem. 77 (2005) 4604-4610.

[15] M. T. Dulay, L. S. Eberlin, R. N. Zare, Anal. Chem. 87 (2015) 12324-12330.

[16] S. Jiang, Z. Zhang, L. Li, J. Chromatogr. A 1412 (2015) 75-81.

[17] W. K. Chui, I. W. Wainer, Anal. Biochem. 201 (1992) 237-245.

[18] F. Xu, W. Wang, Y. Tan, M. L. Bruening, Anal. Chem. 82 (2010) 10045-10051.

[19] J. Dong, W. Ning, W. Liu, M. L. Bruening, Analyst 142 (2017) 2578-2586.

[20] L. Liu, B. Zhang, Q. Zhang, Y. Shi, L. Guo, L. Yang, J. Chromatogr. A 1352 (2014) 80-86.

[21] L. Sun, G. Zhu, N. J. Dovichi, Anal. Chem. 85 (2013) 4187-4194.

[22] Y. Li, R. Wojcik, N. J. Dovichi, J. Chromatogr. A 2018 (2011) 2007-2011.

[23] T. Wang, J. Ma, G. Zhu, Y. Shan, Z. Liang, L. Zhang, Y. Zhang, J. Sep. Sci. 33 (2010) 3194-3200.

[24] S. Moore, S. Hess, J. Jorgenson, J. Chromatogr. A 1476 (2016) 1-8.

[25] G. W. Slysz, D. C. Schriemer, Anal. Chem. 77 (2005) 1572-1579.

[26] S. Feng, M. Ye, X. Jiang, W. Jin, H. Zou, J. Proteome Res. 5 (2006) 422-428.

[27] H. Yuan, Y. Zhou, S. Xia, L. Zhang, X. Zhang, Q. Wu, Z. Liang, Y. Zhang, Anal. Chem. 84 (2012) 5124-5132.

[28] H. Yuan, S. Zhang, B. Zhao, Y. Weng, X. Zhu, S. Li, L. Zhang, Y. Zhang, Anal. Chem. 89 (2017) 6324-6329.

[29] F. Wang, X. Wei, H. Zhou, J. Liu, D. Figeys, H. Zou, Proteomics 12 (2012) 3129-3137.

[30] J. Spross, A. Sinz, Anal. Chem. 82 (2010) 1434-1443.

[31] H. K. Hustoft, O. K. Brandtzaeg, M. Rogeberg, D. Misaghian, S. B. Torsetnes, T. Greibrokk, L. Reubsaet, S. R. Wilson, E. Lundanes, Sci. Rep. 3 (2013) 3511.

[32] M. Ye, Y. Pan, K. Cheng, H. Zou, Nat. Methods 11 (2014) 220-222.

[33] Y. Pan, K. Cheng, J. Mao, F. Liu, J. Liu, M. Ye, H. Zou, Anal. Bioanal. Chem. 406 (2014) 6247-6256.

[34] J. L. Hsu, S. Y. Huang, N. H. Chow, S. H. Chen, Anal. Chem. 75 (2003) 6843-6852.

[35] F. Wang, R. Chen, J. Zhu, D. Sun, C. Song, Y. Wu, M. Ye, L. Wang, H. Zou, Anal. Chem. 82 (2010) 3007-3015.

[36] T. Šlechtová, M. Gilar, K. Kalíková, E. Tesařová, Anal. Chem. 87 (2015) 7636-7643.

[37] T. Geiger, A. Wehner, C. Schaab, J. Cox, M. Mann, Mol. Cell. Proteomics 11 (2012) M111.014050.

[38] Q. Zhao, F. Fang, Y. Shan, Z. Sui, B. Zhao, Z. Liang, L. Zhang, Y. Zhang, Anal. Chem. 89 (2017) 5179-5185.

[39] C. Ding, J. Jiang, J. Wei, W. Liu, W. Zhang, M. Liu, T. Fu, T. Lu, L. Song, W. Ying, C. Chang, Y. Zhang, J. Ma, L. Wei, A. Malovannaya, L. Jia, B. Zhen, Y. Wang, F. He, X. Qian, J. Qin, Mol. Cell. Proteomics. 12 (2013) 2370-2380.

[40] P. Mertins, J. W. Qiao, J. Patel, N. D. Udeshi, K. R. Clauser, D. R. Mani, M. W. Burgess, M. A. Gillette, J. D. Jaffe, S. A. Carr, Nat. Methods 10 (2013) 634-637.

[41] D. Chen, X. Shen, L. Sun, Analyst 42 (2017) 2118-2127.

[42] G. Zhu, L. Sun, N. J. Dovichi, Talanta 146 (2016) 839-843.

[43] L. Sun, G. Zhu, Y. Zhao, X. Yan, S. Mou, N. J. Dovichi, Angew. Chem. Int. Ed. 52 (2013) 13661-13664.

[44] R. Wojcik, O. O. Dada, M. Sadilek, N. J. Dovichi, Rapid Commun. Mass Spectrom. 24 (2010) 2554-2560.

[45] L. Sun, G. Zhu, Z. Zhang, S. Mou, N. J. Dovichi, J. Proteome Res. 14 (2015) 2312-2321.

[46] J. Cox, M. Mann, Nat. Biotechnol. 26 (2008) 1367-1372.

[47] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, M. Mann, J. Proteome Res. 10 (2011) 1794-1805.

[48] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, M. Mann, Mol. Cell. Proteomics 13 (2014) 2513-2526.

[49] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, J. Cox, Nat. Methods 13 (2016) 731-740.

[50] D. Huang, B. T. Sherman, R. Lempicki, Nat. Protoc. 4 (2008) 44-57.

[51] D. Huang, B. T. Sherman, R. Lempicki, Nucleic Acids Res. 37 (2009) 1-13.

[52] D. Zhang, L. Yuwen, L. Peng, J. Chem. 2013 (2013) 1.

[53] K. Sharma, S. Schmitt, C. Bergner, S. Tyanova, N. Kannaiyan, N. Manrique-Hoyos, K. Kongi, L. Cantuti, U. Hanisch, M. Philips, M. Rossner, M. Mann, M. Simons, Nat. Neurosci. 18 (2015) 1819-1831.

[54] E. I. Chen, D. McClatchy, S. K. Park, J. R. Yates III, Anal. Chem. 80 (2008) 8694-8701.

[55] J. Wiśniewski, A. Zougman, M. Mann, J. Proteome Res. 8 (2009) 5674-5678.

[56] Z. Ning, D. Seebun, B. Hawley, C. K. Chiang, D. Figeys, J. Proteome Res. 12 (2013) 1512-1519.

[57] P. E. Geyer, N. A. Kulak, G. Pichler, L. M. Holdt, D. Teupser, M. Mann, Cell Syst. 2 (2016) 185-195.

[58] J. K. Goodman, C. G. Zampronio, A. M. E. Jones, J. R. Hernandez-Fernaud, Proteomics 18 (2018) 1800236.

CHAPTER 3. Coupling Capillary Zone Electrophoresis to Activated Ion-Electron Capture Dissociation (AI-ECD) for Top-Down Characterization of Protein Mixtures

3.1 Introduction

Delineation of proteoforms in cells plays a central role in accurate understanding of protein function in biological processes because different proteoforms from the same gene can have divergent functions [1-6]. Mass spectrometry (MS)-based denaturing top-down proteomics (dTDP) aims to comprehensively characterize proteoforms in cells, which needs high-capacity liquid-phase separation and extensive gas-phase fragmentation of proteoforms [7,8].

Liquid chromatography-MS (LC-MS), typically reversed-phase LC (RPLC), is routinely used for dTDP [9-17]. Proteomes are super complex regarding the number of proteoforms. For example, over one million proteoforms have been predicted in the human proteome [18]. The high sample complexity leads to a high need for liquid-phase separation methods with much better separation capacity for proteoforms. Capillary zone electrophoresis (CZE)-MS has been investigated by our group and others for high-capacity separation of proteoforms, enabling large-scale delineation of proteoforms in complex biological systems [19-27]. CZE-MS has been proven as an alternative tool to RPLC-MS for dTDP due to its several valuable features, such as better sensitivity than RPLC-MS [28,29], high separation efficiency for proteoforms [21], and great potential for accurate prediction of proteoforms' electrophoretic mobility [30-32]. The tremendous progress of developing robust and highly sensitive CE-MS interfaces has laid the solid foundation for deploying CZE-MS for dTDP [33-37].

Extensive gas-phase fragmentation of proteoforms requires new fragmentation methods. Collision-based methods, *i.e.,* collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD), are the routine approaches for fragmentation of biomolecules [11-13,19-22]. However, CID and HCD have some bias in backbone cleavages, impeding complete cleavages of proteoforms' backbones. Alternative gas-phase fragmentation techniques have been developed in recent years to provide better characterization of large biomolecules, including but not limited to electron-transfer dissociation (ETD) [10,38-40], electron-capture dissociation (ECD) [41-45], and ultraviolet photodissociation (UVPD) [46-49].

ECD for protein fragmentation was pioneered by the McLafferty group in the late 1990s [50,51]. ECD-based protein fragmentation is a nonergodic process, in which electrons are captured at the protonated sites of positively charged protein ions, energetic hydrogen atoms ($H^{\cdot}$) are ejected from the protein ions and are captured at high-affinity sites of the protein ions such as backbone amide, leading to backbone cleavages with the production of c and $z^{\cdot}$ ions [51]. ECD fragmentation can be improved by activating the ECD fragment ions, *e.g.*, collision with gas molecules, to break their intramolecular noncovalent bonds and this modified ECD was called activated-ion ECD (AI-ECD) [52]. For example, Horn *et al.* has achieved the cleavage of 116 backbone bonds in a 29-kDa protein using AI-ECD on a Fourier transform (FT) ion cyclotron resonance (ICR) mass spectrometer in 2000 [52]. In another example, Ge *et al.* has obtained the efficient characterization of large intact proteins (45 kDa) using the AI-ECD method in 2002 on an FT-ICR mass spectrometer [53]. The FT-ICR mass spectrometer equipped with ECD has also been employed for the characterization of integral membrane proteins and large protein complexes [41,54]. More recently, the ECD cell has been integrated into QqQ [42], Q-TOF [43], ion mobility [55], and Orbitrap [44, 56, 57] mass spectrometers for peptide, protein, and protein

complex fragmentation. Fort *et al.* demonstrated that ECD outperformed HCD for fragmentation of ubiquitin and myoglobin on an orbitrap mass spectrometer as measured by backbone cleavage coverage [56]. Shaw *et al.* reported a 93% backbone cleavage coverage for carbonic anhydrase II (29 kDa) using ECD on an Orbitrap mass spectrometer with a direct-infusion approach [44], demonstrating the great potential of ECD to advance dTDP via offering extensive protein fragmentation. Direct-infusion MS is typically deployed for ECD-based TDP and in-front liquid-phase separation is needed to analyze complex protein mixtures.

In this work, for the first time, we coupled CZE to ECD on a Q-TOF mass spectrometer for highly efficient liquid-phase separation and extensive gas-phase fragmentation of intact proteins. CID was integrated with ECD to activate the ECD ions and to produce more extensive protein fragmentation. We employed the online CZE-(AI-ECD)-Q-TOF platform for characterization of a standard protein mixture in a mass range of 8-30 kDa. We investigated the effect of CID potential on the backbone cleavage coverage of proteins from AI-ECD, studied how protein precursor's charge state influenced backbone cleavage coverage from AI-ECD, and showed that combining AI-ECD fragment ions from different charge states could boost the backbone cleavage coverage of proteins drastically compared to that from a single charge state. Finally, we achieved baseline separation and nearly complete backbone cleavages for the standard protein mixture using the online CZE-AI-ECD on an Agilent 6545XT AdvanceBio Q-TOF mass spectrometer, suggesting the great potential of the new platform for advancing dTDP.

3.2 Experimental

3.2.1 Materials and reagents

All standard proteins, ammonium acetate (NH_4Ac), dithiothreitol (DTT), iodoacetamide (IAA), and Microcon-30kDa centrifugal filter units for buffer exchange were purchased from Sigma-Aldrich (St. Louis, MO). LC/MS grade water, methanol, formic acid (FA) and acetic acid (AA) were purchased from Fisher Scientific (Pittsburgh, PA). Urea was purchased from Alfa Aesar (Tewksbury, MA). Hydrofluoric acid (HF) and acrylamide were purchased from Acros Organics (NJ, USA). The fused silica capillary (50 μm i.d., 360 μm o.d.) was purchased from Polymicro Technologies (Phoenix, AZ).

3.2.2 Sample preparation

A mixture of standard proteins consisting of ubiquitin (bovine, 0.05 mg/mL), myoglobin (equine, 0.1 mg/mL), carbonic anhydrase (CA, bovine, 0.5 mg/mL), and bovine serum albumin (BSA, 2.0 mg/mL) was prepared in 50 mM NH_4HCO_3 (pH 8.0) for the CZE-MS experiment. Carbonic anhydrase and its impurity superoxide dismutase (SOD, bovine) [21] were denatured with 8 M urea at 37 °C, reduced with DTT and alkylated with IAA, followed by buffer exchange with a Microcon-30 kDa centrifugal filter unit. For the buffer exchange, 200 μg protein material was loaded on the membrane and centrifuged at 14,000 *g* to remove the sample buffer. Then the sample was washed with 200 μL 50 mM NH_4HCO_3 (pH 8.0) for three times, followed by protein recovery from the membrane using 30 μL 50 mM NH_4HCO_3 (pH 8.0) with pipetting and vortexing gently.

E. coli (strain K-12, substrain MG1655) was cultured in Lysogeny broth (LB) medium at 37 °C with 225 rpm shaking until the OD600 value reached 0.7. The bacteria were collected by centrifugation (4,000 rpm, 10 min), then washed three times with phosphate-buffered saline

(PBS). Afterward, the *E. coli* pellet was suspended in the lysis buffer containing 8 M urea, protease inhibitor (Roche), phosphatase inhibitor (Roche), and 100 mM ammonium bicarbonate (pH 8.0). The cells were lysed for 1 min using a homogenizer 150 (Fisher Scientific) and then sonicated on ice for 5 min twice with a Branson Sonifier 250 (VWR Scientific). The *E. coli* lysate was centrifuged at 14,000 *g* for 10 min to collect the supernatant containing extracted proteins. The concentration of total proteins was measured by a bicinchoninic acid (BCA) kit (Fisher Scientific) according to manufacturer's instructions. After denaturation, reduction and alkylation, the buffer exchange of protein sample was conducted by centrifugation with Microcon-30 kDa centrifugal filter (Merck Millipore) at 14,000 *g* for 10 min and then washing three times with 50 mM ammonium bicarbonate (pH 8.0). Finally, the proteins retained on the centrifugal filter membrane were re-dissolved in 50 mM ammonium bicarbonate (pH 8.0). The final concentration was 1 mg/ml.

3.2.3 CZE-ESI-MS/MS analysis

A 7100 CE System from Agilent Technologies (Santa Clara, CA) was used for automated operation of CZE. An EMASS-II CE-MS Ion Source commercialized by CMP Scientific (Brooklyn, NY) was used to couple CZE to a 6545XT AdvanceBio Q-TOF (Agilent Technologies) mass spectrometer, **Figure 3.1A** [34,35]. The ECD fragmentation was realized by a built-in electromagnetostatic ExD cell (e-MSion, Corvallis, OR) between the quadrupole and the collision induced dissociation (CID) cell, **Figure 3.1B**. The ESI emitters of the CE-MS interface were pulled from borosilicate glass capillaries (1.0 mm o.d., 0.75 mm i.d., 10 cm length) with a Sutter P-1000 flaming/brown micropipet puller. The opening size of the ESI emitters was 20-30 μm. Voltage for ESI ranged from +2.0 to +2.3 kV.

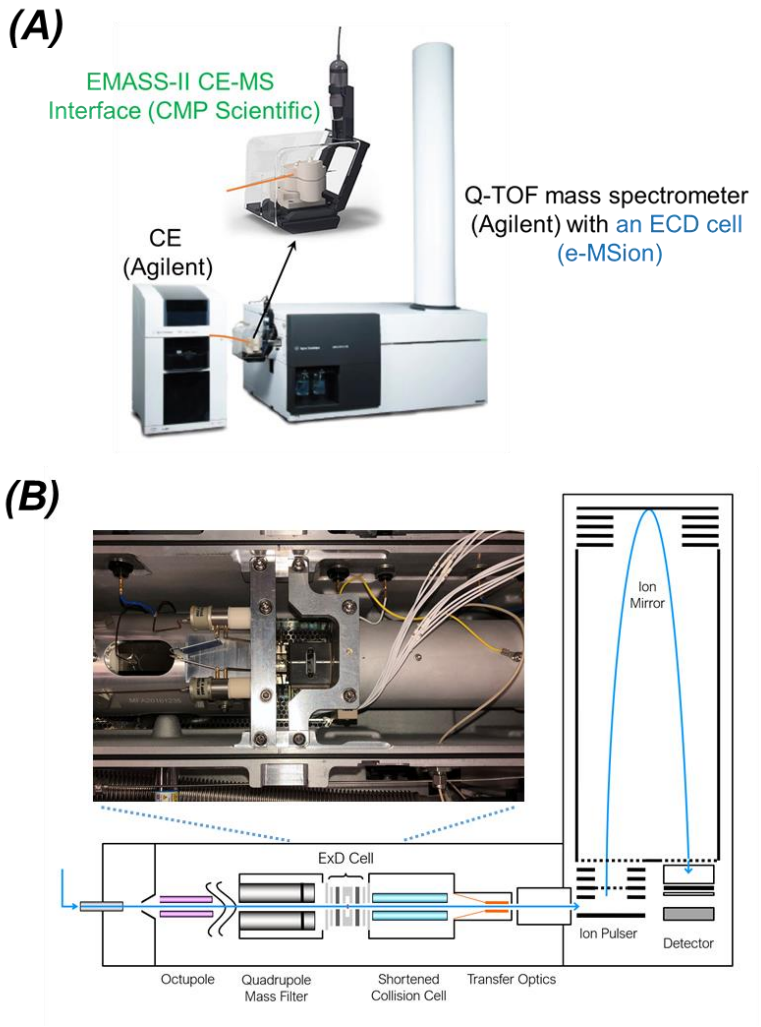

Figure 3.1. (A) Image of the CZE-MS system including a 7100 Agilent CE system, an EMASS-II CE-MS interface from the CMP Scientific, and an Agilent 6545XT Q-TOF mass spectrometer with an ECD cell. The image was adapted from https://www.agilent.com/cs/library/applications/application-nistmab-charge-variants-cief-ms-5994-1079en-agilent.pdf. (B) Schematic of Agilent 6545XT AdvanceBio Q-TOF mass spectrometer with built-in ExD cell (e-MSion). The inset shows an image of the ExD cell installed between quadrupole and shortened collision cell. The figure was kindly provided by the e-MSion.

A 75-cm long capillary (50 μm i.d., 360 μm o.d.) coated with linear polyacrylamide (LPA) was used for separation of the standard protein mixture. A 1.5-m long LPA-coated capillary (50 μm i.d., 360 μm o.d.) was used for *E. coli* separation. The LPA coating was prepared on the inner wall of the capillary based on the literature [58-59]. One end of the capillary was etched with HF to reduce the outer diameter of the capillary to ~ 70 μm [60]. The background electrolyte (BGE) for CZE was 5% (*v*/*v*) AA (pH ~ 2.4). The sheath buffer was 0.2% (*v*/*v*) FA containing 10% (*v*/*v*) methanol. High voltage (+30 kV) was applied for CZE separation. For each CZE-MS/MS run of the standard protein mixture, 120 nL of the sample was injected into the capillary. For CZE-MS/MS run of *E. coli* proteome, 500 nL of the sample was injected into the capillary. The injection was realized by applying 100 mbar air pressure and the injection volume was calculated based on Poiseuille's law.

The 6545XT AdvanceBio Q-TOF (Agilent) was used for the experiments. The gas temperature and flow rate of nitrogen drying gas was 325 ^{o}C and 1 L/min. The voltage applied on the ion transfer capillary was 0 V. The fragmentor was 175 V and the skimmer was 65 V. The mass range was set as Standard (3200 *m/z*). The slicer mode was High Resolution. The instrument mode was Extended Dynamic Range (2 GHz). For MS, the mass range was 600-3000 *m/z*, and the scan rate was 1 spectrum/sec. For MS/MS, the mass range was 300-3000 *m/z*, and the scan rate was 1 spectrum/sec. The precursor ion isolation width for MS/MS was set as wide (~9 amu). For auto MS/MS, max precursors per cycle was 3. Active exclusion for precursor selection was not enabled and the precursors were sorted by abundance only. Only precursors with more than 3000 counts in abundance were isolated for MS/MS. For targeted MS/MS, the max time between MS1 spectra was 5 s. Five MS/MS spectra were collected for each targeted precursor. AI-ECD was used for fragmentation.

3.2.4 Electromagnetostatic ExD Cell

The e-MSion ExD cell mounted on a shortened collision cell replaced Agilent's standard CID cell (**Figure 3.1**). The ExD cell consists of a hot rhenium filament producing electrons and two high-temperature magnets that restrain electrons radially to the central axis. The analyte ions are guided through the cell without trapping by seven DC electrostatic lens. An auxiliary electronics control module controlling the ExD cell was interfaced to the instrument computer. The ExD cell was tuned with direct infusion of Substance P, ubiquitin, and CA using our CZE system. Briefly, the CZE capillary was first filled with the sample. After that, the sample was pushed out of the capillary slowly via applying a small pressure (50-100 mbar) at the sample injection end of the capillary for ESI-MS. The ExD cell was first tuned to achieve full ion transmission without ECD in MS1. Then, the ExD cell and filament current were optimized to achieve the maximum ECD fragment ion intensity in MS2. The optimized ECD conditions including electrostatic potentials and filament current settings are shown in **Table 3.1**.

Table 3.1. Optimized ExD cell settings for the ECD (ECD on) and positive transmission without ECD (ECD off).

Settings	ECD on	ECD off
Lens 1 (V)	28.0	20.0
Lens 2 (V)	-23.5	1.2
Lens 3 (V)	33.0	25.8
Lens 4 (V)	41.0	27.2
Lens 5 (V)	31.5	29.3
Lens 6 (V)	26.0	24.7
Filament Bias (V)	23.0	21.8
Filament current (A)	2.6	2.6

3.2.5 Data analysis

For annotation of the MS/MS spectra of standard proteins, MS/MS spectra for standard proteins were first averaged manually over the electrophoretic peak of each protein in Agilent MassHunter Qualitative Navigator B.08.00. The information in the averaged MS/MS spectra including *m/z* and intensity of ions were exported and saved as a .mgf file for each protein. After that, each .mgf file was loaded into the LcMsSpectator (https://omics.pnl.gov/software/lcmsspectator) for fragment ion match and annotation. Matched fragment ion types were *b*, *y*, *c*, z (*z* and *z*·), and *w* with a 20-ppm mass tolerance and minimum S/N threshold as 1.5. The annotated MS/MS spectra were also manually checked. The sequences of standard proteins were obtained from the UniProt (https://www.uniprot.org/). Fragmentation patterns and backbone cleavage coverages were generated by the LcMsSpectator.

Deconvolution of large proteoforms from *E. coli* proteome was performed with Agilent MassHunter BioConfirm 10.0 using Maximum Entropy algorithm. The mass step was 0.05 Da. Other parameters for deconvolution were set as default.

3.3 Results and discussion

3.3.1 Effect of CID potential on the performance of AI-ECD for protein backbone cleavage

We deployed CID to activate the ECD fragment ions to destroy the intramolecular noncovalent bonds, leading to more extensive backbone cleavage coverage [52]. We set the ECD filament current as 2.6 A for efficient ECD fragmentation. The CID was mainly used to activate the ECD fragment ions. Five types of fragment ions (b, y, c, z and w ions) were considered in the data

analysis using the LcMsSpectator. The z ion in this work represents the z and z· ions. The regular low electron-energy (<1 eV) ECD mainly produces c and z ions and the high electron-energy (3-13 eV) ECD can induce protein backbone fragmentation via different pathways, yielding c, z, b, y, a, and w ions [44, 61]. The w ions are from the secondary fragmentation of z· ions through side chain neutral loss due to the high electron energy, and they are very useful for distinguishing isomeric amino acid residues, like leucine (L) and isoleucine (I), in protein sequences [61]. I and L offer distinctive side chain neutral loss, $\cdot C_2H_5$ (29 Da) and $\cdot C_3H_7$ (43 Da), respectively.

We speculated that too high CID potential would produce obvious CID fragmentation of the protein precursor ions and ECD fragment ions, leading to much more complicated MS/MS spectra and challenges for data interpretation. Therefore, we investigated how CID potential affected AI-ECD fragmentation to achieve an optimized CID potential using the CZE-MS system and a standard protein mixture containing ubiquitin (8.5 kDa), myoglobin (17 kDa), CA (29 kDa), BSA (66 kDa), and one protein impurity SOD (16 kDa). We chose three CID potentials, 10 V, 30 V, and 50 V. We did not test 0-V CID potential because a little bit of CID potential is needed to facilitate ion transmission and achieve sufficient fragment ion signal in our system.

First, CZE achieved baseline and reproducible separation of the standard proteins with relative standard deviations (RSDs) of migration time less than 2%, **Figure 3.2A**. Second, the CZE-AI-ECD method with the 10-V CID potential yielded extensive backbone cleavages of ubiquitin (97%), myoglobin (>80%), CA (~60%), and SOD (~70%) under auto MS/MS mode. For the BSA, we only gained a limited backbone cleavage coverage (<20%) using the AI-ECD due to its large mass (66 kDa) and folded structure with many internal disulfide bonds. We analyzed both completely unfolded (denatured) and folded (nondenatured) CA using AI-ECD, obtaining much better cleavage coverage for the unfolded CA compared to the folded one (~60% *vs.* ~40%). We

noted that the c, z, and w ions dominated the generated fragment ions of ubiquitin (82%), myoglobin (69%), and BSA (66%) when the CID potential was 10 V. For SOD and CA, about 50% of the fragment ions were b and y ions. Third, when the CID potential was increased from 10 V to 50 V, the backbone cleavage coverage of proteins except the BSA was reduced substantially, **Figure 3.2B**. We speculated the reason for the decrease of cleavage coverage was most likely because high CID potential resulted in the over-fragmentation of proteins, in which the generated fragment ions were too small for identification. The number of sequence-informative fragment ions of proteins from AI-ECD decreased as a function of CID potential from 10 to 50 V, **Figure 3.2C**. We further analyzed the mass and intensity of generated fragment ions from three proteins (CA, SOD, and myoglobin), which had the most significant reduction of backbone cleavage coverage as a function of CID potential. The increase of CID potential resulted in fragment ions with obvious lower mass (**Figure 3.2D**) and drastically lower intensity (**Figure 3.2E**). Take CA as an example, the correlation between mass and intensity of fragment ions for CID 10 V and 50 V further elucidated the overall reduction of intensity of fragment ions across the whole mass range, **Figure 3.2F**.

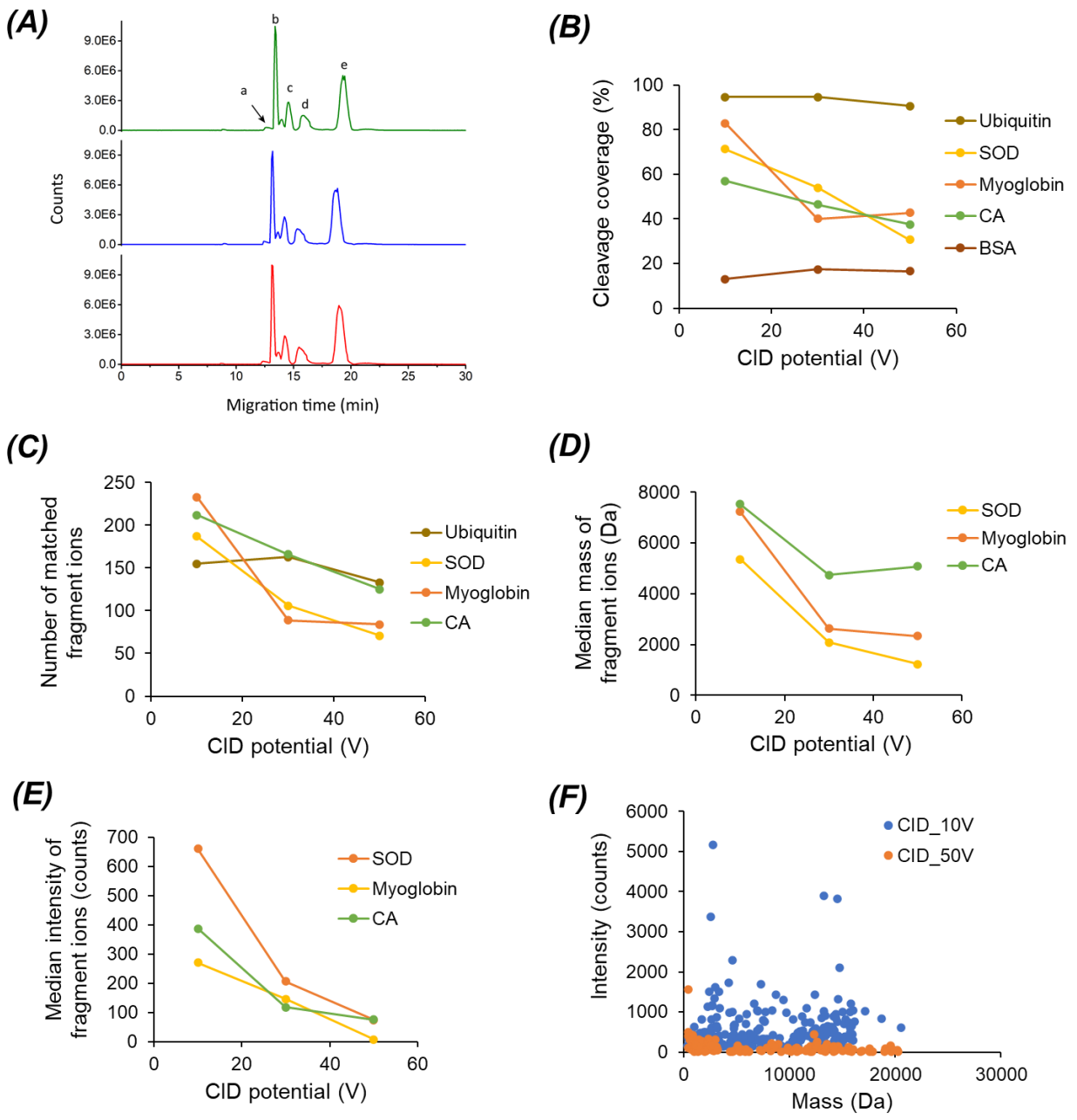

Figure 3.2. (A) Base peak electropherograms of the mixture of a) BSA, b) ubiquitin, c) myoglobin, d) CA and e) SOD after triplicate CZE-MS/MS analyses. (B-E) Changes of backbone cleavage coverage, number of matched fragment ions, median mass of fragment ions and median intensity of fragment ions for standard proteins across different CID potential. (F) Correlation between mass and intensity of fragment ions from CA with 10-V and 50-V CID potentials.

Further comparison of annotated AI-ECD spectra of CA under two different conditions (CID 10 V and 50 V) showed that the 50-V CID produced more noisy MS/MS spectra compared to the 10-V CID potential in the low *m/z* region, **Figure 3.3**. All the evidence suggests that AI-ECD with high CID potential (30 V and 50 V) leads to over fragmentation of proteins lower than 30 kDa. We considered AI-ECD with the 10-V CID as the optimized AI-ECD condition and used it in all the following studies.

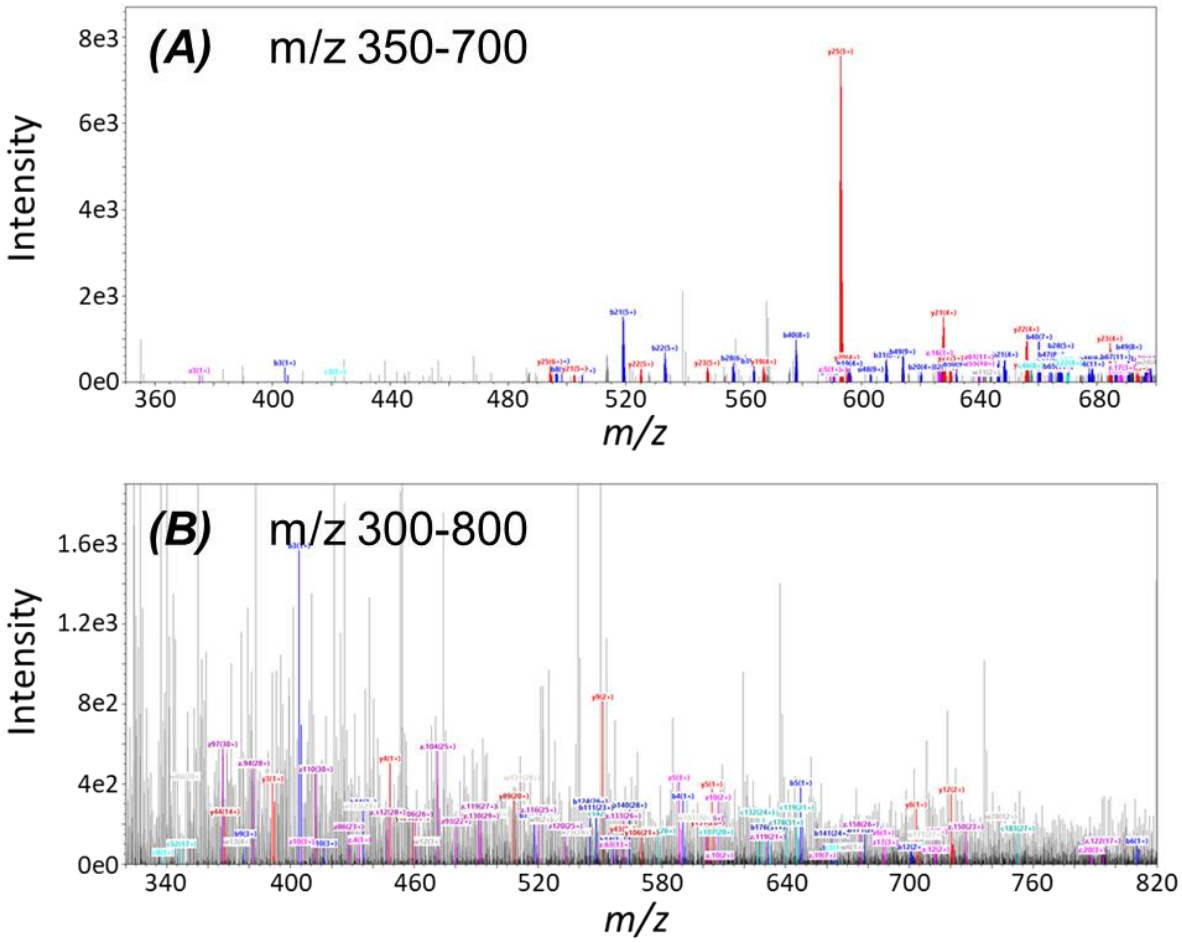

Figure 3.3. Annotated MS/MS spectra of CA from AI-ECD (A) using a 10-V CID potential and (B) using a 50-V CID potential. Auto MS/MS was used. The spectra were averaged from all MS/MS spectra with different precursor ions. Blue: b ions; red: y ions; cyan: c ions; pink: z ions; grey: w ions.

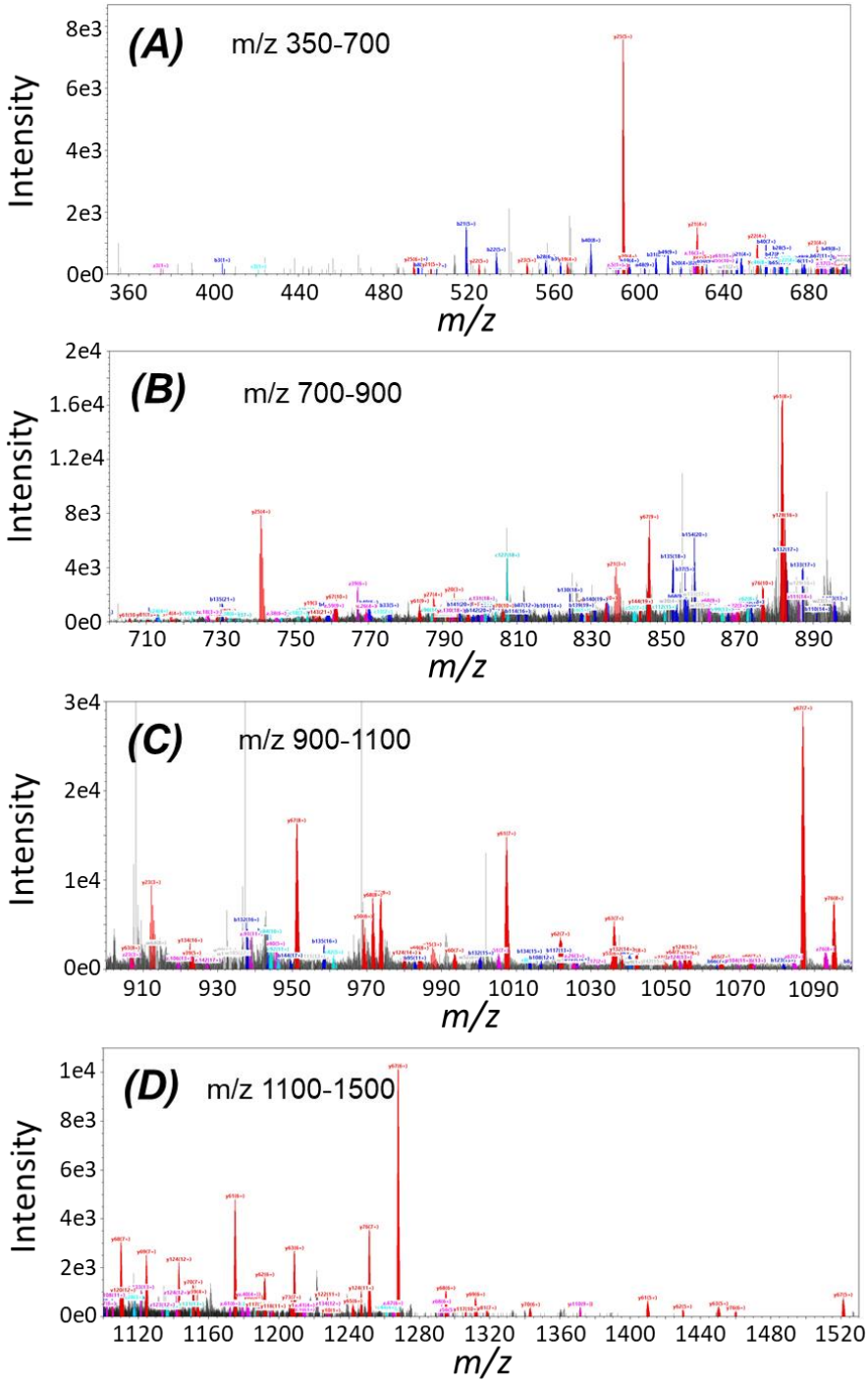

Figure 3.4. Annotated MS/MS spectra of CA obtained using CZE-MS/MS with AI-ECD fragmentation. The CID potential was 10 V. Auto MS/MS was used. The spectra were averaged from all MS/MS spectra with different precursor ions. Blue: b ions; red: y ions; cyan: c ions; pink: z ions; grey: w ions.

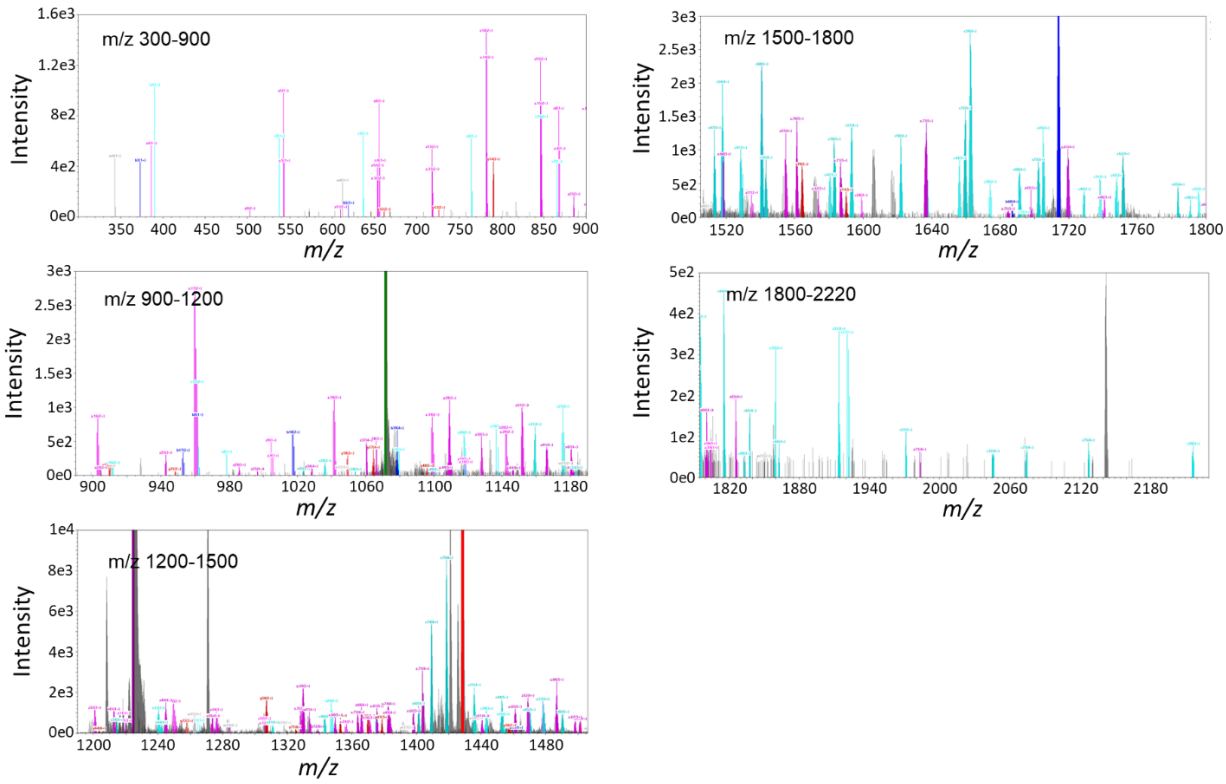

Figure 3.5. Annotated MS/MS spectra of ubiquitin obtained using CZE-MS/MS with AI-ECD fragmentation. The CID potential was 10 V. Auto MS/MS was used. The spectra were averaged from all MS/MS spectra with different precursor ions. Blue: b ions; red: y ions; cyan: c ions; pink: z ions; grey: w ions.

We further manually evaluated the annotated MS/MS spectra of the proteins under the optimized AI-ECD condition, **Figure 3.4** and **Figure 3.5**. For the large protein CA (29 kDa) and the small protein ubiquitin (8.5 kDa), most of the high abundant ions in these MS/MS spectra were annotated as common AI-ECD fragment ions (b, y, c, z, and w ions).

3.3.2 Effect of protein precursor's charge state on AI-ECD fragmentation of proteins

During ECD fragmentation, positively charged protein ions capture electrons emitted from the internal heated rhenium filament, leading to protein fragmentation. Electron capture becomes

more probable with higher charge states of protein ions, which can capture more electrons and generally results in more efficient protein fragmentation. Therefore, we investigated the protein fragmentation from AI-ECD as a function of the protein charge state for three of the five proteins in the protein mixture (Myoglobin, SOD, and CA). The chosen charge states and *m/z* are summarized in **Table 3.2**. We selected three charge states (low, medium, and high) for each protein to make sure that these charge states were significantly different from each other and had comparable precursor ion abundances. The medium charge state was the most abundant charge state in each spectrum. We used the CZE-AI-ECD to separate the standard protein mixture and fragment the specific charge states of the three proteins in targeted MS/MS mode. About 20-60 MS/MS spectra were acquired for each charge state of each protein. The MS/MS spectra were averaged, followed by fragment identification using the LcMsSpectator software.

Table 3.2. Charge states and *m/z* of myoglobin, CA and SOD for studying the effect of protein charge state on AI-ECD fragmentation.

	Low charge	Medium charge	High charge
Myoglobin	**1060 *m/z*, +16**	893 *m/z*, +19	772 *m/z*, +22
CA	1210 *m/z*, +24	**1076 *m/z*, +27**	937 *m/z*, +31
SOD	**928 *m/z*, +17**	830 *m/z*, +19	751 *m/z*, +21

The protein charge state altered the number of sequence-informative fragment ions (**Figures 3.6A-C**) and the backbone cleavage coverage (**Figures 3.6D-F**) materially. The total number of fragment ions dropped obviously as the charge state changes from low to high for myoglobin and SOD. For CA, the low and medium charge states yielded a comparable number of fragment ions, which are significantly better than that from the high charge state. When we examined the changes in the number of different types of fragment ions as a function of protein charge state, we observed that the numbers of c, z and w ions for the three proteins all declined dramatically at the high

charge states compared to the low charge states. However, the numbers of b and y ions for myoglobin and CA show different trends from the c, z and w ions. The high charge states of CA and myoglobin produced more b and y ions than their low charge states. The corresponding protein backbone cleavage coverage data depicted in **Figures 3.6D-F** agree well with the number of fragment ions data discussed above. The low and high charge states generated the highest and lowest backbone cleavage coverages, respectively, for all the three proteins when we considered all types of fragment ions or only c, z and w ions. When we only considered b and y ions, the high charge states yielded better cleavage coverage than the low charge states for CA and myoglobin.

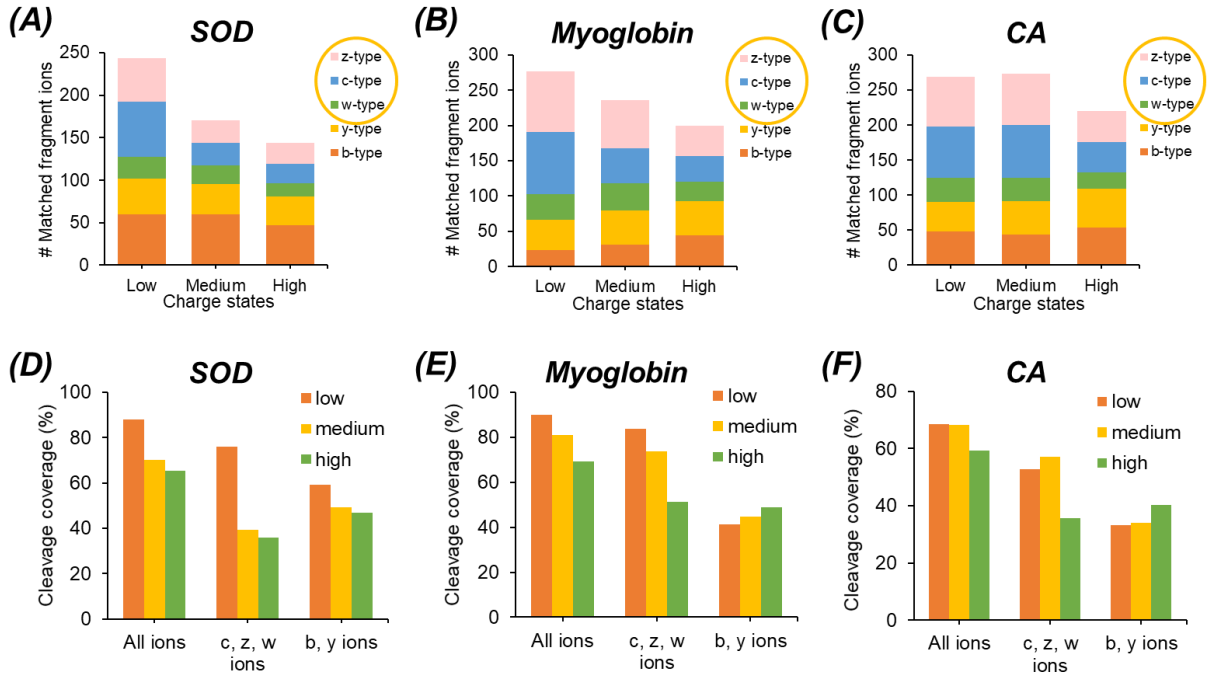

Figure 3.6. Number of matched fragment ions (A-C) and backbone cleavage coverage (D-F) from AI-ECD fragmentation of SOD, myoglobin, and CA as a function of the precursor's charge state. The CID potential was 10 V.

The drastic decrease of overall backbone cleavage coverage and the total number of fragment ions for the three proteins as a function of protein ion charge state might be due to over

fragmentation. The high charge state protein ions produced smaller fragment ions compared to the low charge state protein ions, **Figure 3.7**, offering some evidence of over fragmentation. We then asked why myoglobin and CA showed different trends from SOD regarding the backbone cleavage coverage from only b and y ions when the protein ion charge state increased from low to high. Although myoglobin has a similar mass to SOD, it was not as fully denatured as SOD in our experiment. CA is a much larger protein than SOD. Therefore, higher charge states of myoglobin and CA can facilitate the unfolding of the gas-phase protein ions to a slightly better extent, offering a higher chance for collision-based fragmentation. On the other hand, a protein ion with a high charge state can capture electrons more efficiently compared to the medium and low charge states due to the higher charge density, which might lead to absorption of too much energy for proteins and eventually over fragmentation. SOD had more severe over fragmentation than myoglobin and CA, which is evidenced by the more profound drop of fragment ion mass of SOD than that of CA and myoglobin as a function of charge state, **Figure 3.7**.

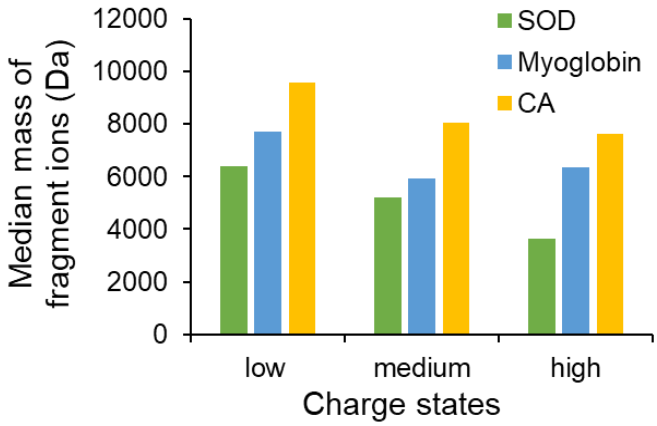

Figure 3.7. Median mass of fragment ions from AI-ECD fragmentation of SOD, myoglobin, and CA as a function of the precursor's charge state. The CID potential was 10 V.

We noted that although the charge state of protein ions could influence their backbone cleavage coverage from AI-ECD significantly, our CZE-AI-ECD-Q-TOF system provided reasonably extensive protein backbone cleavages for the three proteins in a mass range of 15-30 kDa across a wide range of charge states, which is extremely useful for the dTDP of complex protein mixtures in the widely used data-dependent acquisition (DDA) mode. The system produced over 60% (high, +21) to 88% (low, +17) backbone cleavage coverage for SOD, nearly 70% (high, +22) to 90% (low, +16) backbone cleavage coverage for myoglobin, and a little bit lower than 60% (high, +31) to 69% (low, +24) backbone cleavage coverage for CA.

3.3.3 Combining AI-ECD fragment ions from different charge states of proteins for improved backbone cleavage coverages

We observed that the generated fragment ions for different charge states of proteins using AI-ECD are complementary to each other. As shown in **Figure 3.8**, AI-ECD fragmentation of the low (+24), medium (+27) and high (+31) charge states of CA produced 269, 273, and 220 fragment ions. Only about 50% of the fragment ions from the medium or high charge state of CA were the same as that from its low charge state. Combining the AI-ECD data from the three charge states of CA produced an 85% backbone cleavage coverage and 454 fragment ions in total, including 88 b ions, 105 c ions, 68 w ions, 80 y ions, and 113 z ions, **Figure 3.9D**. Noticeably, the w ions allowed us to confidently distinguish the isomeric leucine (L) and isoleucine (I) residues at 10 positions in the CA sequence. The backbone cleavage coverage and number of fragment ions from the combined data for CA were 23% (85% *vs.* 69%) and 69% (454 *vs.* 269) higher than that from the low charge state of CA (+24) only. By combining data from the three different charge states, we obtained extremely high backbone cleavage coverages for myoglobin (97%) and SOD (94%), which offered 7% and 6% gains in backbone cleavage

coverage, respectively, compared to the best data of the single charge states, **Figure 3.9B** and **Figure 3.9C**. For ubiquitin, AI-ECD fragmentation of a single charge state (+7) already produced 97% cleavage coverage and 154 fragment ions, including 8 b ions, 54 c ions, 19 y ions, 61 z ions, and 12 w ions, **Figure 3.9A**. The w ions enabled distinguishment of the isomeric I and L residues at 4 positions in the protein sequence. To our best knowledge, our work is one of the first examples of coupling online liquid-phase separation to ESI-MS/MS for top-down MS characterization of protein mixtures with nearly complete backbone cleavages. This work certainly represents the first example of dTDP using CZE-ECD. The data demonstrate that our CZE-AI-ECD Q-TOF platform has a great potential to advance dTDP via offering highly efficient protein separation and extensive protein fragmentation.

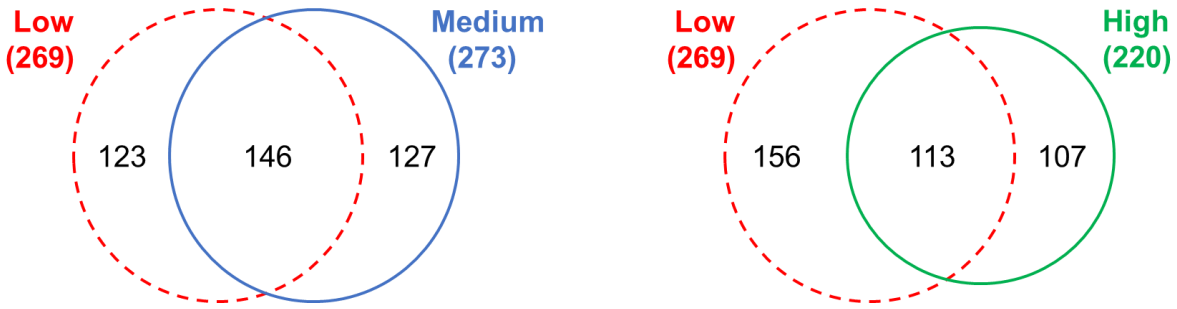

Figure 3.8. Overlap of AI-ECD fragment ions of CA between different charge states (low, medium, and high).

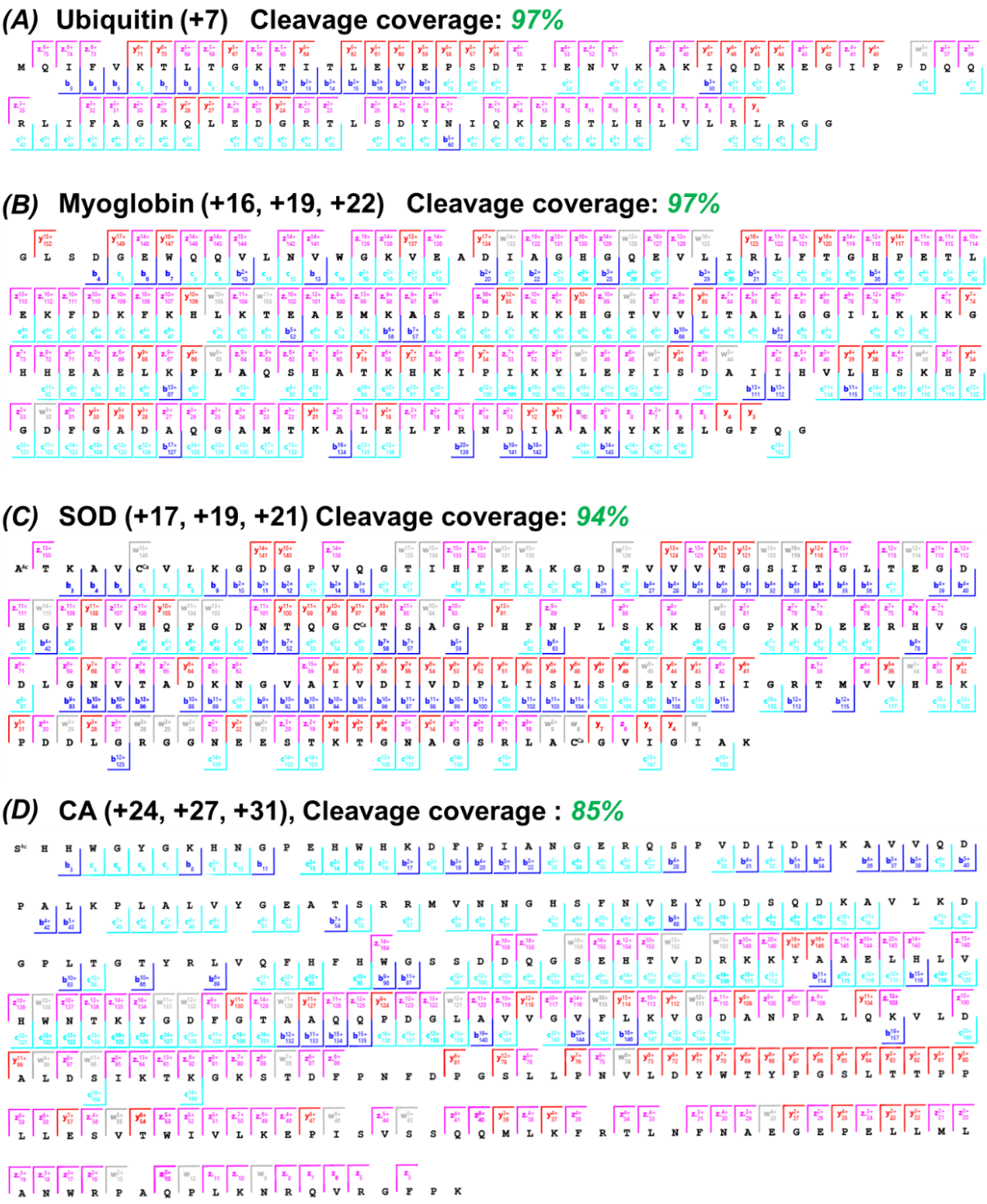

Figure 3.9. Sequences and fragmentation patterns of (A) ubiquitin, (B) myoglobin, (C) SOD and (D) CA obtained using CZE-AI-ECD. For each protein, three charge states were isolated separately

Figure 3.9 (cont'd). for AI-ECD fragmentation and the backbone cleavage coverage were calculated by combining all the fragment ions from the three charge states. B, y, c, z, and w ions were considered for the AI-ECD fragmentation. The CID potential was 10 V. Blue: b ions; red: y ions; cyan: c ions; pink: z ions; grey: w ions.

3.3.4 CZE-ESI-Q-TOF for large proteoform detection from a complex sample

Before our next step to apply the system for large-scale dTDP of a complex sample, we first want to prove that CZE-ESI-Q-TOF is capable to detect proteoforms from complex samples and characterize proteoforms larger than 30 kDa. The *E. coli* proteome (1 mg/mL) was used here to test the system. The low sample loading amount and narrow separation window are two main obstacles that limit CZE-MS/MS for large-scale TDP. To increase the sample loading amount, we utilized dynamic pH junction as the online preconcentration method, which allowed us to inject 500 nL protein sample (about 500 ng protein material) into the capillary. Then we employed a 1.5-m LPA coated capillary for proteoform separation to extend the separation window. As shown in **Figure 3.10**, the CZE separation could generate a 120 min separation window. The CZE-ESI-Q-TOF analysis was reproducible in quadruplicated runs in terms of separation profile and peak intensity. We observed several large proteoforms from the data and **Figure 3.11** shows two examples. The spectrum of the first example (eluting at 71.5 min) is complex and difficult to determine the charge states due to the overlap of signals from several proteoforms (**Figure 3.11A**). However, we could clearly see three proteoform masses about 40 kDa after deconvolution (**Figure 3.11C**). These three proteoforms should belong to one proteoform family as they were co-eluted from CZE; their masses increase by 183 Da, probably the modification by AEBSF protease inhibitor. The second example proteoform was observed at 72.8 min in the electropherogram with a deconvolution mass about 45 kDa (**Figure 3.11B, D**).

The results demonstrated the ability of CZE-ESI-Q-TOF platform for separation and detection of complex sample analysis. The detection of large proteoforms with low signal intensity also showed the high sensitivity of CZE-ESI-Q-TOF system. We will employ AI-ECD in the platform for dTDP of *E. coli* proteome in the future to identify the observed proteoforms.

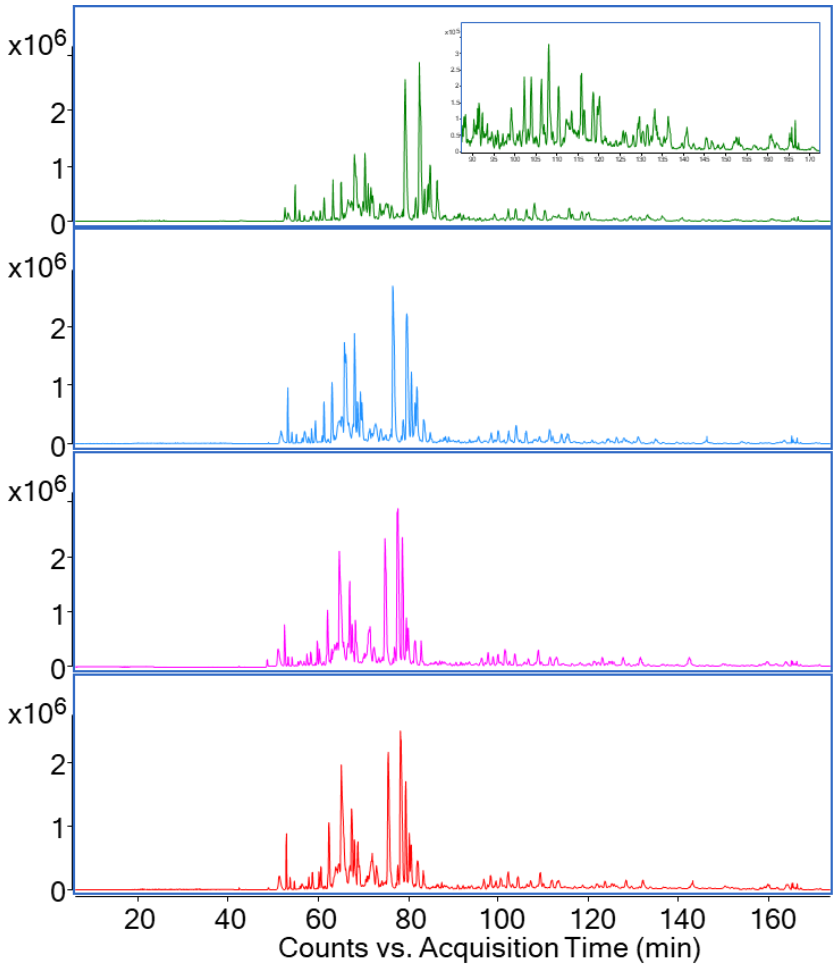

Figure 3.10. Electropherograms of *E. coli* protein sample analyzed by quadruplicated CZE-ESI-Q-TOF with 1.5-m LPA-coated capillary and 500 nL injection volume. The inset is the zoom-in electropherogram from 90-170 min.

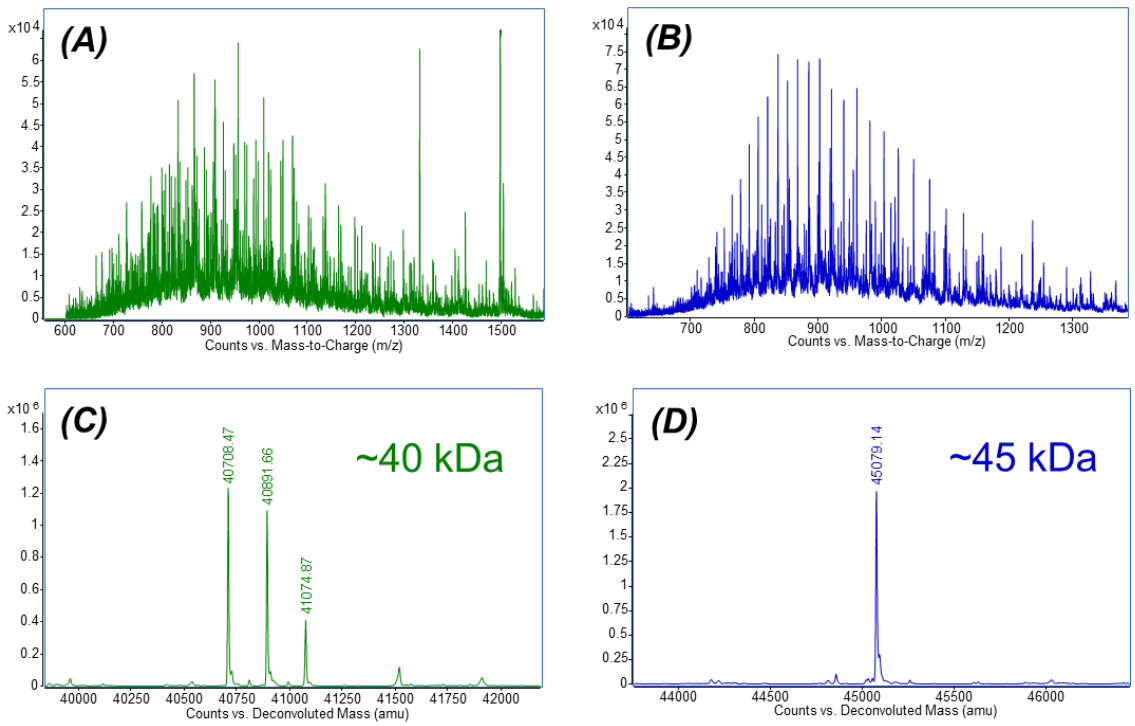

Figure 3.11. MS spectra (A, B) and deconvolution spectra (C, D) of two large proteoforms observed from CZE-ESI-Q-TOF analysis of *E. coli* proteome.

3.4 Conclusions

We presented a novel analytical tool for dTDP of protein mixtures by combining highly efficient CZE separation and extensive AI-ECD fragmentation of proteins on an Agilent 6545XT Q-TOF mass spectrometer. The CID potential and the charge state of proteins could alter the cleavage coverage and the number of sequence-informative fragment ions of proteins from AI-ECD fragmentation significantly. Under an optimized condition, the CZE-AI-ECD-Q-TOF system produced a baseline separation and nearly complete backbone cleavages of a mixture of standard proteins in a mass range of 8-30 kDa. The CZE-ESI-Q-TOF platform also shows high-capacity and robust separation performance and high sensitivity for the complex sample. The

CZE-AI-ECD will be a new tool for dTDP with high utility to advance both the separation and fragmentation of proteins.

However, some improvement in the technique and data analysis software need to be done to allow routine and large-scale dTDP using the CZE-AI-ECD-Q-TOF system. First, we need to enable the real-time mass calibration during CZE-MS on the Agilent Q-TOF system, which will ensure high mass accuracy of fragment ions, improving the confidence of fragment ion matching. Second, spectral averaging is extremely useful for improving the cleavage coverage of proteins from ECD. Incorporation of some spectral averaging function in the available dTDP software packages will allow the automated analysis of the ECD data. Third, the widely used dTDP software packages for proteoform identification via database search were developed mainly based on the data from Orbitrap and FT-ICR mass spectrometers. To analyze the CZE-AI-ECD Q-TOF data, some efforts need to be made to modify the current software tools and better fit the Q-TOF data.

3.5 Acknowledgments

We thank Agilent, e-MSion and CMP Scientific for their kind support for this project. In particular, we thank John Sausen (Director of Strategic Initiatives-Mass Spectrometry) and Dr. David Wong at Agilent and Drs. Joe Beckman, Valery Voinov, Blake Hakkila, Mike Hare at *e-MSion* for their help on the ECD cell and data analysis as well as useful discussions about the AI-ECD data. We thank the support from the National Science Foundation (CAREER Award, Grant DBI1846913) and the National Institutes of Health (Grant R01GM125991).

REFERENCES

REFERENCES

[1] L. M. Smith, N. L. Kelleher, Science 359 (2018) 1106-1107.

[2] X. Yang, J. Coulombe-Huntington, S. Kang, G. M. Sheynkman, T. Hao, A. Richardson, S. Sun, F. Yang, Y. A. Shen, R. R. Murray, K. Spirohn, B. E. Begg, M. Duran-Frigola, A. MacWilliams, S. J. Pevzner, Q. Zhong, S. A. Trigg, S. Tam, L. Ghamsari, N. Sahni, S. Yi, M. D. Rodriguez, D. Balcha, G. Tan, M. Costanzo, B. Andrews, C. Boone, X. J. Zhou, K. Salehi-Ashtiani, B. Charloteaux, A. A. Chen, M. A. Calderwood, P. Aloy, F. P. Roth, D. E. Hill, L. M. Iakoucheva, Y. Xia, M. Vidal, Cell 164 (2016) 805-817.

[3] Y. I. Li, B. van de Geijn, A. Raj, D. A. Knowles, A. A. Petti, D. Golan, Y. Gilad, J. K. Pritchard, Science 352 (2016) 600–604.

[4] H. A. Costa, M. G. Leitner, M. L. Sos, A. Mavrantoni, A. Rychkova, J. R. Johnson, B. W. Newton, M. C. Yee, F. M. De La Vega, J. M. Ford, N. J. Krogan, K. M. Shokat, D. Oliver, C. R. Halaszovich, C. D. Bustamante, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 13976–13981.

[5] I. Ntai, L. Fornelli, C. J. DeHart, J. E. Hutton, P. F. Doubleday, R. D. LeDuc, A. J. van Nispen, R. T. Fellers, G. Whiteley, E. S. Boja, H. Rodriguez, N. L. Kelleher, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) 4140–4145.

[6] T. Jenuwein, Science 293 (2001) 1074-1080.

[7] T. K. Toby, L. Fornelli, N. L. Kelleher, Annu. Rev. Anal. Chem. 9 (2016) 499-519.

[8] B. Chen, K. A. Brown, Z. Lin, Y. Ge, Anal. Chem. 90 (2017), 110-127.

[9] W. Cai, T. Tucholski, B. Chen, A.J. Alpert, S. McIlwain, T. Kohmoto, S. Jin, Y. Ge, Anal. Chem. 89 (2017) 5467–5475.

[10] C. Ansong, S. Wu, D. Meng, X. Liu, H. M. Brewer, B.L. Deatherage Kaiser, E.S. Nakayasu, J.R. Cort, P. Pevzner, R.D. Smith, F. Heffron, J.N. Adkins, L. Pasa-Tolic, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 10153–10158.

[11] J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman, K. R. Durbin, J. D. Tipton, A. Vellaichamy, J. F. Kellie, M. Li, C. Wu, S. M. M. Sweet, B. P. Early, N. Siuti, R. D. LeDuc, P. D. Compton, P. M. Thomas, N. L. Kelleher, Nature 480 (2011) 254–258.

[12] Y. Shen, N. Tolić, P. D. Piehowski, A. K. Shukla, S. Kim, R. Zhao, Y. Qu, E. Robinson, R. D. Smith, L. Paša-Tolić, J. Chromatogr. A 1498 (2017) 99-110.

[13] A. D. Catherman, K. R. Durbin, D. R. Ahlf, B. P. Early, R. T. Fellers, J. C. Tran, P. M. Thomas, N. L. Kelleher, Mol. Cell. Proteomics 12 (2013) 3465-3473.

[14] L. C. Anderson, C. J. DeHart, N. K. Kaiser, R. T. Fellers, D. F. Smith, J. B. Greer, R. D. LeDuc, G. T. Blakney, P. M. Thomas, N. L. Kelleher, C. L. J. Proteome Res. 16 (2016) 1087-1096.

[15] L. V. Schaffer, J. W. Rensvold, M. R. Shortreed, A. J. Cesnik, A. Jochem, M. Scalf, B. L. Frey, D. J. Pagliarini, L. M. Smith, J. Proteome Res. 17 (2018) 3526-3536.

[16] D. Yu, Z. Wang, K. A. Cupp-Sutton, X. Liu, S. Wu, J. Am. Soc. Mass Spectrom. 30 (2019) 2502-2513.

[17] L. Fornelli, K. R. Durbin, R. T. Fellers, B. P. Early, J. B. Greer, R. D. LeDuc, P. D. Compton, N. L. Kelleher, J. Proteome Res. 16 (2017) 609–618.

[18] R. Aebersold, J. N. Agar, I. J. Amster, M. S. Baker, C. R. Bertozzi, E. S. Boja, C. E. Costello, B. F. Cravatt, C. Fenselau, B. A. Garcia, Y. Ge, J. Gunawardena, R. C. Hendrickson, P. J. Hergenrother, C. G. Huber, A. R. Ivanov, O. N. Jensen, M. C. Jewett, N. L. Kelleher, L. L. Kiessling, N. J. Krogan, M. R. Larsen, J. A. Loo, R. R. Ogorzalek Loo, E. Lundberg, M. J. MacCoss, P. Mallick, V. K. Mootha, M. Mrksich, T. W. Muir, S. M. Patrie, J. J. Pesavento, S. J. Pitteri, H. Rodriguez, A. Saghatelian, W. Sandoval, H. Schlüter, S. Sechi, S. A. Slavoff, L. M. Smith, M. P. Snyder, P. M. Thomas, M. Uhlén, J. E. Van Eyk, M. Vidal, D. R. Walt, F. M. White, E. R. Williams, T. Wohlschlager, V. H. Wysocki, N. A. Yates, N. L. Young, B. Zhang, Nat. Chem. Biol. 14 (2018) 206– 214.

[19] R. A. Lubeckyj, E. N. McCool, X. Shen, Q. Kou, X. Liu, L. Sun, Anal. Chem. 89 (2017) 12059-12067.

[20] E. N. McCool, R. A. Lubeckyj, X. Shen, D. Chen, Q. Kou, X. Liu, L. Sun, Anal. Chem. 90 (2018) 5529– 5533.

[21] R. A. Lubeckyj, A. R. Basharat, X. Shen, X. Liu, L. Sun, J. Am. Soc. Mass Spectrom. 30 (2019) 1435-1445.

[22] X. Shen, Z. Yang, E. N. McCool, R. A. Lubeckyj, D. Chen, L. Sun, TrAC, Trends Anal. Chem. 120 (2019) 115644.

[23] X. Han, Y. Wang, A. Aslanian, M. Bern, M. Lavallée-Adam, J. R. Yates, Anal. Chem. 86 (2014) 11006-11012.

[24] R. Haselberg, G. J. de Jong, G. W. Somsen, Anal. Chem. 85 (2013) 2289-2296.

[25] Y. Li, P. D. Compton, J. C. Tran, I. Ntai, N. L. Kelleher, Proteomics 14 (2014) 1158-1164.

[26] D. R. Bush, L. Zang, A. M. Belov, A. R. Ivanov, B. L. Karger, Anal. Chem. 88 (2016) 1138-1146.

[27] Y. Zhao, L. Sun, M. M. Champion, M. D. Knierman, N. J. Dovichi, Anal. Chem. 86 (2014) 4873-4878.

[28] G. A. Valaskovic, N. L. Kelleher, F. W. McLafferty, Science. 273 (1996) 1199-1202.

[29] X. Han, Y. Wang, A. Aslanian, B. Fonslow, B. Graczyk, T. N. Davis, J. R. Yates, J. Proteome Res. 13 (2014) 6078-6086.

[30] O. V. Krokhin, G. Anderson, V. Spicer, L. Sun, N. J. Dovichi, Anal. Chem. 89 (2017) 2000-2008.

[31] D. Chen, K. R. Ludwig, O. V. Krokhin, V. Spicer, Z. Yang, X. Shen, A. B. Hummon, L. Sun, Anal. Chem. 91 (2019) 2201-2208.

[32] D. Chen, R. A. Lubeckyj, Z. Yang, E. N. McCool, X. Shen, Q. Wang, T. Xu, L. Sun, Anal. Chem. 92 (2020) 3503-3507.

[33] E. J. Maxwell, X. Zhong, H. Zhang, N. van Zeijl, D. D. Y. Chen, Electrophoresis 31 (2010) 1130-1137.

[34] R. Wojcik, O. O. Dada, M. Sadilek, N. J. Dovichi, Rapid Commun. Mass Spectrom. 24 (2010) 2554-2560.

[35] L. Sun, G. Zhu, Z. Zhang, S. Mou, N. J. Dovichi, J. Proteome Res. 14 (2015) 2312-2321.

[36] S. B. Choi, M. Zamarbide, M. C. Manzini, P. Nemes, J. Am. Soc. Mass Spectrom. 28 (2016) 597-607.

[37] M. Moini, Anal. Chem. 79 (2007) 4241-4246.

[38] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, D. F. Hunt, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 9528-9533.

[39] N. M. Riley, J. W. Sikora, H. S. Seckler, J. B. Greer, R. T. Fellers, R. D. LeDuc, M. S. Westphall, P. M. Thomas, N. L. Kelleher, J. J. Coon, Anal. Chem. 90 (2018) 8553-8560.

[40] E. N. McCool, J. M. Lodge, A. R. Basharat, X. Liu, J. J. Coon, L. Sun, J. Am. Soc. Mass Spectrom. 30 (2019) 2470-2479.

[41] H Li, H. H. Nguyen, R. R Ogorzalek-Loo, I. D. G. Campuzano, J. A. Loo, Nat. Chem. 10 (2018) 139-148.

[42] V. G. Voinov, M. L. Deinzer, D. F. Barofsky, Rapid Commun. Mass Spectrom. 22 (2008) 3087-3088.

[43] V. G. Voinov, M. L. Deinzer, J. S. Beckman, D. F. Barofsky, J. Am. Soc. Mass Spectrom. 22 (2011) 607-611.

[44] J. B. Shaw, N. Malhan, Y. V. Vasil'ev, N. I. Lopez, A. A. Makarov, J. S. Beckman, V. G. Voinov, Anal. Chem. 90 (2018) 10819-10827.

[45] T. Tucholski, S. J. Knott, B. Chen, P. Pistono, Z. Lin, Y. Ge, Anal. Chem. 91 (2019) 3835-3844.

[46] J. B. Shaw, W. Li, D. D. Holden, Y. Zhang, J. Griep-Raming, R. T. Fellers, B. P. Early, P. M. Thomas, N. L. Kelleher, J. S. Brodbelt, J. Am. Chem. Soc. 135 (2013) 12646-12651.

[47] S. M. Greer, J. S. Brodbelt, J. Proteome Res. 17 (2018) 1138-1145,

[48] L. Fornelli, K. Srzentić, T. K. Toby, P. F. Doubleday, R. Huguet, C. Mullen, R. D. Melani, H. dos Santos Seckler, C. J. DeHart, C. R. Weisbrod, K. R. Durbin, J. B. Greer, B. P. Early, R. T. Fellers, V. Zabrouskov, P. M. Thomas, P. D. Compton, N. L. Kelleher, Mol. Cell. Proteomics 19 (2019) 405-420.

[49] E. N. McCool, D. Chen, W. Li, Y. Liu, L. Sun, Methods 11 (2019) 2855-2861.

[50] R. A. Zubarev, N. L. Kelleher, F. W. McLafferty, J. Am. Chem. Soc. 120 (1998) 3265-3266.

[51] R. A. Zubarev, N. A. Kruger, E. K. Fridriksson, M. A. Lewis, D. M. Horn, B. K. Carpenter, F. W. McLafferty, J. Am. Chem. Soc. 121 (1999) 2857-2862.

[52] D. M. Horn, Y. Ge, F. W. McLafferty, Anal. Chem. 72 (2000) 4778-4784.

[53] Y. Ge, B. G. Lawhorn, M. ElNaggar, E. Strauss, J. H. Park, T. P. Begley, F. W. McLafferty, J. Am. Chem. Soc. 124 (2002) 672-678.

[54] V. Zabrouskov, J. P. Whitelegge, J. Proteome Res. 6 (2007) 2205-2210.

[55] J. P. Williams, L. J. Morrison, J. M. Brown, J. S. Beckman, V. G. Voinov, F. Lermyte, Anal. Chem. 92 (2020) 3674-3681.

[56] K. L. Fort, C. N. Cramer, V. G. Voinov, Y. V. Vasil'ev, N. I. Lopez, J. S. Beckman, A. J. R. Heck, J. Proteome Res. 17 (2018) 926-933.

[57] M. Zhou, W. Liu, J. B. Shaw, Anal. Chem. 92 (2019) 1788-1795.

[58] G. Zhu, L. Sun, N. J. Dovichi, Talanta 146 (2016) 839-843.

[59] E. N. McCool, R. Lubeckyj, X. Shen, Q. Kou, X. Liu, L. Sun, J. Vis. Exp. 140 (2018) e58644.

[60] L. Sun, G. Zhu, Y. Zhao, X. Yan, S. Mou, N. J. Dovichi, Angew. Chem. Int. Ed. 52 (2013) 13661-13664.

[61] F. Kjeldsen, K. F. Haselmann, B. A. Budnik, F. Jensen, R. A. Zubarev, Chem. Phys. Lett. 356 (2002) 201-206.

CHAPTER 4. Native Proteomics in Discovery Mode using Size Exclusion Chromatography-Capillary Zone Electrophoresis-Tandem Mass Spectrometry

4.1 Introduction

Modern proteomics has already approached 10,000 protein identifications (IDs) from mammalian cell lines with bottom-up strategy and obtained thousands of proteoform IDs from human cell lines with top-down strategy [1-7]. However, the majority of proteins in a cell function as protein complexes, and typical bottom-up and top-down strategies cannot directly measure the dynamics of proteomes in cells at protein complex level because the protein-protein interactions are destroyed during sample preparation and reversed-phase liquid chromatography (RPLC)-electrospray ionization (ESI)-tandem mass spectrometry (MS/MS) analysis. Comprehensive characterization of complex proteomes under native conditions, termed native top-down proteomics (native proteomics), will ultimately produce a full picture of endogenous protein complexes in a cell [8].

Native proteomics requires high-resolution and liquid-phase separation of a complex proteome prior to native electrospray ionization (nESI)-MS and MS/MS. NESI-MS has been widely used for the characterization of purified protein complexes, antibodies and virus assemblies via direct infusion [9-16]. Some work has been done using liquid-phase separation-nESI-MS for the

---

Part of this chapter was adapted with permission from: X. Shen, Q. Kou, R. Guo, Z. Yang, D. Chen, X. Liu, H. Hong, L. Sun, Anal. Chem. 90 (2018) 10095–10099.

characterization of standard protein complexes or samples with very low complexity [17-23]. Recently, Skinner *et al.* coupled off-line ion exchange chromatography or clear native gel-eluted liquid fraction entrapment electrophoresis [24] to direct infusion nESI-MS/MS for native proteomics of mouse hearts and four human cell lines, leading to the identification of 164 proteins and 125 protein complexes from 600 fractions [8]. This is the first example of native proteomics. However, the workflow is labor- and time-consuming. Coupling an online and high-resolution separation technique to nESI-MS and MS/MS is required to boost the throughput and scale of native proteomics.

Capillary zone electrophoresis (CZE)-MS/MS has a great potential for native proteomics due to the high separation efficiency of CZE for intact proteins [24-27], the mature CE-MS interfaces [28-31], and its capability for high-resolution separation and highly sensitive detection of protein complexes under native conditions [20,23, 32-34]. Nguyen *et al.* established a native CZE-MS system based on a sheathless CE-MS interface for characterization of protein complexes, detecting carbonic anhydrase II-Zn complex, carbonic anhydrase I-Zn complex and hemoglobin A (tetramer) from human red blood cells (RBCs) [20]. It is worth noting that those three complexes span a concentration dynamic range of ~3 orders of magnitude in RBCs. This work clearly demonstrated the great potential of CZE-MS for highly sensitive characterization of protein complexes in native conditions. Leize-Wagner group recently demonstrated a sheathless CE-MS interface based native CZE-MS system as a powerful and highly sensitive nanoESI infusion platform for the analysis of antibody-drug conjugates, monoclonal antibodies and other various protein complexes [22,35,36]. Very recently, Belov *et al.* established a native CZE-MS and MS/MS platform based on the sheath-less CE-MS interface and applied the platform for the analysis of standard protein complexes, monoclonal antibodies and a ribosomal isolate from *E.*

coli. [23]. Although native CZE-MS has been well recognized as a useful platform for the analysis of protein complexes, there is still no report on evaluating CZE-MS/MS for native proteomics of complex proteomes. In this work, for the first time we established a native SEC-CZE-ESI-MS/MS platform based on the commercialized electro-kinetically pumped sheath flow interface for characterization protein complexes from a complex proteome sample (*E. coli* cell lysate) via native proteomics.

4.2 Experimental

4.2.1 Materials and reagents

Carbonic anhydrase from bovine erythrocytes, pyruvate kinase from rabbit muscle, 3-(Trimethoxysilyl)propyl methacrylate, ammonium persulfate, ammonium acetate (NH_4Ac) and the Microcon-30kDa centrifugal filter units for buffer exchange were purchased from Sigma-Aldrich (St. Louis, MO). Recombinant streptavidin, hydrofluoric acid (HF) and LC/MS grade water were purchased from Fisher Scientific (Pittsburgh, PA). Acrylamide were purchased from Acros Organics (NJ, USA). Bare fused silica capillaries (50-μm i.d., 360-μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ).

4.2.2 Preparation of separation capillary for CZE

The LPA coating was conducted based on the protocol described in references [37] and [38]. A bare fused silica capillary was successively flushed with 1 M hydrochloric acid, water, 1 M sodium hydroxide, water, and methanol, followed by treatment with 3-(trimethoxysilyl) propyl methacrylate to introduce carbon-carbon double bonds on the inner wall of the capillary. The treated capillary was filled with degassed acrylamide solution in water containing ammonium

persulfate, followed by incubation at 50 ˚C water bath for 35 to 40 min with both ends sealed by silica rubber. After that, the capillary was flushed with water to remove the unreacted reagents. The details of the linear carbohydrate polymer (LCP)-coating are discussed in **section 5.2.2**. Then one end of the LPA-coated capillary and the LCP-coated capillary was etched with HF based on the protocol in reference [31] to reduce its outer diameter to around 70 μm.

4.2.3 Sample preparation

E. coli (strain MG1655) was cultured in Lysogeny broth medium at 37 ˚C until OD600 reached 0.7. After washed with PBS three times, the cells were lysed in a PBS buffer plus 10 mM magnesium chloride, 2 mM calcium chloride and complete protease inhibitors (Roche) and homogenized for 30 s, followed by sonication with a Branson Sonifier 250 (VWR Scientific, Batavia, IL) on ice for 2 min. After centrifugation, the supernatant containing the extracted proteins was collected. A small aliquot of the diluted sample was used for the bicinchoninic acid (BCA) assay to determine the protein concentration.

One aliquot of the *E. coli* lysate containing about 600 μg of proteins (~2 mg/mL) was fractionated with size exclusion chromatography (SEC) on an Agilent Infinity II HPLC system. The AdvanceBio SEC column (4.6 x 300 mm, 2.7 μm particles, 300 Å pores) was from Agilent. The mobile phase was 100 mM NH4Ac (pH 7.0), and the flow rate was 0.15 mL/min. Eight fractions were collected from 11-19 min (1 min for each fraction) for relatively small proteins based on our preliminary experiment. Then each fraction was loaded onto a Microcon-30 kDa centrifugal filter unit, respectively, followed by centrifugation to remove the lysis buffer. We washed the membrane with 50 mM NH_4Ac (pH 6.9) for buffer exchange, followed by adding 40 μL of 50 mM NH_4Ac (pH 6.9) into each filter unit to extract the proteins on the membrane. We gently vortexed the filter units for 5 min and took the protein solution from the filter units for

native CZE-MS/MS analysis. The use of Microcon-30 kDa centrifugal filter unit for buffer exchange was based on the recent native proteomics work from the Kelleher group [8].

The mixture of standard protein complexes was prepared with carbonic anhydrase and its impurity superoxide dismutase (CA and SOD, 1 mg/mL), pyruvate kinase (PK, 5 mg/mL) and streptavidin (2 mg/mL). The protein complex mixture was purified and buffer exchanged into 10 mM NH_4Ac (pH 6.9) by 3-time centrifugation with a Microcon-100 kDa centrifugal filter unit.

4.2.4 SDS-PAGE

In order to evaluate the sample loss during the buffer exchange with Microcon-30 kDa centrifugal filter units, we analyzed the *E. coli* whole cell lysate before and after the buffer exchange as well as the flow through using SDS-PAGE. About 400 µg of *E. coli* proteins in 50 µL of the lysis buffer were loaded onto one membrane filter, followed by centrifugation at 10 000 *g* for 10 min. The membrane was washed with 100 µL of 50 mM NH_4Ac (pH 6.9). After centrifugation, 100 µL of 50 mM NH_4Ac (pH 6.9) was added onto the membrane to extract the proteins. The membrane filter was gently vortexed for 5 min. After that, the protein solution on the membrane was collected and lyophilized to about 50 µL for the SDS-PAGE experiment. The flow-through during the buffer exchange (~150 µL) was collected and lyophilized to about 50 µL for the SDS-PAGE experiment. We performed the buffer exchange experiment twice as technical duplicate. The samples from the technical duplicate were loaded onto an SDS-PAGE gel for analysis. Two microliters of the *E. coli* sample before and after the buffer exchange (~16 µg of proteins in theory) and 2 µL of the flow-through sample were analyzed by SDS-PAGE. The Coomassie blue solution was used for staining.

4.2.5 Native CZE-ESI-MS and MS/MS analysis for *E. coli* proteome

An ECE-001 capillary electrophoresis autosampler (CMP Scientific, Brooklyn, NY) was used for automated operation of CZE. A commercialized electrokinetically pumped sheath flow interface (CMP Scientific) was used to couple CZE to MS. A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used for the experiments. The commercialized electrokinetically pumped sheath flow interface (CMP Scientific) was directly attached to the Q-Exactive HF mass spectrometer for experiments. The ESI emitters of the CZE-MS interface were pulled from borosilicate glass capillaries (1.0 mm o.d., 0.75 mm i.d., 10 cm length) with a Sutter P-1000 flaming/brown micropipet puller. The opening size of the ESI emitters was 20 μm. The spray emitter with ~4 cm length was typically used. The voltage for ESI was ~2 kV.

A 1-m LPA-coated capillary (50-μm i.d. and 360-μm o.d.) was used for the CZE separation of *E. coli* samples. A 70-cm linear carbohydrate polymer (LCP)-coated capillary was used for the CZE separation of the mixture of standard protein complexes. The background electrolyte (BGE) for CZE was 50 mM NH_4Ac (pH 6.9), and the sheath buffer was 25 mM NH_4Ac (pH 6.9). 15 kV was applied at the sample injection end and 1 psi was applied at the meantime for CZE separation. The *E. coli* sample was injected into the separation capillary for CZE-MS/MS with 5-psi pressure for 20 s.

A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used for all of the experiments. The transfer capillary temperature was 200 ^{o}C, and the S-lens RF level was 50. A top 3 data-dependent acquisition (DDA) method was used. The number of microscans was 3 for both MS and MS/MS. The resolution for MS and MS/MS was 240,000 and 120,000 (*m/z* 200), respectively. The AGC target was 3E6 for MS and 1E6 for MS/MS. The maximum injection time was 200 ms for MS and 500 ms for MS/MS. The mass range for MS scans was 1000-4000

m/z. Three most abundant protein peaks in the mass spectrum were sequentially isolated with isolation window as 4 *m/z* in the quadrupole, followed by fragmentation with normalized collisional energy (NCE) as 35. "Exclude isotopes" was turned on and the dynamic exclusion was 30 s.

4.2.6 Native CZE-ESI-MS for the mixture of standard protein complexes

A 7100 CE System from Agilent Technologies (Santa Clara, CA) was used for automated operation of CZE. The interface setting was the same as native CZE-ESI-MS analysis for *E. coli* proteome unless labelled otherwise. The opening size of the ESI emitters was 30-40 μm. Voltage for ESI ranged from +2.2 to +2.5 kV.

A 70-cm linear carbohydrate polymer (LCP)-coated capillary and a 70-cm LPA-coated capillary was used for the CZE separation of the mixture of standard protein complexes. The background electrolyte (BGE) for CZE was 25 mM NH_4Ac (pH 6.9), and the sheath buffer was 10 mM NH_4Ac (pH 6.9). 30 kV was applied at the sample injection end and 50 mbar was applied at the mean time for CZE separation. The protein complex mixture was injected into the separation capillary for CZE-MS/MS with 100 mbar pressure. For LPA experiment, 50 nL sample was injected for CZE separation, while for LCP experiment, 30 nL sample was injected for CZE separation. The injection volume was calculated based on Poiseuille's law.

A 6545XT AdvanceBio LC/Q-TOF mass spectrometer (Agilent Technologies, Santa Clara, CA) with an electromagnetostatic ExD cell (e-MSion, Corvallis, OR) was used for the experiments. The ExD cell was set for positive transmission without ECD fragmentation (ECD off). CID potential of 10 V was required for the maximum transmission efficiency of mAbs in the system with the ExD cell. A regular ESI spray shield was used in the experiments. The gas

temperature and flow rate of nitrogen drying gas was 300 ^{o}C and 1 L/min. The fragmentor voltage was 300 V, and the skimmer voltage was 250 V. The voltage applied on the ion transfer capillary was 0 V. The mass range option was set as High (1000 *m/z*). The slicer mode was High Resolution. The mass range of detection was 3000-12000 *m/z*, and the scan rate was 0.25 spectrum/sec.

4.2.7 Data analysis

The RAW files were first converted into mzML files using Msconvert tool [39]. Then, TopFD (TOP-Down Mass Spectrometry Feature Detection) was used for the spectral deconvolution to produce msalign files. TopFD (http://proteomics.informatics.iupui.edu/software/toppic/) is an improved version of MS-Deconv [40]. Finally, TopPIC (version 1.1.3) [41] was used for database search with msalign files as input. The *E. coli* UniProt database (UP000000625, 4307 entries) was used for database search. The false discovery rates (FDRs) were estimated using the target-decoy approach [42]. The database search parameters were as follows: the maximum number of unexpected modifications as 2, the precursor and fragment mass error tolerances as 15 ppm, and the mass shift of unknown modifications as -200 to 10000 Da. In order to reduce the redundancy of proteoform identifications (IDs), proteoforms identified by multiple spectra were considered as one proteoform ID if those spectra match the same proteoform feature reported by TopFD or those proteoforms belong to the same protein and have similar precursor masses (within 1.2 Da).

Two rounds of analyses were performed. TopPIC was employed to search each raw file against the *E. coli* database separately, and no filter was applied in this step. Then, all the proteoform spectrum-matches (PrSMs) identified from the 8 SEC fractions were combined and

filtered out with a 1% spectrum-level FDR. The identified proteoforms were further filtered with a 5% proteoform-level FDR.

Deconvolution of standard protein complexes was performed using Agilent MassHunter BioConfirm 10.0 using Maximum Entropy algorithm. The mass step was 0.5 Da. Other parameters for deconvolution were set as default.

4.2.8 Workflow for identification of protein complexes

First, we performed a regular data-dependent acquisition (DDA) experiment on the fractionated *E. coli* samples to acquire MS and MS/MS spectra of the proteins and protein complexes. We isolated a protein or a whole protein complex with the quadrupole, followed by HCD fragmentation of the protein or protein complex. Second, we performed a database search of the acquired MS and MS/MS spectra using TopPIC to identify proteoforms.

Third, we believe if one proteoform is a complex with some co-factors, there should be a detected mass shift that matches with the mass of the co-factor after database search. We obtained a potential protein cofactor list from the UniProt *E. coli* database, as shown in **Table 4.1**. Here we take RNA polymerase-binding transcription factor DksA as an example. We identified this protein by TopPIC and obtained the proteoform as shown in **Figure 4.1**. We found it has an unknown modification of ~63.5 Da, which is close to the average isotopic mass of zinc or copper. We think the proteoform should be a potential protein complex with a zinc ion or a copper ion.

Table 4.1. The names and masses of the major protein co-factors in the UniProt *E. coli* database.

Cofactor	MW (Da)
Mg(2+)	24
Chloride	35.5
K(+)	39
Ca(2+)	40
Mn(2+)	55
Fe(2+)	56
Ni(2+)	58.7
Co(2+)	59
Hydrogencarbonate	61
Cu(2+)	64
Zn(2+)	65
pyruvate	87.05
[2Fe-2S] cluster	175.8
(R)-lipoate	206
pyridoxal 5'-phosphate	245.126
[3Fe-4S] cluster	296
pyrroloquinoline quinone	327.182
[4Fe-4S] cluster	352
pantetheine 4'-phosphate	356.333
dipyrromethane	416
thiamine diphosphate	422.29
FMN	453.321
FMNH2	456.344
Mo-molybdopterin	519.26
heme b	614.471
NAD(+)	663.43
NADP(+)	744.41
FAD	782.5
siroheme	908.597
methylcob(III)alamin	1344.38
adenosylcob(III)alamin	1579.58
Mo-bis(molybdopterin guanine dinucleotide)	1584.99

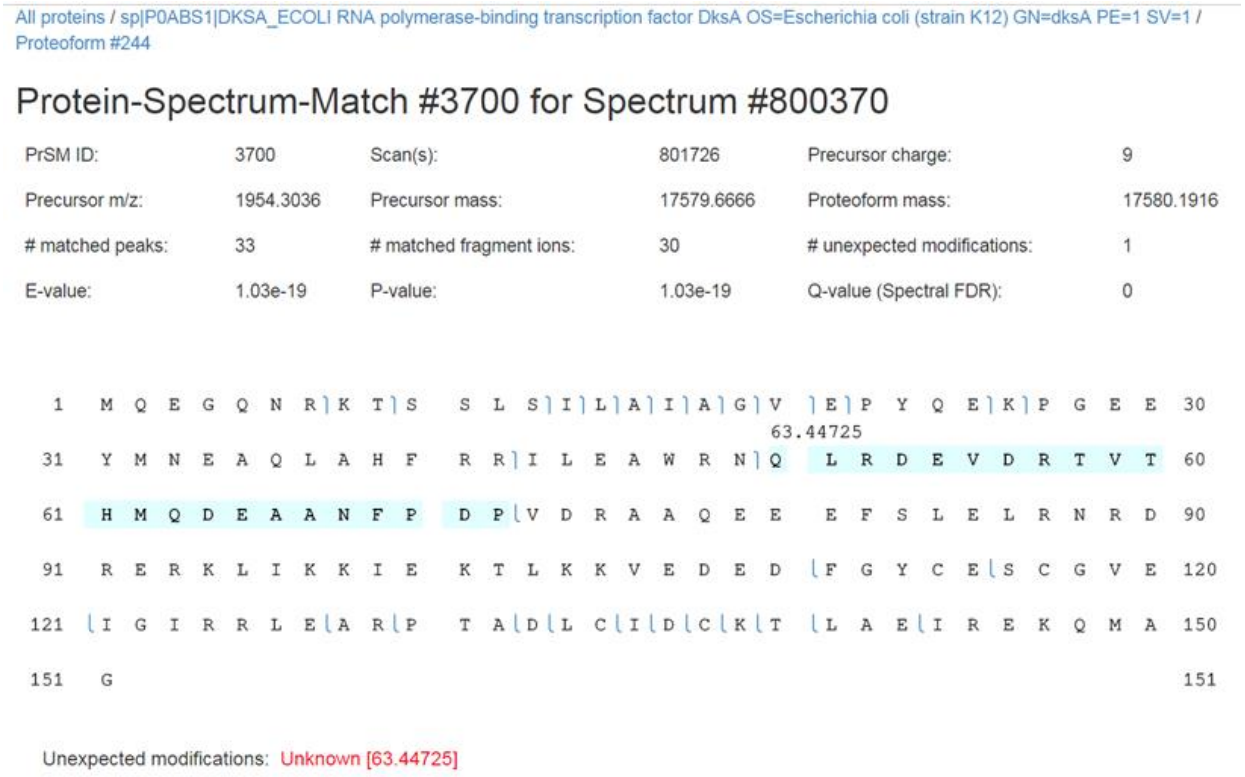

Figure 4.1. The sequence of the RNA polymerase-binding transcription factor DksA, the observed fragmentation pattern, and the mass shift detected through the database search.

Fourth, in order to confirm this modification (+63.5 Da) is not an unusual covalent modification, we compared the proteoform with our recently published large-scale top-down proteomics dataset of *E. coli* under a denaturing condition [43]. If the proteoform matches with some proteoform identified under the denaturing condition in terms of the mass shift within a 4-Da mass tolerance, we think the modification (+63 Da) should be some covalent modification and the proteoform is not a protein-metal complex. If we did not observe any proteoform similar to the proteoform identified in this work, we conclude the proteoform should be a protein complex.

Finally, we went back to UniProt and tried to seek some information in the literature on the protein complex. In this case, we found the RNA polymerase-binding transcription factor DksA had been reported to bind with a zinc ion and has no other modifications of the same mass. Then we conclude the identification of the protein complex with a zinc ion. If we did not get literature information of some protein complexes, we reported those protein complexes as unreported protein complexes.

For the identification of homodimers, we used the similar workflow. Because the mass shift in this case is very big, it should not correspond to a co-factor. If the mass shift of some proteoform is 50% of the detected proteoform mass, we think the proteoform should represent a homodimer. We considered hetero-oligomers in the experiment via manually evaluating the proteoforms with large mass shifts but we only found small homodimers.

4.3 Results and discussion

4.3.1 Native proteomics of *E. coli* proteome with SEC-CZE-ESI-MS/MS

In this work, we coupled size exclusion chromatography (SEC) prefractionation to online CZE-MS/MS for native proteomics in discovery mode, **Figure 4.2**. *E. coli* cells were lysed in PBS buffer. The extracted proteins were fractionated with SEC into 8 fractions. The mobile phase was 100 mM ammonium acetate (NH_4Ac, pH 7.0). After simple protein concentration and buffer exchange with Microcon-30 kDa centrifugal filter units, the SEC fractions were analyzed by CZE-MS/MS. We evaluated the sample loss during the buffer exchange using SDS-PAGE, **Figure 4.3**. We did not observe significant differences in protein abundance before and after the buffer exchange. The protein abundance in the flow-through sample was ignorable compared

with the original sample. For the CZE-MS/MS, the commercialized electro-kinetically pumped sheath flow interface (CMP Scientific, Brooklyn, NY) was used to couple CZE to MS. The background electrolyte (BGE) and the sheath buffer were 50 mM NH_4Ac (pH 6.9) and 25 mM NH_4Ac (pH 6.9), respectively. A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used. The SEC-CZE-MS/MS platform is straightforward. The CZE-MS/MS analyses of the 8 SEC fractions took 16 h. One example electropherogram is shown in **Figure 4.2B**. The native CZE-MS/MS run obtained 15 major peaks and approached a 1-h separation window. TopPIC software was used for the database search of the acquired MS/MS spectra for proteoform identification [41,42]. We note that the Q-Exactive HF cannot isolate ions with the mass-to-charge ratio (*m/z*) higher than 2500 for fragmentation. In this proof-of-principle work, we focused on the identification of protein complexes with mass lower than 30 kDa.

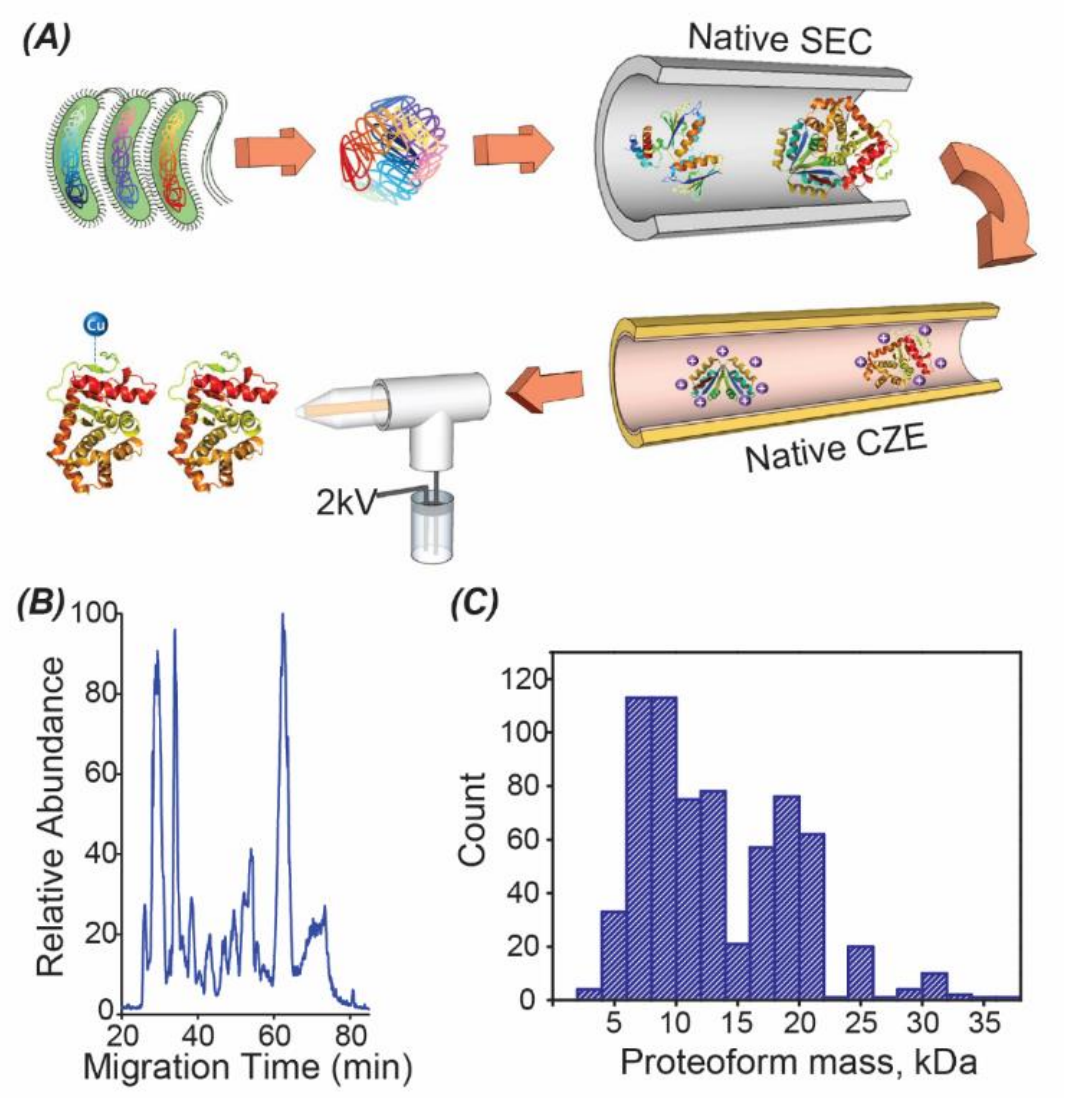

Figure 4.2. (A) The SEC-CZE-ESI-MS/MS platform for native proteomics. (B) An example base peak electropherogram of an SEC fraction of the *E. coli* lysate after CZE-MS/MS analysis. (C) The mass distribution of the identified proteoforms from the *E. coli* proteome.

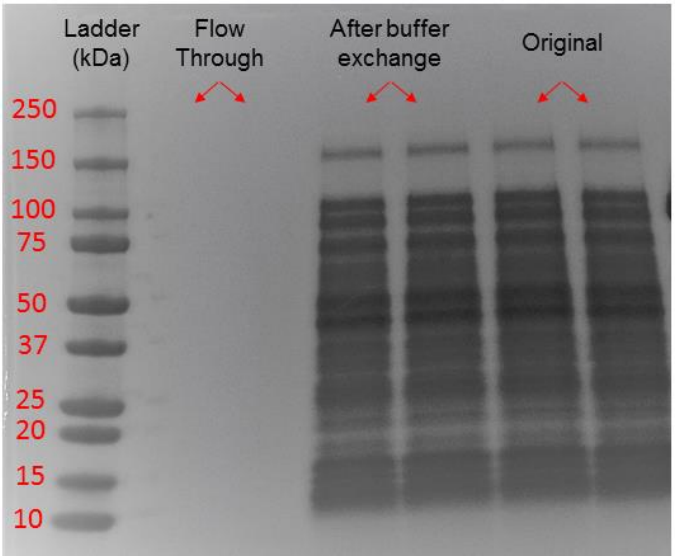

Figure 4.3. Image of the SDS-PAGE results. *E. coli* cell lysate before (Original) and after the buffer exchange with Microcon-30 kDa centrifugal filter units were analyzed by SDS-PAGE. About 16 µg of proteins were loaded in theory. The flow-through during buffer exchange was also analyzed. The buffer exchange experiment was performed in technical duplicate and the data were shown as the two channels.

A total of 144 proteins and 672 proteoforms were identified from the *E. coli* lysate with a 1% spectrum-level false discovery rate (FDR) and a 5% proteoform-level FDR. The number of protein identifications from each SEC fraction and the protein-level overlap between adjacent SEC fractions are shown in **Figure 4.4**. The data indicate that SEC can reach a reasonable protein separation under the native condition. Most of the identified proteoforms have mass lower than 30 kDa, **Figure 4.2C**.

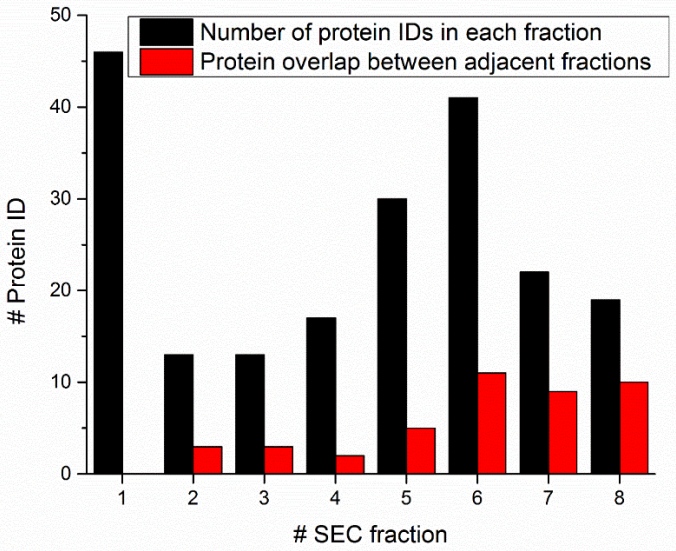

Figure 4.4. The number of protein identifications (IDs) from each SEC fraction and protein overlaps between adjacent SEC fractions.

4.3.2 Identification of protein complexes from *E. coli* proteome

23 protein complexes from 17 proteins were identified, including four homodimers, 16 protein-metal complexes, two protein-[2Fe-2S] complexes, and one protein-glutamine complex. 14 out of the 23 protein complexes have not been reported before. The details of those protein complexes are listed in **Table 4.2**. The SEC-CZE-MS/MS performed native proteomics in discovery mode because the identities of protein complexes were unknown before analysis.

The protein complexes were identified through several steps. First, during CZE-MS/MS analysis, a protein or a whole protein complex was isolated in the quadrupole, followed by high energy collision dissociation (HCD). The MS and MS/MS spectra of the proteins and protein complexes were acquired. Second, proteoforms were identified through the database search of the acquired MS and MS/MS spectra against a UniProt *E. coli* database with the TopPIC software. Third, the mass shifts of the identified proteoforms were compared with the masses of

known protein co-factors in the UniProt *E. coli* database manually. The co-factors are shown in **Table 4.1**. If they matched with each other within a 4-Da mass difference, we assumed that the mass shift corresponded to the specific cofactor. We obtained a list of proteoforms that were potential protein complexes with the cofactors. Fourth, we compared those proteoforms with that identified in our recent deep top-down proteomics work. We identified nearly 6000 proteoforms and 850 proteins from the *E. coli* proteome using denaturing top-down approach [45]. Because of the denaturing conditions, the non-covalently bound cofactors were lost during that experiment. If the potential protein complexes detected in this work matched well with some proteoforms in reference [43], the corresponding mass shifts should represent some covalent modifications. Those potential protein complexes were removed from the initial protein complex list. After this step, we obtained a list of identified protein complexes with bound cofactors. Finally, we searched the UniProt database to find information on those identified protein complexes in the literature. Protein complexes without literature information were considered as unreported protein complexes. The workflow using RNA polymerase-binding transcription factor DksA-zinc complex as an example was described in **section 4.2.7**. **Figure 4.5A** shows one deconvoluted spectrum of the DksA-zinc complex. **Figure 4.1** shows the sequence, observed fragmentation pattern, and detected mass shift of the DksA-zinc complex after the database search.

Table 4.2. The list of the identified protein complexes with the SEC-CZE-MS/MS from the *E. coli* proteome.

Protein complex	UniProt accession #	Protein name	Mass difference (observed-theoretical, Da)	First amino acid	Last amino acid	E-value	**Unreported**
Homodimer	P0AES9	Acid stress chaperone HdeA	18	22	110	4.51E-13	X
	P0AES9	Acid stress chaperone HdeA	-3.5	22	110	1.06E-11	
	P0AES9	Acid stress chaperone HdeA	58	22	110	1.29E-06	X
	P0AA04	Phosphocarrier protein HPr	-0.31	1	85	1.35E-25	X
Zinc ion binding	P0AAZ7	UPF0434 protein YcaR	-2.2	1	60	1.74E-07	
	P0ABS1	RNA polymerase-binding transcription factor DksA	-1.9	1	151	1.03E-19	
	P0AEG4	Thiol:disulfide interchange protein DsbA	-3.7	20	208	1.66E-18	
	P0AEG4	Thiol:disulfide interchange protein DsbA	0.74	146	208	1.33E-09	X
Copper ion binding	P0AA25	Thioredoxin 1	-0.23	2	109	6.17E-10	
	P64534	Nickel/cobalt homeostasis protein RcnB	1.5	27	112	2.06E-12	
	P0AA57	Protein YobA	-1.6	27	124	1.40E-11	
Zinc/ copper ion binding	P09372	Protein GrpE	-2.7/-0.89	2	197	7.85E-25	X
	P0A800	DNA-directed RNA polymerase subunit omega	-0.43/1.4	26	91	6.54E-08	X
	P0A9X9	Cold shock protein CspA	-0.40/1.4	2	70	1.70E-10	X
	P0AA04	Phosphocarrier protein HPr	0.57/2.4	1	85	1.67E-07	X
	P0AC59	Glutaredoxin 2	-0.62/1.2	1	215	9.29E-08	X
	P0ADU5	Protein YgiW	-3.1/-1.2	21	130	7.51E-16	X
	P0AEQ3	Glutamine-binding periplasmic protein	-1.6/0.26	23	248	2.35E-09	X
	P0AF36	Cell division protein ZapB	0.58/2.4	4	81	1.47E-21	X
	P76402	UPF0339 protein YegP	-1.9/-0.060	2	110	1.23E-14	X
[2Fe-2S] binding	P0A9R4	2Fe-2S ferredoxin	-2.6	2	111	4.55E-10	
	P0A9R4	2Fe-2S ferredoxin	21	2	111	3.49E-08	X
Glutamine binding	P0AEQ3	Glutamine-binding periplasmic protein	-0.23	23	248	1.61E-14	

Using this approach, we identified 17 protein complexes including 16 protein-zinc/copper complexes and one protein-[2Fe-2S] complex, **Table 4.2**. Besides the detected 2Fe-2S ferredoxin

complex, we observed another form of the complex with additional 21-Da modification. Seven out of the identified protein complexes have been reported in the literature including YcaR-zinc complex [44], DksA-zinc complex [45], DsbA-zinc complex [46], thioredoxin 1-copper complex [47], RcnB-copper complex [48], YobA-copper complex [49], and ferredoxin-[2Fe-2S] complex [50]. The data clearly indicate that those non-covalent interactions can be preserved during the SEC-CZE-MS analysis. The 11 unreported protein complexes include the 2Fe-2S ferredoxin complex with additional 21-Da modification, a truncated DsbA-zinc complex, and protein-metal complexes from nine novel zinc/copper-binding proteins.

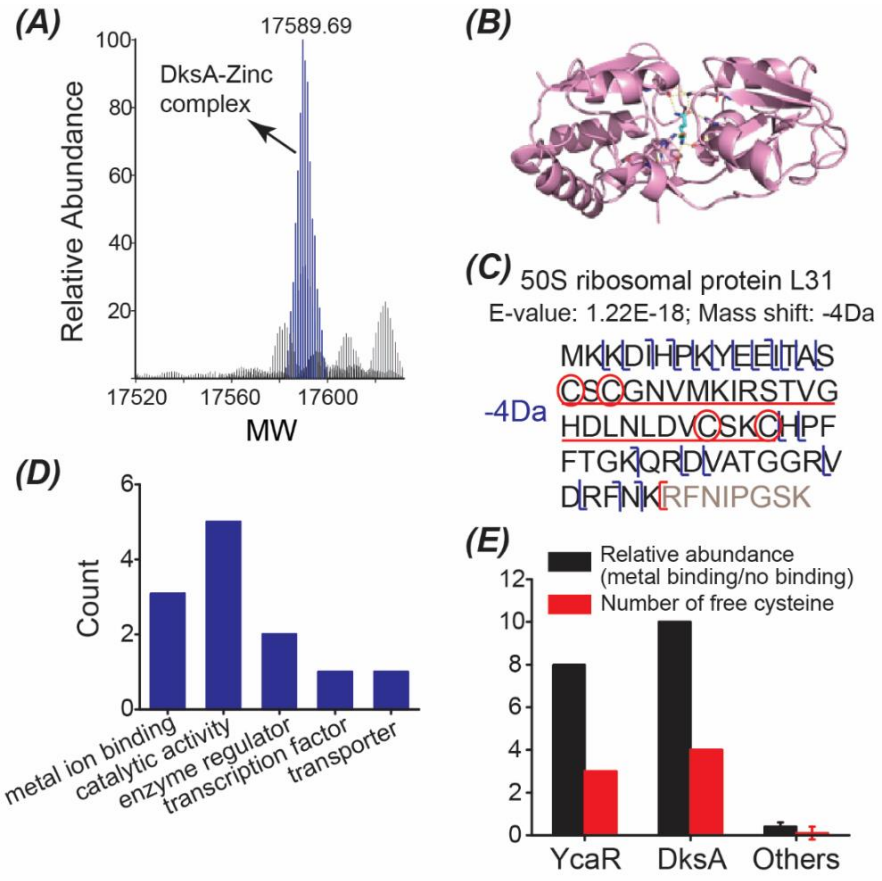

Figure 4.5. (A) One deconvoluted spectrum of the identified RNA polymerase-binding transcription factor DksA-zinc complex. The averaged mass spectrum across the peak of the

complex was used for the mass deconvolution with the Xtract software (Thermo Fisher Scientific) using the default settings. The x-axis is molecular weight (MW). (B) The crystal structure of glutamine-binding periplasmic protein bound with a glutamine molecule. The image of the crystal structure was obtained from the Protein Data Bank in Europe (https://www.ebi.ac.uk/pdbe/). (C) The sequence, observed fragmentation pattern, and detected mass shift of the 50S ribosomal protein L31 through the database search. The location of the mass shift and the cysteine amino acids are highlighted. (D) The molecular function distribution of the identified metalloproteins. The Retrieve/ID mapping tool on the UniProt website (http://www.uniprot.org/uploadlists/) was used to obtain the molecular function information. (E) The metal binding stoichiometry of some identified metalloproteins. The detailed information is shown in **Table 4.3**. The error bars for "Others" represent the standard deviations of relative abundance and cysteine count from 13 metalloproteins.

We noted that the mass spectrometer used in this work limited our capability to localize the protein co-factors in the protein sequences. For example, as shown in **Figure 4.1**, the zinc ion is localized between the 50^{th} and 72^{nd} amino acids based on the database search result. However, the zinc ion should bind with the four cysteine amino acids at positions 114,117,135, and 138 based on the UniProt database. During the HCD fragmentation, the zinc ion and the DksA protein fell apart, leading to a challenge for accurately localizing the zinc ion in the protein sequence. During the database search with the TopPIC, the mass shift corresponding to the co-factor was assigned to a region that no fragment ion could cover. In order to improve the localization of protein co-factors, we will employ mass spectrometers with electron transfer dissociation (ETD) [51] or electron capture dissociation (ECD) [16] in our future work.

Table 4.3. The metal binding stoichiometry of some identified metalloproteins.

Cofactor	Protein name	Relative abundance (metal binding/no binding)*	Number of C/H/D/E in the protein sequence**
Zinc ion	UPF0434 protein YcaR	>8.0	3/1/4/6
	RNA polymerase-binding transcription factor DksA	>10	4/2/10/22
	Thiol:disulfide interchange protein DsbA	0.50	2 (S-S)/3/12/12***
Copper ion	Thioredoxin 1	0.60	2(S-S)/1/11/5***
	Nickel/cobalt homeostasis protein RcnB	0.70	0/3/7/4
	Protein YobA	0.20	0/6/4/4
Zinc/copper ion	Protein GrpE	0.60	0/3/13/26
	DNA-directed RNA polymerase subunit omega	0.20	0/0/5/12
	Cold shock protein CspA	0.80	0/1/6/2
	Phosphocarrier protein HPr	0.50	0/2/1/9
	Glutaredoxin 2	0.20	2 (S-S)/4/19/9***
	Protein YgiW	0.20	1/1/11/6
	Glutamine-binding periplasmic protein	0.30	0/2/22/10
	Cell division protein ZapB	0.30	0/2/1/16
	UPF0339 protein YegP	0.10	0/1/2/8

* The relative abundance was calculated based on the intensity of the proteoforms with and without metal binding. The averaged mass spectra across the proteoform peaks were used for the calculation. ** C for cysteine, H for histidine, D for aspartic acid, and E for glutamic acid. *** The two cysteine amino acids form a disulfide bond based on the database search results and/or the UniProt *E. coli* database.

4.3.3 Identification of homodimers from *E. coli* proteome

We identified four homodimers from two proteins, acid stress chaperone HdeA and phosphocarrier protein HPr, **Table 4.2**. The masses of these homodimers are 19487 Da (HdeA), 19466 Da (HdeA), 19527 Da (HdeA), and 18227 Da (HPr). For those two proteins, the mass shifts of some proteoforms are about 50% of the proteoform mass. For example, a mass shift of 9731 Da was detected from one HdeA proteoform that had a mass of 19466 Da. The proteoform

is the homodimer of HdeA. The data agree well with the literature [52]. HdeA is homodimer at neutral pH and dissociates into monomer at pH 4. Using the same approach, we detected the homodimer of phosphocarrier protein HPr. Another two forms of HdeA homodimer were identified with additional 18-Da and 58-Da modifications. Those three protein complexes have not been reported previously. We identified one proteoform of glutamine-binding periplasmic protein (glnH) with a mass shift of 146 Da. The mass shift matches well with the mass of glutamine. The proteoform represents the glnH-glutamine complex. GlnH is involved in glutamine transport. One crystal structure of the glnH-glutamine complex has been reported [53], **Figure 4.5B**. The data highlight the capability of our platform for the identification of various protein complexes.

4.3.4 Characterization of cofactor interaction in protein complexes from *E. coli* proteome

We noted that NfuA-[4Fe-4S] complex and 50S ribosomal protein L31-zinc complex have been reported with bound cofactors [54,55]. However, we only identified proteoforms corresponding to those two proteins without the cofactors through the database search and did not identify the whole protein complexes. Cysteine(C)149 and C152 of NfuA are known to play central roles in binding the [4Fe-4S] cluster, and C16 of 50S ribosomal protein L31 is crucial for zinc ion binding. We detected mass shifts as -4 and -2 Da from the identified 50S ribosomal protein L31 and NfuA proteoforms, **Figure 4.5C** and **Figure 4.6**.

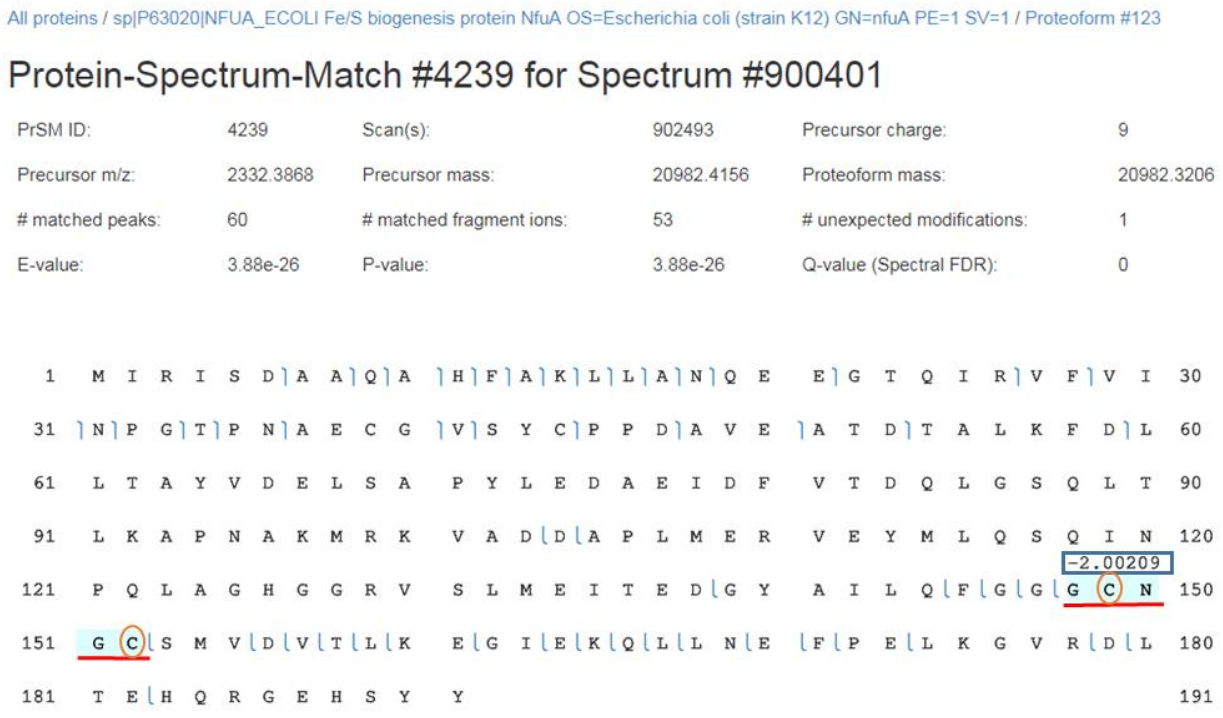

Figure 4.6. The sequence of the Fe/S biogenesis protein NfuA, the observed fragmentation pattern, and the mass shift detected through the database search. The mass shift, location of the mass shift, and the cysteine amino acids were highlighted.

Based on those mass shifts and their location in the protein sequences, we concluded that those mass shifts represented two disulfide bonds among the four cysteines (C16, C18, C37, and C40) in 50S ribosomal protein L31 and one disulfide bond (C149-C152) in NfuA. Therefore, 50S ribosomal protein L31 and NfuA were detected without the zinc ion and [4Fe-4S]. We further performed mass deconvolution on the averaged mass spectra across the peaks of the identified 50S ribosomal protein L31 and NfuA proteoforms without the cofactors, **Figures 4.7** and **4.8**. It is clear that 50S ribosomal protein L31 with two disulfide bonds and NfuA with one disulfide bond dominate the spectra. There are not very strong protein peaks corresponding to the 50S ribosomal protein L31-zinc complex and NfuA-[4Fe-4S] complex in **Figures 4.7** and **4.8**. The

results demonstrate that a larger fraction of 50S ribosomal protein L31 and NfuA exist as apo forms lacking the cofactors in the *E. coli* cells used in the experiment, providing new insight into the cofactor binding of these two protein complexes.

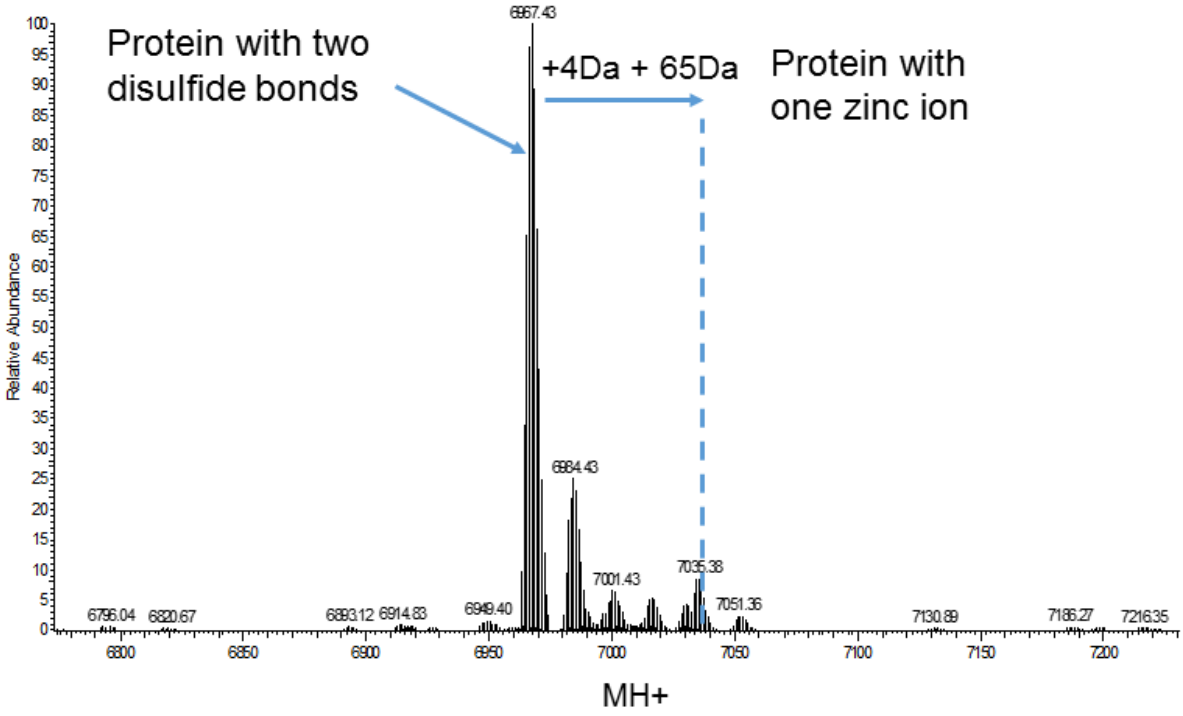

Figure 4.7. The deconvoluted spectrum from the averaged mass spectrum across the peak of the identified 50S ribosomal protein L31 proteoform without the zinc cofactor. The Xtract software from Thermo Fisher Scientific was used for the mass deconvolution with the default settings.

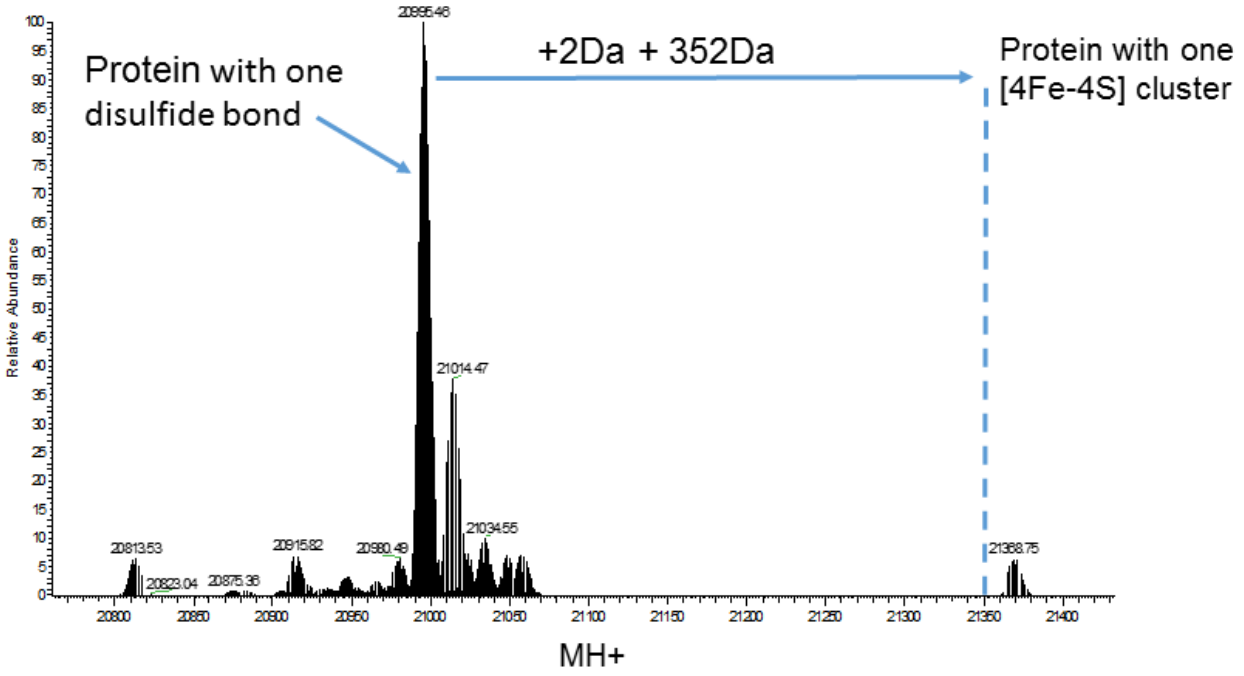

Figure 4.8. The deconvoluted spectrum from the averaged mass spectrum across the peak of the identified Fe/S biogenesis protein NfuA proteoform without the [4Fe-4S] cofactor. The Xtract software from Thermo Fisher Scientific was used for the mass deconvolution with the default settings.

The SEC-CZE-MS/MS platform identified 16 protein-metal complexes. These metalloproteins are involved in metal ion binding, catalysis, enzyme regulation, transcription, and transmembrane transport, **Figure 4.5D**. This work agrees with the literature regarding the molecular function distribution of metalloproteins [56]. The platform enabled us to determine the metal binding stoichiometry of most of those metalloproteins, **Table 4.3** and **Figure 4.5E**. Protein YcaR and DksA bind zinc ion through sulfur from cysteine (C) [44,45], and others most likely bind metal ions through nitrogen from histidine (H) and/or oxygen from acidic amino acids (aspartic acid (D) and glutamic acid (E)). Zinc ion binding through sulfur and nitrogen is generally more stable than that through nitrogen and oxygen [57]. Our metal binding

stoichiometry results agree well with the general concept from the literature. For YcaR and DksA, the abundance of the metal-binding form is at least 8 times higher than the non-binding form. For other proteins, the metal-binding form has lower abundance than the non-binding form. Our results highlight the potential of the native SEC-CZE-MS/MS platform for high throughput characterization of metal ion binding on metalloproteins directly from complex proteomes.

4.3.5 Characterization of PTMs on proteoforms under native condition

We identified many post-translational modifications (PTMs) including N-terminal acetylation, phosphorylation, C-terminal thiocarboxylation, 4′-phosphopantetheine, biotinylation, and disulfide bond. An example of those PTMs is shown in **Figure 4.9**. Some of the PTMs and corresponding proteins are listed in **Table 4.4**. We identified 55 proteoforms with N-terminal acetylation. Disulfide bonds were identified on eight proteins, and four of them are reported for the first time.

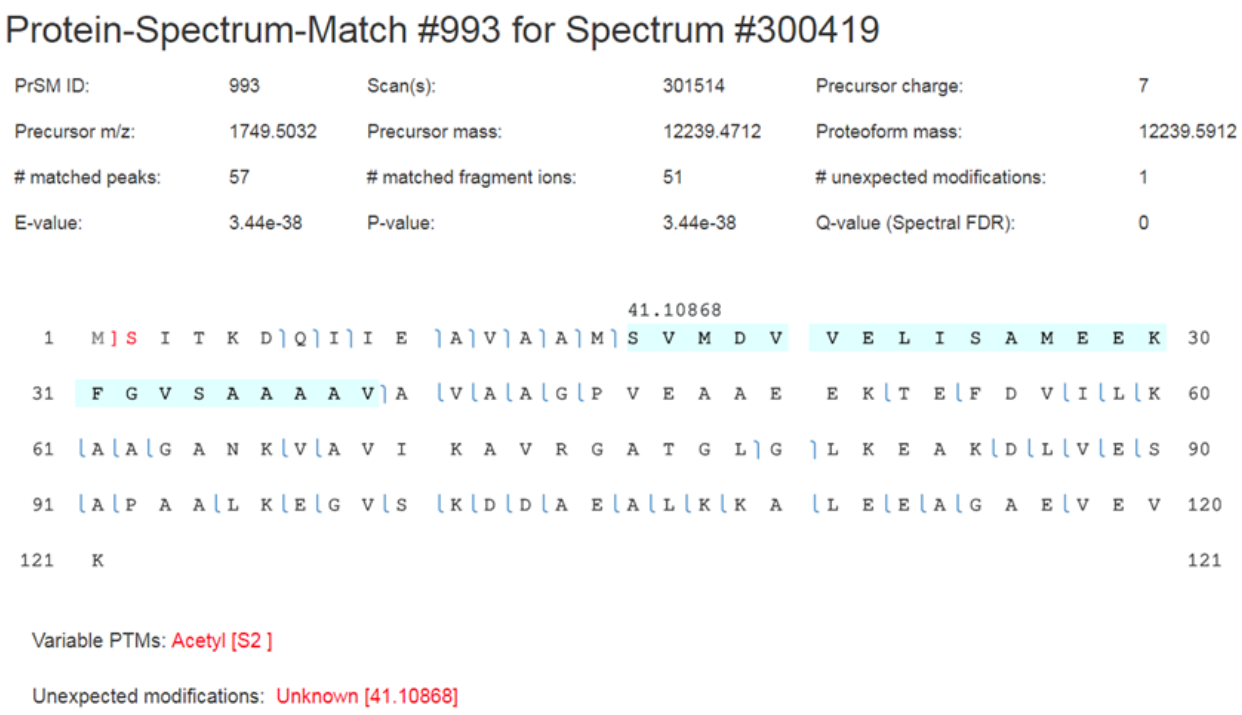

Figure 4.9. The sequence of the 50S ribosomal protein L7/L12, the observed fragmentation pattern, and the modifications through the database search. The initial methionine excision, N-terminal acetylation, and one +41 Da modification were labeled.

Table 4.4. The list of some of the PTMs detected in this work.

PTMs	Protein name	Mass error (Da)	E-value	Unreported*
C-terminal thiocarboxylation	Molybdopterin synthase sulfur carrier subunit	-0.09	2.90E-08	
Phosphorylation on histidine	PTS system glucose-specific EIIA component	-0.7	1.79E-29	
Biotinylation	Biotin carboxyl carrier protein of acetyl-CoA carboxylase	-0.2	2.28E-17	
4′-Phosphopantetheine	Acyl carrier protein	-0.92	6.42E-41	
Disulfide bond	Peroxiredoxin Bcp	-0.94	4.35E-22	
	Thiol:disulfide interchange protein DsbA	-2.4	1.84E-18	
	Thioredoxin 1	-0.01	2.90E-34	
	Glutaredoxin 3	0.00	2.40E-27	
	Fe/S biogenesis protein NfuA	0.00	3.88E-26	
	Putative sulfur carrier protein YeeD	-0.02	9.86E-14	
	Uncharacterized protein YbgS	-0.02	1.91E-11	
	50S ribosomal protein L31	-0.03	1.22E-18	

* The disulfide bonds in the proteins highlighted in green have not been reported in the literature.

We detected unreported signal peptide cleavage and initial methionine excision on 25 proteins, **Table 4.5**. An example of unreported signal peptide cleavage is shown in **Figure 4.10**.

Table 4.5. The list of proteins with unreported signal peptide cleavage and initial methionine excision.

Unreported signal peptide cleavage		
Protein name	First amino acid	Last amino acid
Maltose operon periplasmic protein	27	306
30S ribosome-binding factor	30	133
DNA-directed RNA polymerase subunit omega	26	91
Phosphocarrier protein HPr	11	85
Protein YcgL	12	108
Biotin carboxyl carrier protein of acetyl-CoA carboxylase	8	156
Uncharacterized protein YhhA	19	146
Cell division protein ZapB	4	81
Glycine betaine/proline betaine-binding periplasmic protein	47	109
Putative cryptic phosphonate transport system permease protein PhnE1	49	113
Inner membrane protein YihN	13	128
Nickel/cobalt homeostasis protein RcnB	27	112
50S ribosomal protein L25	19	94
PTS system glucose-specific EIIA component	8	169
PTS system glucose-specific EIIA component	9	169
Uncharacterized protein YkfA	11	144
DTW domain-containing protein YfiP	34	100
Unreported initial methionine excision		
Protein GrpE	2	197
UPF0234 protein YajQ	2	163
Glutaredoxin 4	2	115
Iron-sulfur cluster assembly scaffold protein IscU	2	128
Protein IscX	2	66
Putative sulfur carrier protein YeeD	2	75
Putative selenoprotein YdfZ	2	67
UPF0339 protein YegP	2	110

Figure 4.10. The sequence of the 50S ribosomal protein L25, the observed fragmentation pattern, and the modifications through the database search. The first 18 amino acids are cleaved as a signal peptide, which has not been reported in the literature. The signal peptide cleavage and one +2.3 Da modification were labeled.

4.3.6 Characterization of large protein complexes via native CZE-MS

Detection of large protein complexes is a big challenge in native proteomics. As mentioned above, all proteoforms and protein complexes we identified were less than 35 kDa (**Figure 4.2C**) due to the limitation of the Q-Exactive-HF mass spectrometer. However, most protein complexes are usually much larger than 30 kDa. In order to detect large protein complexes, we made two improvements in the native CZE-ESI-MS platform. First, we coupled our native CZE separation with an Agilent Q-TOF mass spectrometer to expand the *m/z* range of detection. Second, we developed a new coating for CZE separation to reduce the adsorption of proteins on the inner wall of the capillary. We noted that the traditional LPA coating still has interaction with intact proteins in native conditions, leading to the peak broadening and the lost identifications of low

abundance proteoforms. Thus, we designed a new linear carbohydrate polymer (LCP)-based neutral coating through the collaboration with Prof. Wenjun Du at Central Michigan University. A mixture of four protein complexes from 29 kDa to 232 kDa was used to evaluate the improved system. We first compared the CZE separation performance of the protein complex mixture with LPA and LCP coating, **Figure 4.11**. Similar sample amounts were used for the CZE separation. Obviously, the LCP coating showed better separation performance in terms of peak width and separation profile when comparing to traditional LPA coating. Four protein complexes and several protein impurities were baseline separated in the LCP-coated capillary. The peak intensity with LCP coating is also 5 times higher than that with the LPA coating. Moreover, we observed the peak of PK, a homotetramer about 232 kDa, with the new coating (peak 1 in **Figure 4.11**). A clear spectrum of PK was observed in the *m/z* range from 7000-11500 with a charge state distribution from 21+ to 33+ (**Figure 4.12**), which means the homotetramer of PK was in its native condition. Besides the full-length form, we also observed a truncated form, which is consistent with the literature report [58]. It is noted that the signal of PK was not observed in LPA-coated capillary, probably because of the strong adsorption of PK on the LPA coating, which significantly lowered the signal intensity of PK. The results indicated that the new LCP coating has less interaction and adsorption of proteins during the separation and can benefit the detection of native protein complexes, especially for larger protein complexes.

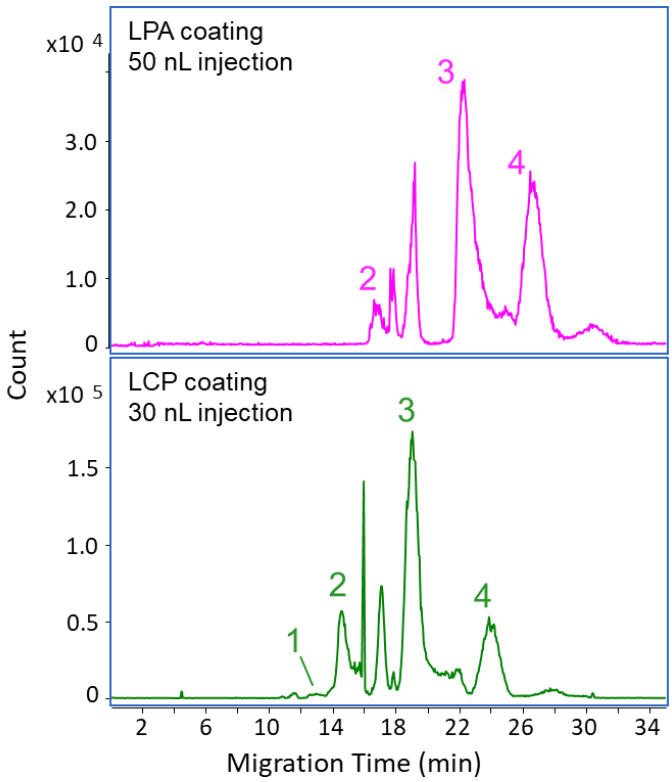

Figure 4.11. Electropherograms of CZE separation for four protein complexes with LPA-coated and LCP-coated capillaries. 1: tetrameric PK (232 kDa); 2: tetramer streptavidin (53 kDa); 3: CA-Zn^{2+} complex (29 kDa); 4: Dimeric SOD-Zn^{2+}, Cu^{2+} complex (31 kDa).

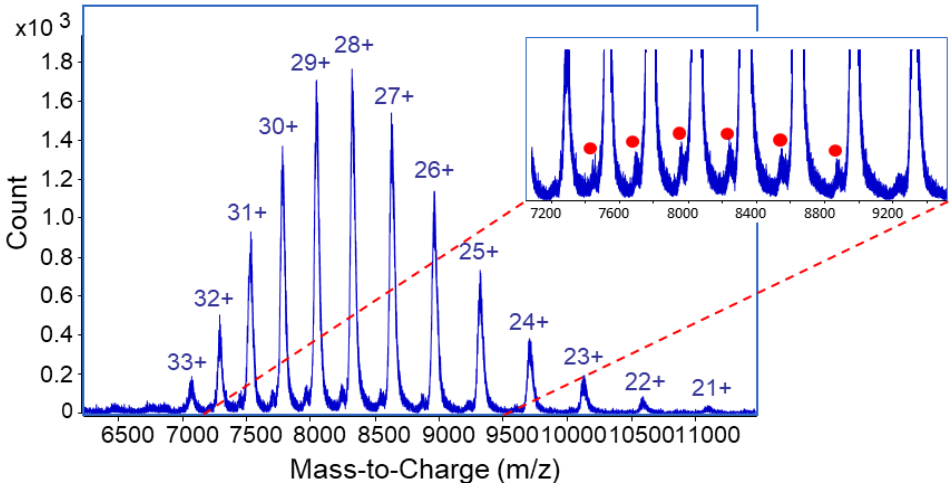

Figure 4.12. Averaged spectrum of the homotetramer of PK across the electropherographic peak in CZE separation with LCP coating. The inset is the zoom-in spectrum of *m/z* 7200-9400.

The streptavidin (peak 2 in **Figure 4.11**) was also well characterized in the LCP-based native CZE-ESI-MS analysis. Both monomer and homotetramer of streptavidin were observed in the averaged spectrum. The zoom-in spectrum of one charge state of the monomer shows a pair of peaks with the mass difference as 131 Da, which corresponds to the initiator methionine removal. The homotetramer can be assembled by either form of the monomer with four combinations. As a result, multiple peaks were observed in the zoom-in spectrum of the homotetramer. They represent the homotetramer with 0-3 methionine residues on the monomers. The peak of the tetramer with four methionine residues was not detected, probably because the abundance of this form was too low compared to other forms. The other two protein complexes detected in the CZE separation were CA-Zn^{2+} complex and dimeric SOD-Zn^{2+}, Cu^{2+} complex (peak 3 and 4 in **Figure 4.11**). The deconvolution mass and the theoretical mass of these two protein complexes were very close (CA: 29089.18 Da vs. 29088.00 Da; SOD: 31434.05 Da vs. 31432.58 Da). The mass errors were less than 1.5 Da and 47 ppm, indicating the high mass accuracy of the system.

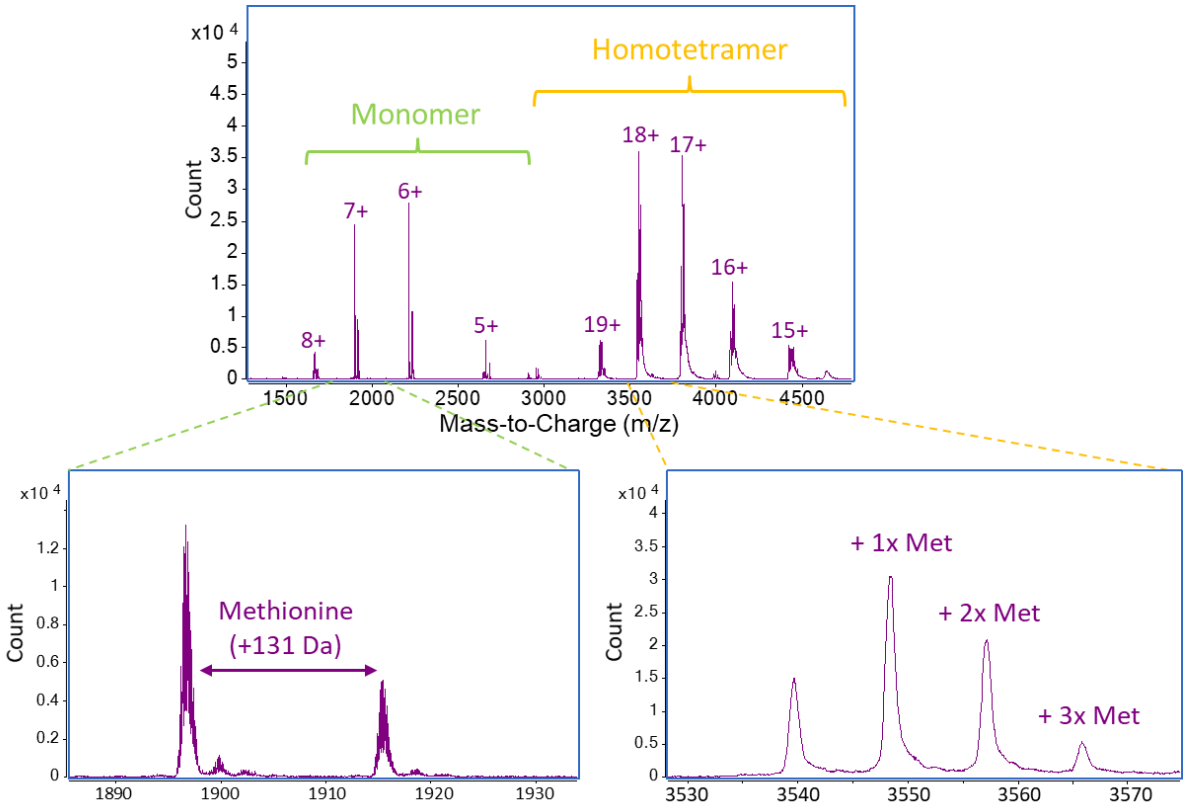

Figure 4.13. Averaged spectrum of streptavidin across the electropherographic peak in CZE separation with LCP coating. The insets are the zoom-in spectra of *m/z* 1890-1930 for the monomer and *m/z* 3530-3570 for the homotetramer.

We then employed the native CZE-ESI-MS platform with LCP coating to study the protein complex stability in gas phase. We performed 8 CZE runs with the same sample, and the CID potential were varied for each run from 0 to 200 V. The total intensity of all forms from the same protein complex was normalized to 100% so that the relative abundance of different forms could be investigated. The example we want to highlight in this study is PK and streptavidin, **Figure 4.13**. They are both homotetramers, but they went through different paths of dissociation as we changed the CID potential. The tetramer of pyruvate kinase tended to dissociate into a more stable form of dimer as the CID potential increased (**Figure 4.13A**). The dimer was the major

form after 30 V, while the trimer and monomer were much less abundant. Also, clearly the protein-protein interaction in pyruvate kinase is pretty strong, since the tetramer still existed at 200 V CID potential, though streptavidin is less stable compared with pyruvate kinase. We did not observe the tetramer at CID potentials beyond 50 V (**Figure 4.13B)**, which is most likely because streptavidin is much smaller than PK in mass (53 kDa vs. 232 kDa). As a result, it requires less internal energy to be dissociated. We also observed the monomer even when no CID potential was applied. One possible reason is that the complex could be denatured and dissociated in the ion source. It is also possible that the tetramer was dissociated during the producing process.

The dissociation path of streptavidin is also different from PK. The tetramer tended to dissociate into its monomer directly and the signal for trimer and dimer was very low. Over 90% of this protein was in the monomer form after 50 V CID potential. This study proved that native CZE-MS/MS can be a useful and high-throughput method to study protein complex stability.

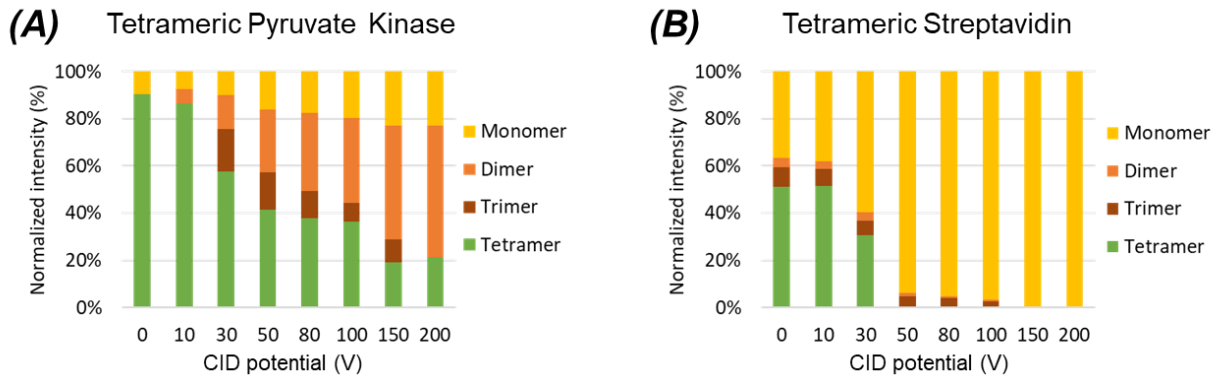

Figure 4.14. Relative abundance of different conformations of (A) PK and (B) streptavidin in gas phase as the CID potential increased.

4.4 Conclusion

We developed a novel, efficient and high-throughput SEC-CZE-MS/MS platform for the characterization of endogenous protein complexes in cells. The proof-of-principle study of the *E. coli* proteome identified 144 proteins, 672 proteoforms, and 23 protein complexes in discovery mode. The platform will be useful for the proteomics community for characterization of complex proteomes under native conditions. The platform can be further improved through optimization of the SEC and CZE conditions for better separation of protein complexes, via use of multiple fragmentation techniques (*e.g.*, HCD [59], ETD[51], ECD [16], and ultraviolet photodissociation [60]) for more comprehensive fragmentation of intact protein complexes, and by employing a high-resolution mass spectrometer that is capable of detection of large protein complexes in high *m/z* range. We have demonstrated the feasibility of native CZE-Q-TOF system for the detection of large protein complexes with a mixture of standard protein complexes. In the future, we will keep optimizing this system and employ the built-in ECD fragmentation for native proteomics to provide more comprehensive information on endogenous protein complexes of complex proteome samples.

4.5 Acknowledgment

We thank Profs. Daniel Jones and Jian Hu at the Department of Biochemistry and Molecular Biology at Michigan State University for kindly providing help on the project. We thank the support from the National Institute of General Medical Sciences, National Institutes of Health (NIH), through Grant R01GM118470 (X. Liu), R01GM125991 (L. Sun and X. Liu) and R01GM118685 (H. Hong).

REFERENCES

REFERENCES

[1] T. Geiger, A. Wehner, C. Schaab, J. Cox, M. Mann, Mol. Cell. Proteomics 11 (2012) M111. 014050.

[2] F. Zhou, Y. Lu, S. B. Ficarro, G. Adelmant, W. Jiang, C. J. Luckey, J. A. Marto, Nat. Commun. 4 (2013).

[3] T. Y. Low, S. van Heesch, H. van den Toorn, P. Giansanti, A. Cristobal, P. Toonen, S. Schafer, N. Hübner, B. van Breukelen, S. Mohammed, E. Cuppen, A. J. R. Heck, V. Guryev, Cell Rep. 5 (2013) 1469–1478.

[4] F. Fang, Q. Zhao, H. Chu, M. Liu, B. Zhao, Z. Liang, L. Zhang, G. Li, L. Wang, J. Qin, Y. Zhang, Mol. Cell. Proteomics 19 (2020) 1724–1737.

[5] J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman, K. R. Durbin, J. D. Tipton, A. Vellaichamy, J. F. Kellie, M. Li, C. Wu, S. M. M. Sweet, B. P. Early, N. Siuti, R. D. LeDuc, P. D. Compton, P. M. Thomas, N. L. Kelleher, Nature 480 (2011) 254–258.

[6] W. Cai, T. Tucholski, B. Chen, A. J. Alpert, S. McIlwain, T. Kohmoto, S. Jin, Y. Ge, Anal. Chem. 89 (2017) 5467–5475.

[7] J. Park, P. D. Piehowski, C. Wilkins, M. Zhou, J. Mendoza, G. M. Fujimoto, B. C. Gibbons, J. B. Shaw, Y. Shen, A. K. Shukla, R. J. Moore, T. Liu, V. A. Petyuk, N. Tolić, L. Paša-Tolić, R. D. Smith, S. H. Payne, S. Kim, Nat. Methods 14 (2017) 909–914.

[8] O. S. Skinner, N. A. Haverland, L. Fornelli, R. D. Melani, L. H. F. Do Vale, H. S. Seckler, P. F. Doubleday, L. F. Schachner, K. Srzentić, N. L. Kelleher, P. D. Compton, Nat. Chem. Biol. 14 (2017) 36–41.

[9] J. L. P. Benesch, B. T. Ruotolo, D. A. Simmons, C. V. Robinson, Chem. Rev. 107 (2007) 3544–3567.

[10] Y. Lu, H. Zhang, D. M. Niedzwiedzki, J. Jiang, R. E. Blankenship, M. L. Gross, Anal. Chem. 88 (2016) 8827–8834.

[11] S. Rosati, R. J. Rose, N. J. Thompson, E. van Duijn, E. Damoc, E. Denisov, A. Makarov, A. J. R. Heck, Angew. Chem. Int. Ed. 51 (2012) 12992–12996.

[12] J. Snijder, R. J. Rose, D. Veesler, J. E. Johnson, A. J. R. Heck, Angew. Chem. Int. Ed. 52 (2013) 4020–4023.

[13] R. S. Quintyn, M. Zhou, J. Yan, V. H. Wysocki, Anal. Chem. 87 (2015) 11879–11886.

[14] A. C. Susa, Z. Xia, E. R. Williams, Angew. Chem. Int. Ed. 56 (2017) 7912–7915.

[15] M. T. Marty, K. K. Hoi, J. Gault, C. V. Robinson, Angew. Chem. Int. Ed. 55 (2015) 550–554.

[16] H. Li, H. H. Nguyen, R. R. Ogorzalek Loo, I. D. G. Campuzano, J. A. Loo, Nat. Chem. 10 (2018) 139–148.

[17] B. Chen, Y. Peng, S. G. Valeja, L. Xiu, A. J. Alpert, Y. Ge, Anal. Chem. 88 (2016) 1885–1891.

[18] K. Muneeruddin, M. Nazzaro, I. A. Kaltashov, Anal. Chem. 87 (2015) 10138–10145.

[19] K. Muneeruddin, J. J. Thomas, P. A. Salinas, I. A. Kaltashov, Anal. Chem. 86 (2014) 10692–10699.

[20] A. Nguyen, M. Moini, Anal. Chem. 80 (2008) 7169–7173.

[21] A. -L. Marie, E. Dominguez-Vega, F. Saller, J. -L. Plantier, R. Urbain, D. Borgel, N. T. Tran, G. W. Somsen, M. Taverna, Anal. Chim. Acta. 947 (2016) 58–65.

[22] N. Said, R. Gahoual, L. Kuhn, A. Beck, Y. -N. François, E. Leize-Wagner, Anal. Chim. Acta. 918 (2016) 50–59.

[23] A. M. Belov, R. Viner, M. R. Santos, D. M. Horn, M. Bern, B. L. Karger, A. R. Ivanov, J. Am. Soc. Mass Spectrom. 28 (2017) 2614-2634.

[24] R. D. Melani, H. S. Seckler, O. S. Skinner, L. H. F. Do Vale, A. D. Catherman, P. C. Havugimana, M. Valle de Sousa, G. B. Domont, N. L. Kelleher, P. D. Compton, Vis. Exp. 108 (2016) e53597.

[25] J. Jorgenson, K. Lukacs, Science 222 (1983) 266–272.

[26] R. Haselberg, G. J. de Jong, G. W. Somsen, Anal. Chem. 85 (2013) 2289–2296.

[27] X. Han, Y. Wang, A. Aslanian, B. Fonslow, B. Graczyk, T. N. Davis, J. R. Yates, J. Proteome Res. 13 (2014) 6078–6086.

[28] R. D. Smith, C. J. Barinaga, H. R. Udseth, Anal. Chem. 60 (1988) 1948–1952.

[29] M. Moini, Anal. Chem. 79 (2007) 4241–4246.

[30] R. Wojcik, O. O. Dada, M. Sadilek, N. J. Dovichi, Rapid Commun. Mass Spectrom. 24 (2010) 2554–2560.

[31] L. Sun, G. Zhu, Y. Zhao, X. Yan, S. Mou, N. J. Dovichi, Angew. Chem. Int. Ed. 52 (2013) 13661–13664.

[32] N. Y. Engel, V. U. Weiss, M. Marchetti-Deschmann, G. Allmaier, J. Am. Soc. Mass Spectrom. 28 (2016) 77–86.

[33] M. Borges-Alvarez, F. Benavente, J. Barbosa, V. Sanz-Nebot, Rapid Commun. Mass Spectrom. 24 (2010) 1411–1418.

[34] M. Moini, Rapid Commun. Mass Spectrom. 24 (2010) 2730–2734.

[35] R. Gahoual, J. -M. Busnel, P. Wolff, Y. N. François, E. Leize-Wagner, Anal. Bioanal. Chem. 406 (2013) 1029–1038.

[36] Y. -N. François, M. Biacchi, N. Said, C. Renard, A. Beck, R. Gahoual, E. Leize-Wagner, Anal. Chim. Acta. 908 (2016) 168–176.

[37] G. Zhu, L. Sun, N. J. Dovichi, Talanta 146 (2016) 839–843.

[38] D. Chen, X. Shen, L. Sun, Analyst 142 (2017) 2118–2127.

[39] D. Kessner, M. Chambers, R. Burke, D. Agus, P. Mallick, Bioinformatics 24 (2008) 2534–2536.

[40] X. Liu, Y. Inbar, P. C. Dorrestein, C. Wynne, N. Edwards, P. Souda, J. P. Whitelegge, V. Bafna, P. A. Pevzner, Mol. Cell. Proteomics 9 (2010) 2772–2782.

[41] Q. Kou, L. Xun, X. Liu, Bioinformatics 32 (2016) btw398.

[42] J. E. Elias, S. P. Gygi, Nat. Methods 4 (2007) 207–214.

[43] E. N. McCool, R. A. Lubeckyj, X. Shen, D. Chen, Q. Kou, X. Liu, L. Sun, Anal. Chem. 90 (2018) 5529–5533.

[44] G. Bourgeois, J. Létoquart, N. van Tran, M. Graille, Biomolecules 7 (2017) 7.

[45] A. Perederina, V. Svetlov, M. N. Vassylyeva, T. H. Tahirov, S. Yokoyama, I. Artsimovitch, D. G. Vassylyev, Cell 118 (2004) 297–309.

[46] K. Inaba, S. Murakami, M. Suzuki, A. Nakagawa, E. Yamashita, K. Okada, K. Ito, Cell 127 (2006) 789–801.

[47] Rudresh, R. Jain, V. Dani, A. Mitra, S. Srivastava, S. P. Sarma, R. Varadarajan, S. Ramakumar, S. Protein Eng. 15 (2002) 627–633.

[48] C. Blériot, M. Gault, E. Gueguen, P. Arnoux, D. Pignol, M. -A. Mandrand-Berthelot, A. Rodrigue, Metallomics 6 (2014) 1400–1409.

[49] C. Rensing, G. Grass, FEMS Microbiol. Rev. 27 (2003) 197–213.

[50] H. E. KNOELL, J. KNAPPE, J. Eur. J. Biochem. 50 (1974) 245–252.

[51] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, D. F. Hunt, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 9528–9533.

[52] K. S. Gajiwala, S. K. Burley, J. Mol. Biol. 295 (2000) 605–612.

[53] Y. J. Sun, J. Rose, B. C. Wang, C. D. Hsiao, J. Mol. Biol. 278 (1998) 219–229.

[54] M. P. Hensley, T. S. Gunasekera, J. A. Easton, T. K. Sigdel, S. A. Sugarbaker, L. Klingbeil, R. M. Breece, D. L. Tierney, M. W. Crowder, J. Inorg. Biochem. 111 (2012) 164–172.

[55] S. Angelini, C. Gerez, S. O. Choudens, Y. Sanakis, M. Fontecave, F. Barras, B. Py, J. Biol. Chem. 283 (2008) 14084–14091.

[56] C. Andreini, I. Bertini, A. Rosato, A. Acc. Chem. Res. 42 (2009) 1471–1479.

[57] R. G. Pearson, J. Am. Chem. Soc. 85 (1963) 3533–3539.

[58] L. F. Schachner, A. N. Ives, J. P. McGee, R. D. Melani, J. O. Kafader, P. D. Compton, S. M. Patrie, N. L. Kelleher, J. Am. Soc. Mass Spectrom. 30 (2019) 1190–1198.

[59] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, M. Mann, Nat. Methods 4 (2007) 709–712.

[60] J. B. Shaw, W. Li, D. D. Holden, Y. Zhang, J. Griep-Raming, R. T. Fellers, B. P. Early, P. M. Thomas, N. L. Kelleher, J. S. Brodbelt, J. Am. Chem. Soc. 135 (2013) 12646–12651.

CHAPTER 5. Investigating Native Capillary Zone Electrophoresis-Mass Spectrometry on a High-End Quadrupole-Time-Of-Flight Mass Spectrometer for the Characterization of Monoclonal Antibodies

5.1 Introduction

Monoclonal antibodies (mAbs) have become a dominant class of therapeutics for the treatment of cancer and autoimmune diseases because of their specificity and affinity to diverse targets [1,2]. Since first commercialized in 1985, over 80 therapeutic mAbs have been approved by FDA [3,4]. However, the complex production process of mAbs usually introduces various post-translational modifications (*e.g.*, glycosylation, oxidation, deamidation, *etc.*) and structure changes (*e.g.*, misfolding, denaturation, aggregation, *etc.*), leading to heterogeneities in the final products, which affect the potency, stability and efficacy of the therapeutics [5-7]. Thus, critical quality attributes of mAbs need to be closely monitored to ensure the desired product quality.

Capillary zone electrophoresis (CZE) has been widely used for the quality control of mAbs in biopharmaceutical fields due to its high separation efficiency and straightforward operation [8]. Compared to other electrophoresis techniques, one advantage of CZE is that it has better compatibility with electrospray ionization-mass spectrometry (ESI-MS). A large number of reports illustrate CZE-ESI-MS for the characterization of mAbs from peptide mapping to intact protein analysis [9-19]. Although the peptide-level analysis can reveal detailed information on primary structure and PTMs of mAbs, intact mAb analysis better defines accurate mass and heterogeneity of mAb proteoforms. Han *et al.* performed an intact mass analysis of reduced and

deglycosylated IgG1 by CZE-ESI-MS implemented by an electrokinetically pumped sheath-flow nanospray interface [16]. Redman *et al.* developed a microfluidic CZE-ESI device with online MS detection for separation and characterization of charge variants of intact Infliximab [14]. We note that most of the intact mAb analyses by CZE-MS are performed under denaturing conditions, which most likely lead to the information loss of mAb's structure changes.

Recently, native CZE-ESI-MS has emerged as a promising technology for the characterization of mAb variants and aggregates as well as complex proteomes under native conditions [20,21]. The Ivanov group published the pioneering works on the analysis of mAbs by native CZE-MS, revealing major proteoforms due to glycosylations as well as low-abundance truncated species and mAb aggregates [18,22]. A sheathless CZE-MS interface [23], a linear polyacrylamide (LPA)-coated capillary, and an Orbitrap EMR mass spectrometer [24] were employed in the study. Le-Minh *et al.* investigated the conformational changes of Infliximab under stressed conditions using native CZE-MS [25]. A co-axial sheath liquid interface [26], a separation capillary with cationic coatings, and a Q-TOF mass spectrometer were utilized. Some challenges still exist in native CZE-MS for the characterization of mAbs. First, the CZE separation of different mAb variants or conformations under native conditions needs to be improved. Second, the sample loading capacity of CZE is low under native conditions, which impedes the detection of low-abundance proteoforms of mAbs. Our group has demonstrated the capability of the dynamic pH junction sample stacking method [27] for substantially boosting the sample loading capacity of CZE for large-scale top-down proteomics under denaturing conditions [28]. However, the dynamic pH junction method is hard to deploy for native CZE because it employs the drastic difference in pH between sample buffer and background electrolyte (BGE) (*i.e.,* pH 8-11 *vs.* 3). Alternative sample stacking methods are required for native CZE-MS.

In this work, we present the successful coupling of CZE with a high-end Q-TOF mass spectrometer using the electrokinetically pumped sheath-flow CE-MS interface [30,31] for the characterization of mAbs under native conditions. We first optimized the Q-TOF instrument parameters and CZE separation conditions, together with employing a new capillary coating to fulfill the maximum signal and mass resolution for native mAb detection. An online sample stacking method based on capillary isoelectric focusing (cIEF) in a narrow pH range was developed to expand the loading capacity as well as improve the CZE separation in the native condition for the first time. With the cIEF-assisted native CZE-MS, we achieved the separation of four conformational variants of SigmaMAb. In the study of NISTmAb, three major peaks were baseline separated. The glyco-proteoforms of both monomer and homodimer of the mAbs were observed in the mass spectra and annotated after deconvolution.

5.2 Experimental

5.2.1 Materials and reagents

The SILu Lite SigmaMAb universal antibody standard human (MSQC4), Pharmalyte 3-10, ammonium persulfate, ammonium acetate, ammonium formate and ammonium bicarbonate were purchased from Millipore Sigma Inc. (St Louis, MO, USA). Micro Bio-Spin™ P-6 gel columns were purchased from Bio-Rad Laboratories (Hercules, CA, USA). Hydrofluoric acid (HF) and acrylamide were purchased from Acros Organics (NJ, USA). The fused silica capillary (50 μm i.d., 360 μm o.d.) was purchased from Polymicro Technologies (Phoenix, AZ).

5.2.2 Sugar monomer synthesis and characterization

All reactions were conducted under dried nitrogen or argon stream. Anhydrous solvents (CH_2Cl_2

99.8%) were purchased in capped DriSolv™ bottles, used without further purification, and stored under argon. All other solvents and reagents were used without further purification. All glassware utilized were flame-dried before use. Glass-backed TLC plates (Silica Gel 60 with a 254 nm fluorescent indicator) were used without further manipulation and stored over desiccant. Developed TLC plates were visualized under a short-wave UV lamp, and/or by heating plates that were dipped in ammonium molybdate/cerium (IV) sulfate solution. Silica gel column chromatography was performed using flash silica gel (32-63 μm) and employed a solvent with polarity correlated with the TLC mobility.

Mass spectrometry was measured with a Waters LCT Premier™ XE unit. ^{1}H NMR spectra were recorded at 300 MHz on a Varian Mercury 300 MHz or a Varian Inova 500 spectrometer, respectively, with tetramethylsilane (TMS) proton signal as the standard. ^{13}C NMR spectra were recorded at 75 MHz on a Varian Mercury 300 spectrometer.

The synthesis of monomer **4** started from commercially available 1,2:5,6-di-*O*-isopropylidene-α-D-glucofuranose **1** (**Figure 5.1**). To a flame-dried round bottom flask was added **1** (1.0 g, 3.84 mmol) in anhydrous DCM (30 mL), acryloyl chloride (0.69g, 7.68 mmol) and triethylamine (TEA, 1.52 g, 15.4 mmol). The reaction mixture was stirred under nitrogen gas at rtf o 18 h. After determining that the starting material was consumed by TLC (hexane/EtOAc = 8/2, v/v, R_f = 0.4), solvent was removed under reduced pressure and the reaction mixture was subjected to silica gel chromatography (hexane/EtOAc = 8/2, v/v, R_f = 0.4) to give the product as an off-white oily liquid **2** (1.1 g, 92%), which was used directly for the synthesis of **3**.

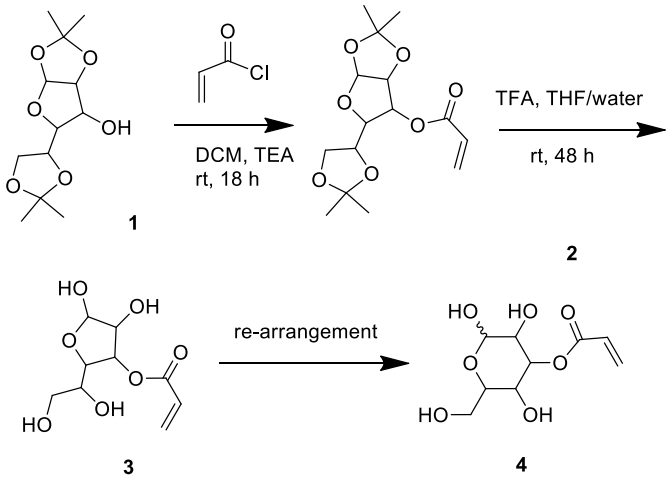

Figure 5.1. The synthesis of glucose-based monomer **4** from 1,2:5,6-di-O-isopropylidene-α-D-glucofuranose **1**.

To a round bottom flask was added compound **2** (1.0 g, 3.2 mmol) was added a mixed solvent of tetrahydrofuran (THF) and trifluoroacetic acid (TFA) (1:1. v/v). The reaction mixture was stirred for 48 h, after determining that the conversion was converted by TLC (further purification using silica gel chromatography (dichloromethane/methanol = 8/2, v/v, R_f = 0.3), the reaction mixture was neutralized by using 5% NaOH solution. The solvent was removed under reduced pressure to almost dry, silica gel (2.0 g) was added and the powder was subjected to silica gel chromatography (dichloromethane/methanol = 8/2, v/v, R_f = 0.3) to afford the final compound **4** as a slightly yellow powder, which underwent rearrangement to give 3-*O*-acryloyl-α/β-D-glucopyranose (0.6 g, 2.56 mmol, 80%). The structure was confirmed by ^{1}H NMR (500 MHz, D_2O) δ 6.31, 6.11, 5.96, 5.96, 5.89, 5.08, 4.89, 4.67, 4.65, 4.62, 4.58, 4.49, 4.46, 3.80, 3.58, 3.29, 3.22, 3.08, **Figure 5.2**. ^{13}C NMR (75 MHz, D_2O) δ 168.02, 162.43, 132.92, 127.56, 95.84, 92.03, 77.48, 75.40, 74.03, 72.66, 71.37, 69.49, 67.73, 60.72, **Figure 5.3**. ESI-MS calc for $C_9H_{14}NaO_7$ $[M+Na]^+$ = 257.0637, found: 257.0110.

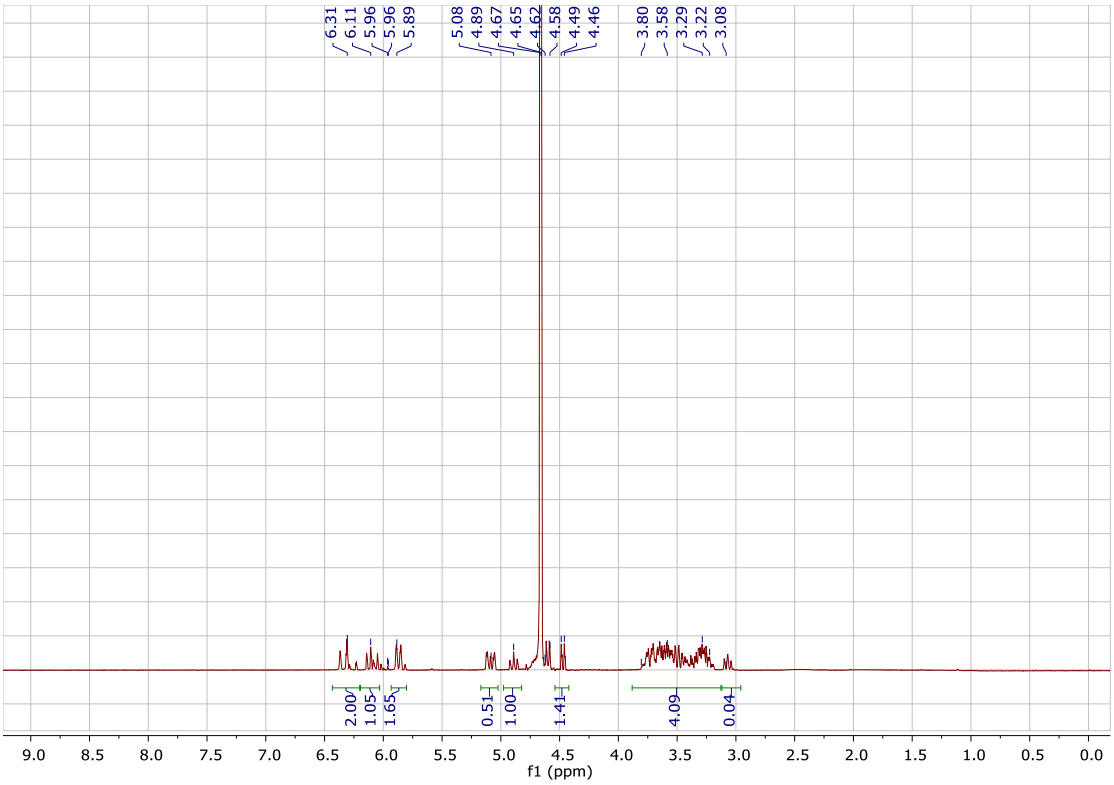

Figure 5.2. [1]H-NMR spectrum of 3-*O*-acryloyl-α/β-D-glucopyranose **4**.

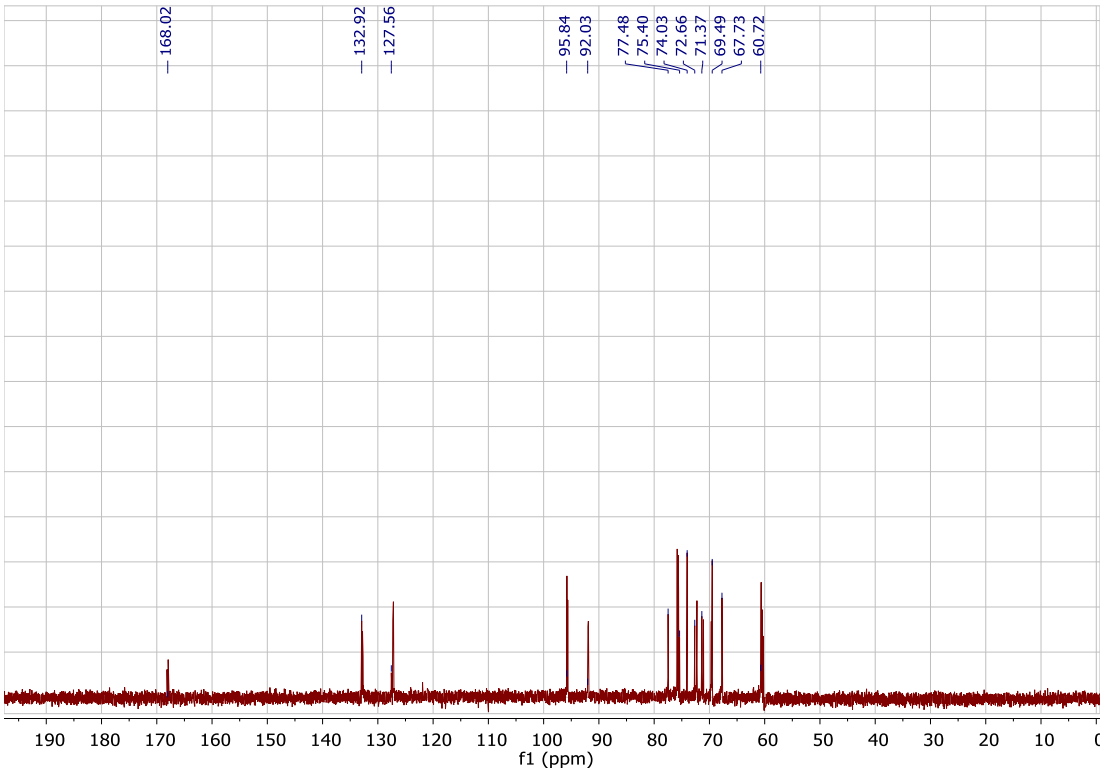

Figure 5.3. [13]C-NMR spectrum of 3-*O*-acryloyl-α/β-D-glucopyranose **4**.

5.2.3 Antibody purification

SigmaMAb lyophilized powder was dissolved in water. NISTmAb was received in solution. Antibody samples were purified and buffer exchanged with 10 mM ammonium acetate (pH 6.8) by Bio-Spin™ P-6 gel columns according to the instruction. Briefly, after the removal of the remaining buffer in the column, the column was washed with 500 µL 10 mM ammonium acetate four times. Then 20 µL of the stock sample was loaded in the column, and the protein sample was collected in the flow-through solution after centrifugation. The samples were diluted to desired concentrations for native CZE-MS analysis.

5.2.4 Native CZE-ESI-MS analysis

A 7100 CE System from Agilent Technologies (Santa Clara, CA) was used for the automated operation of CZE. The commercialized electrokinetically pumped sheath-flow nanospray interface (EMASS-II CE-MS Ion Source, CMP Scientific, Brooklyn, NY) [29,30] was used to couple CZE to a 6545XT AdvanceBio LC/Q-TOF mass spectrometer (Agilent Technologies, Santa Clara, CA). The ESI emitters of the CE-MS interface were pulled from borosilicate glass capillaries (1.0 mm o.d., 0.75 mm i.d., 10 cm length) with a Sutter P-1000 flaming/brown micropipet puller. The opening size of the ESI emitters was 30-40 µm. Voltage for ESI ranged from +2.2 to +2.5 kV.

Two 70-cm-long capillaries (50 µm i.d., 360 µm o.d.) coated with a new linear carbohydrate polymer (LCP) coating from a synthesized sugar monomer and a linear polyacrylamide (LPA) coating from a commercially available acrylamide monomer (**Figure 5.4**) were used for CZE separation. The newly designed sugar monomer (**Figure 5.4B**) was used to form the LCP coating. The LPA coating was prepared on the inner wall of the capillary based on the literature [31,32]. The new coating procedure is similar to the LPA coating. Briefly, the pretreated capillary was

filled with degassed sugar monomer (3-*O*-acryloyl-α/β-D-glucopyranose) solution (0.5 mg/mL) containing ammonium persulfate, followed by incubation at 35 ˚C water bath for 25-30 min with both ends sealed by silica rubber. After that, the capillary was flushed with water to remove the unreacted reagents. Both capillaries were etched with HF to reduce the outer diameter of one end of the capillaries to ~70 μm [33]. High voltage (+30 kV) and 50 mbar assisting pressure were applied for CZE separation unless specified. Samples were injected into the capillary by applying 100-950 mbar air pressure and the injection volume was calculated based on Poiseuille's law.

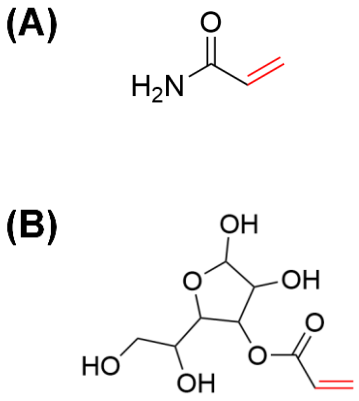

Figure 5.4. Monomer structures of the typical LPA coating (A) and the new carbohydrate polymer coating (B). The carbon double bonds highlighted with red are used for the reaction between monomers for polymerization.

The background electrolyte (BGE) was 25 mM ammonium acetate (pH 6.8), and the sheath liquid (SL) was 10 mM ammonium acetate (pH 6.8) for CZE separation unless specified otherwise. For cIEF, SigmaMAb was dissolved in 10 mM ammonium acetate (pH 6.8) with 0.25% Pharmalyte 3-10. Before sample injection, 160 nL (50 mM) ammonium acetate (AA, pH 9.0) was injected as the catholyte. After sample injection, 12 nL BGE was injected to make sure the sample would not move back into the stock BGE solution. At the beginning of the separation, the sample

was first focused without assisting pressure for 5-20 min depending on the injection volume. Then assisting pressure of 50 mbar was applied to the sample injection end of the separation capillary.

6545XT AdvanceBio LC/Q-TOF mass spectrometers (Agilent Technologies, Santa Clara, CA) with and without an electromagnetostatic ExD cell (e-MSion, Corvallis, OR) were used for the experiments. The instrument without the ExD cell was used for optimizations of MS parameters, CZE background electrolytes (BGEs) and sheath liquid (SL). The instrument with the ExD cell was used for other experiments. The ExD cell was set for positive transmission without ECD fragmentation (ECD off). CID potential of 10 V was required for the maximum transmission efficiency of mAbs in the system with the ExD cell. A regular ESI spray shield and a nanoESI spray shield were used in the experiments. The gas temperature and flow rate of nitrogen drying gas was 365 ^{o}C and 1 L/min. The voltage applied on the ion transfer capillary was 0 V. The mass range option was set as High (10000 m/z). The slicer mode was High Resolution. The mass range of detection was 3000-10000 m/z, and the scan rate was 0.25 spectrum/sec. Fragmentor voltage, skimmer voltage and CID potentials were set as specified.

5.2.5 Data analysis

Native CZE-ESI-MS runs were analyzed with Agilent MassHunter Qualitative Navigator B.08.00. Mass spectra were averaged across the electropherographic peaks. Average charge state (Z_{avg}) of the mAb was calculated based on the equation [34]:

$$Z_{\text{avg}} = \frac{\sum_i^n q_i I_i}{\sum_i^n I_i} \qquad \text{(Equation 5.1)}$$

where q is the net charge of the given charge state, I is the intensity of the given charge state.

Deconvolution was performed using Agilent MassHunter BioConfirm 10.0 using Maximum Entropy algorithm. The mass step was 0.05 Da. Other parameters for deconvolution were set as default.

5.3 Results and Discussion

5.3.1 Optimizations of mass spectrometric parameters

Different from the detection of denaturing proteins, native MS requires mild MS parameters to prevent protein complex dissociation, denaturation, and fragmentation in the system. To achieve high-quality spectra as well as maintain the native condition of native mAbs, we investigated three MS parameters (fragmentor voltage, skimmer voltage, CID potential) of the Q-TOF instrument through direct infusion MS of SigmaMAb (3 mg/mL in 10 mM NH_4Ac, pH 6.8) with CZE system.

Fragmentor voltage is designed to promote ion transmission and perform in-source fragmentation. For large molecules like mAbs, the high fragmentor voltage could improve transmission and sensitivity. Moreover, high salt concentration is usually used for native conditions and caused salt adduction on proteins. The high fragmentor voltage could help decluster salts and water molecules complexing with the proteins. Therefore, we chose the fragmentor voltage of 380 V for the later experiments, which was the largest value we could set for the instrument.

Skimmer voltage and CID potential also had significant impacts on the mass resolution and signal intensity of the mAb. Skimmer is used to sample the analytes into the high vacuum compartment. It can also focalize the ions and reduce ion beam broadening. When the skimmer voltage was raised from 65 to 300 V, we observed a three-fold improvement in mass resolution

and a ten-fold increase in mAb intensity, **Figure 5.5A**. The higher skimmer energy could add more internal energy to the analytes and further remove the salt and solvent adducts on the mAbs. As a result, the heterogeneity of the mAb was reduced and higher mass resolution and signal intensity were achieved.

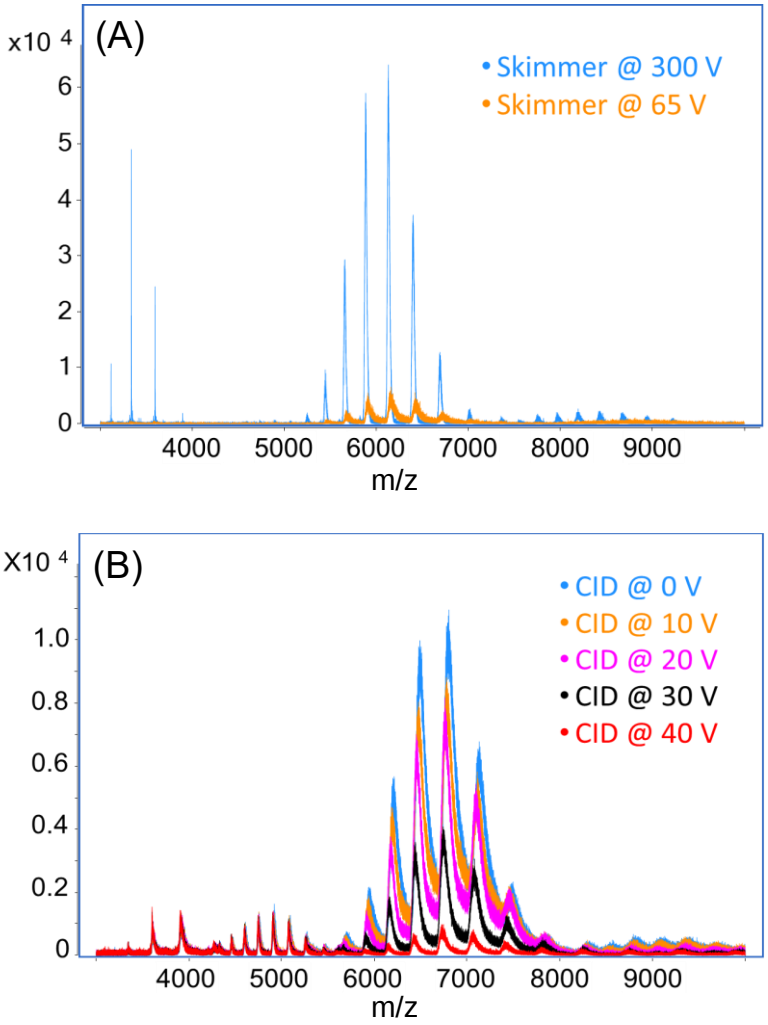

Figure 5.5. Mass spectra of the SigmaMAb through direct infusion MS with the CZE system and the nanospray sheathflow CE interface. (A) Skimmer and (B) CID potential was investigated.

Applying collision energy is also commonly used to help remove salt and neutral adducts from proteins during native MS experiments. We investigated five different CID potential (0 V to 40 V) in the CID cell. As we increased the CID potential, the intensity of both monomer and homodimer of SigmaMAb decreased, **Figure 5.5B**. One possible reason is that we already applied enough energy on the proteins for declustering, and the high CID potential could cause fragmentation of the proteins and lead to the reduction of protein intensity. Therefore, CID potential was not necessary for native MS of the SigmaMAb. We used 10 V for later experiments because we have an ExD cell in our instrument and we needed the addition of CID potential to promise the transmission efficiency.

Another setting that affects the acquired spectra is the spray shield on the inlet of the Q-TOF mass spectrometer. We tested two kinds of spray shield: a regular ESI spray shield and a nanoESI spray shield. With the regular ESI spray shield that has a larger orifice, more ions could be transmitted into the instrument and higher signal intensity was obtained. With the nanoESI spray shield that has a smaller orifice and allows fewer ions to pass into the instrument, the signal intensity dropped about 10 times. However, the mass resolution was greatly improved, and the peak broadening effect was reduced. As a result, we were able to see clear signals of mAb proteoforms due to glycosylations in the spectra with the nanoESI spray shield.

5.3.2 Optimizations of the CZE conditions for mAbs

Volatile salt solutions around neutral pH are usually used to preserve the higher order structure of proteins in native liquid-phase separations. We first investigated two kinds of salt, ammonium acetate (AA) and ammonium formate (AF), for BGE and SL in native CZE-ESI-MS. An LPA-coated capillary was used for CZE separation and SigmaMAb (3 mg/mL) was still used as a standard to test the system. For each CZE run, 60 nL mAb sample was injected. Here, we used the

regular ESI spray shield. Both BGE and SL were prepared with 10 mM concentration of the salt buffers. AA and AF presented almost identical separation profiles, **Figures 5.6A** and **5.6B**, but AA had slightly better separation performance in the labeled part where two minor peaks were separated. The mass spectra of the major peaks in both electropherograms have a charge state distribution from 21+ to 28+ in the range of 5000-7200 *m/z*, **Figure 5.6D**. Besides the monomer, we also observed the homodimer of SigmaMAb with a charge state distribution of 32+ to 39+, demonstrating two salt conditions were gentle enough to preserve the noncovalent interaction. However, when comparing these two spectra, AA presented higher intensity for both monomer and homodimer of the SigmaMAb. We further calculated the average charge state of the mAb in two conditions. The mAb in AA has a marginally lower average charge state than that in AF (24.4 *vs.* 24.5). The result indicates that mAb in AF is a little more denatured, which is consistent with previous reports that AF has a destabilizing effect and can cause structure unfolding of proteins [34,35].

We also tested 50 mM AA for BGE and SL and the electropherogram is shown in **Figure 5.6C**. The CZE separation in 50 mM AA has a similar separation profile, but longer migration time and wider peak width compared to that in 10 mM AA, probably due to the increased viscosity of BGE as the salt concentration increased. From the mass spectra, we observed the reduction of mAb monomer intensity as well as the mass resolution with increased concentration of AA, **Figure 5.6E**, which could be explained by the higher salt concentration interfered with the ionization and caused ion suppression. Nonetheless, it should be noted that the intensity of mAb dimer is higher in 50 mM AA, **Figure 5.6E**. Also, the average charge state of mAb in 50 mM AA is lower than that in 10 mM AA (23.8 *vs.* 24.4), because the higher salt concentration could better maintain protein higher order structure. After an overall consideration of signal intensity and native conformation,

we finally decided to use 10 mM AA as SL and 25 mM AA as BGE for the following native CZE-ESI-MS experiments.

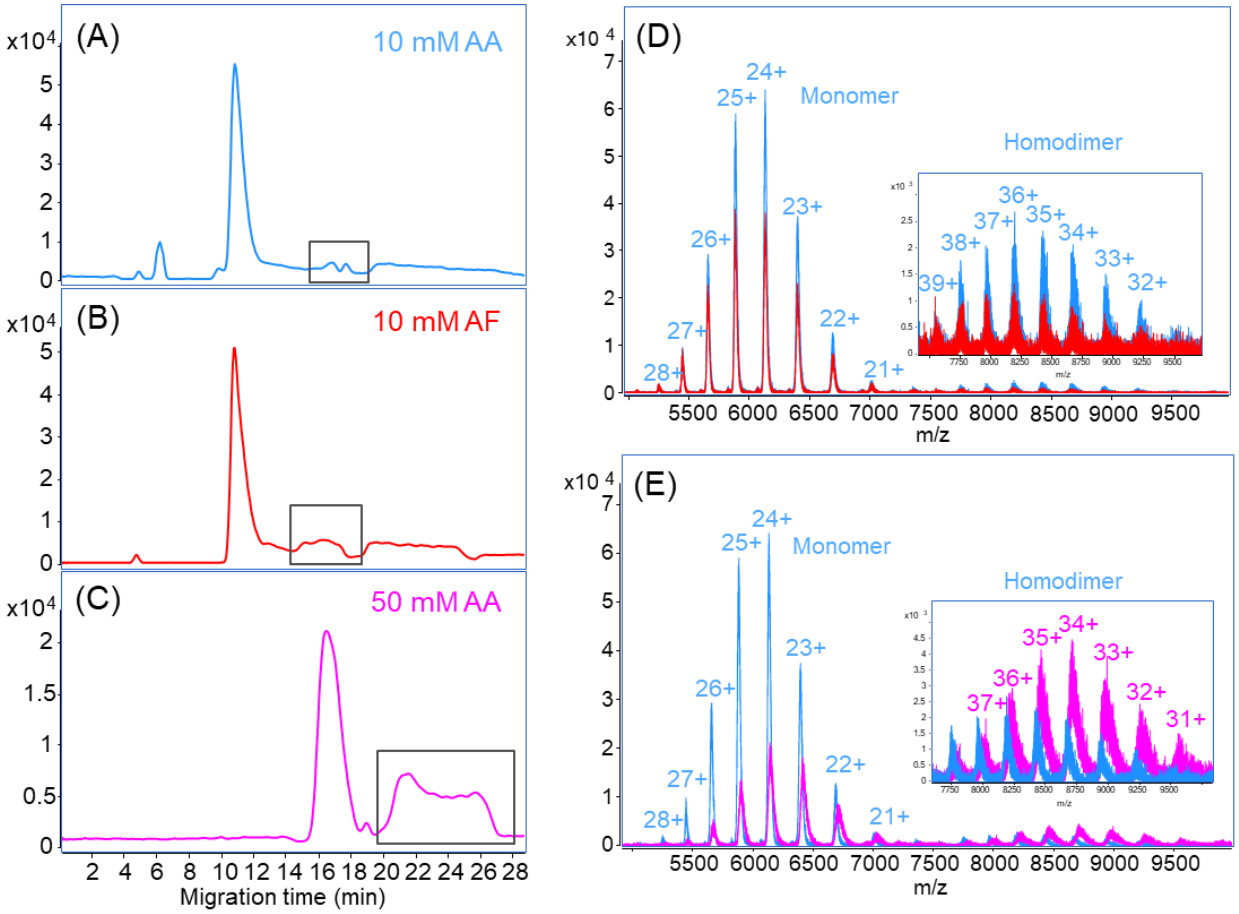

Figure 5.6. Investigation of native CZE separation conditions for the SigmaMAb. (A-C) Base peak electropherograms of native CZE-MS for SigmaMab with 10 mM AA, 10 mM AF and 50 mM AA as the BGE and SL. The peaks labeled with black boxes in the electropherograms represent the same mAb species. The spectra of the main peak in (D) 10 mM AA *vs.* 10 mM AF and (E) 10 mM AA *vs.* 50 mM AA are overlapped for comparisons. The insets are the zoom-in spectra of 7500-9500 m/z range. An LPA-coated capillary was used.

The assisting pressure used in CZE separation was also studied. In native condition, the mAb is folded and carries much fewer charges compared to that in denaturing conditions. Consequently,

the electrophoretic mobility of the mAb in native CZE is lower than regular CZE, and its migration time can be very long. We applied assisting pressure in native CZE separation of the mAb to shorten the migration time and increase the throughput of experiments. **Figure 5.7** shows the electropherograms of SigmaMAb with 0-50 mbar assisting pressure during the native CZE separation. The injection amount is 60 nL. One thing we need to additionally indicate is that from here we changed the regular ESI spray shield to a nanoESI spray shield. Thus, the peak intensity was decreased significantly compared to the experiments above. As the assisting pressure decreased, the migration time of the mAb turned longer. At the same time, the peak width became wider and peak shape became worse. When applying 50 mbar assisting pressure, two peaks were observed in the electropherogram. However, only one peak was observed with lower assisting pressure. When no assisting pressure was applied, we could not even find the mAb signal. The longer migration time gave the analytes more chance to diffuse during the separation and eventually led to the peak broadening. Considering both the throughput and separation performance, assisting pressure of 50 mbar was used for the following experiments.

The capillary coating is another key factor in CZE separation. The LPA coating has been widely used in both peptide and intact protein analysis to eliminate electroosmotic flow in the capillary and improve CZE separation performance. However, we noticed from the data above that the LPA-coated capillary still had protein adsorption on the inner wall as evidenced by wide peaks of the mAb. It has been demonstrated that the polymers having the nitrogen element lead to significant protein adsorption [36] and carbohydrates-based polymers have excellent resistance to protein non-specific adsorption [37,38]. Recently, we developed a new linear carbohydrate polymer (LCP)-based neutral coating, which is based on a glucose monomer (for details, see ESI) and applied this new coating for mAb studies. With the same MS settings and CZE conditions, the new

carbohydrate coating showed a 6-fold increment in mAb intensity (**Figure 5.8** *vs.* **Figure 5.7A**). The result suggests that the new LCP coating produces less interaction with the mAb during separation, boosting the sensitivity of native CZE-MS for the mAb significantly. We employed capillaries with the LCP coating for the rest of the experiments.

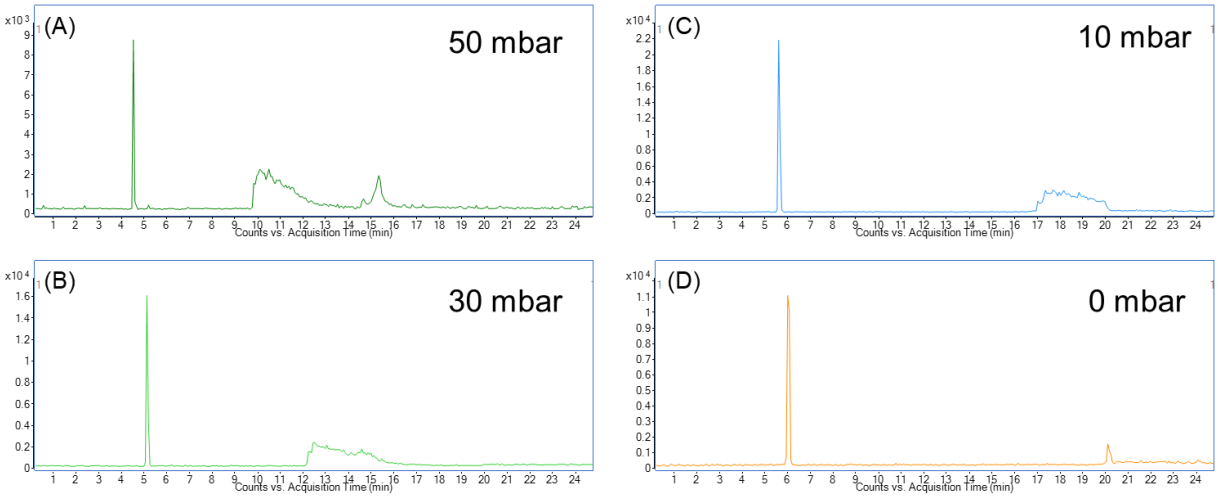

Figure 5.7. Base peak electropherograms of native CZE-MS for the SigmaMAb with (A) 50 mbar, (B) 30 mbar, (C) 10 mbar, (D) 0 mbar assisting pressure during separation. An LPA-coated capillary was used.

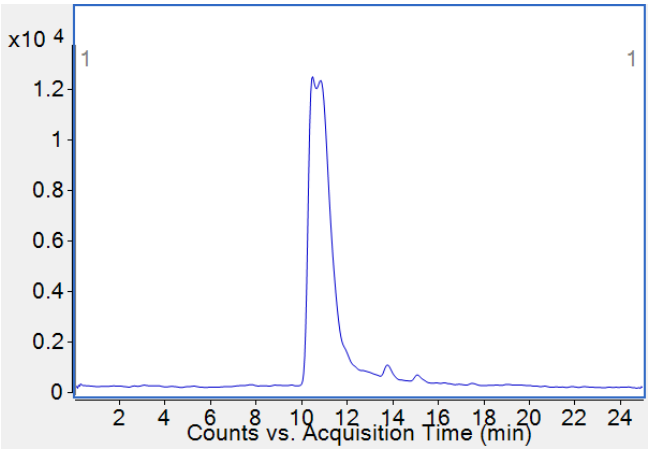

Figure 5.8. Base peak electropherogram of native CZE-MS for SigmaMAb with the new carbohydrate-coated capillary.

5.3.3 Evaluating native capillary isoelectric focusing (cIEF)-assisted CZE-MS for mAbs

One drawback of CZE is the low sample loading capacity. Less than 1% of the total capillary volume is typically filled with the sample to obtain high separation efficiency, which limits the detection of low-abundance species in the sample. Using online sample stacking methods could help solve this problem. In addition, it could reduce peak width and lead to higher separation resolution. Many sample stacking methods have been evaluated for denaturing CZE-MS characterization of proteins, *e.g.,* dynamic pH junction and field-amplified sample stacking (FASS) [29]. However, it is difficult to apply them in native conditions efficiently. An efficient sample stacking method under native conditions is urgently needed for native CZE-MS.

Here we investigated the possibility of employing cIEF for online sample concentration in native CZE-MS. cIEF separates analytes based on their isoelectric points (pIs). Although cIEF is mostly used for protein and peptide separation under denaturing conditions, several studies proved its feasibility for protein complex separation under native environments without destroying the native conformation and noncovalent interactions [39-43]. Its feature of focusing and ability of operation in native conditions provides us the possibility to utilize cIEF in a narrow pH range (*i.e.*, pH 6-9) as a sample stacking method in the native CZE separation. First, a short plug of 50 mM AA (pH 9.0, about 160 nL) was injected as the catholyte for cIEF. Then, we injected a 30-nL SigmaMAb sample (3 mg/mL) dissolving in 0.25% Pharmalyte and 10 mM AA into the capillary. After that, 12-nL BGE (25 mM ammonium acetate, pH 6.8) was injected. After a high voltage was applied across the capillary, the SigmaMAb was first focused by native cIEF in the sample plug. After the focusing was completed, the mAb was further separated by native CZE. The native cIEF stacking had an obvious contribution to the CZE separation, **Figure 5.9A**. Four major peaks of SigmaMAb were separated, which was not observed by regular native CZE separations in **Figure 5.6**.

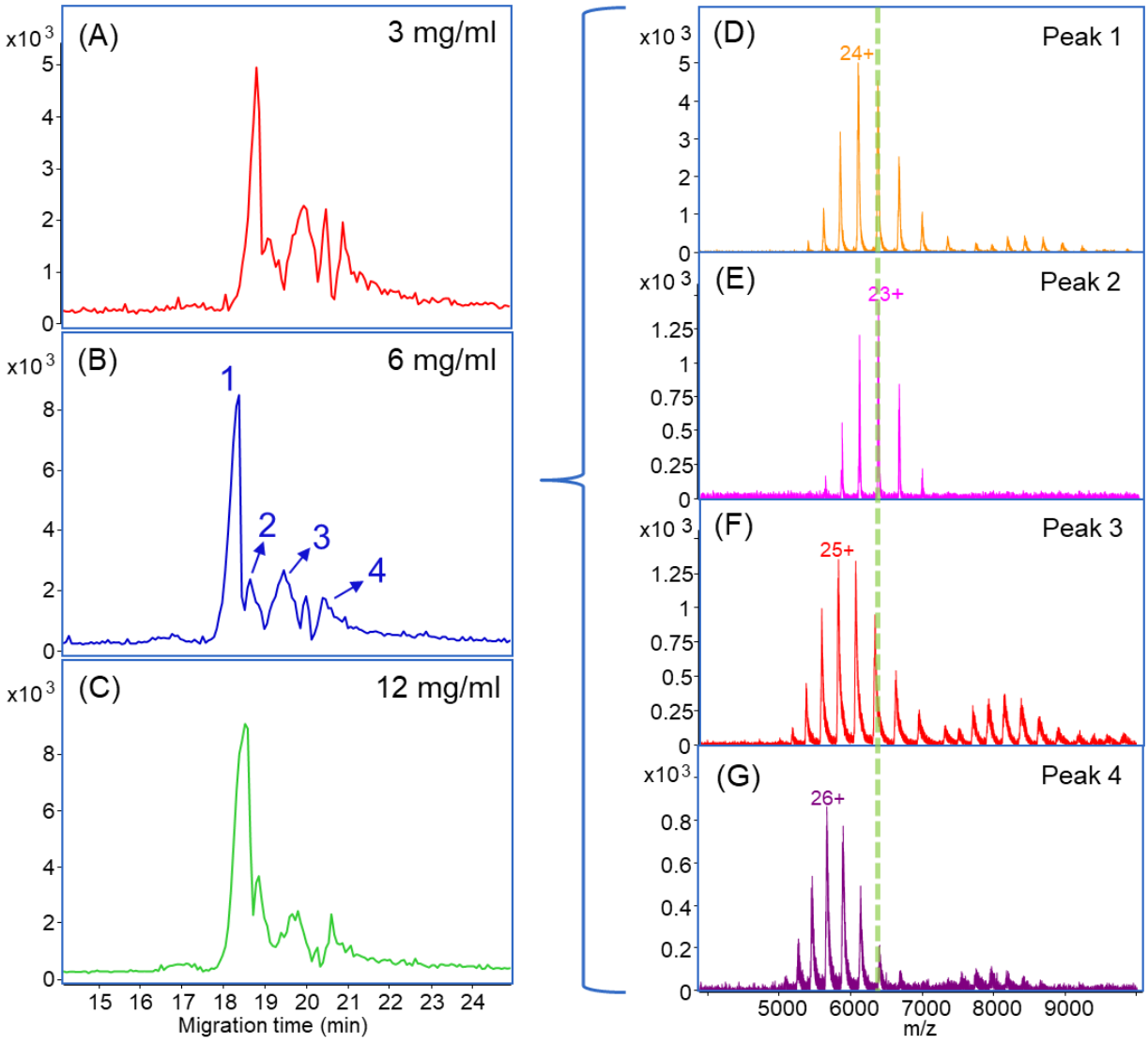

Figure 5.9. (A-C) Base peak electropherograms of native cIEF-assisted CZE-MS for 3, 6 and 12 mg/mL SigmaMAb; (D-G) Averaged mass spectra of the four major peaks separated in (B). Herein the LCP-coated capillary was used.

We increased the sample concentration for higher intensity as shown in **Figures 5.9B** and **5.9C**. The electropherograms of the three concentrations are reproducible in terms of migration time and separation profile. The highest intensity was achieved with 6-mg/mL SigmaMAb. The higher protein concentration (12 mg/mL) did not further improve the signal intensity but widened the

peak instead. We observed four major peaks of SigmaMAb with the 6 mg/mL sample (**Figure 5.9D-G**). The mass spectra show charge state distributions (CSDs) of mAb monomer from 20+ to 28+ and homodimer from 32+ to 39+. The zoom-in mass spectrum of the 23+ charge state from **Figure 5.9E** is shown in **Figure 5.10A**. The proteoforms due to different glycosylations could be resolved in the spectrum. The deconvoluted mass spectrum of **Figure 5.9E** shows five known glycosylated proteoforms of SigmaMAb (**Figure 5.10B**, **Table 5.1**), which were not achieved in the previous work with native CZE separation and Q-TOF instrument [25]. Interestingly, the four major peaks of SigmaMAb correspond to very similar masses after deconvolution but are different in CSDs, **Figures 5.9D-G**. Ions in peaks 1 and 2 carried significantly fewer charges than peaks 3 and 4. The charge envelopes of peaks 3 and 4 shifted to the higher charge states, which implied their structures were partially unfolded leading to more positive charges. Thus, our native cIEF-assisted CZE-MS most likely separated four different conformations of the SigmaMAb. These potential conformational changes could be caused by environmental stress such as long-time storage or exposure to light and room temperature. It could happen in both production and sample preparation.

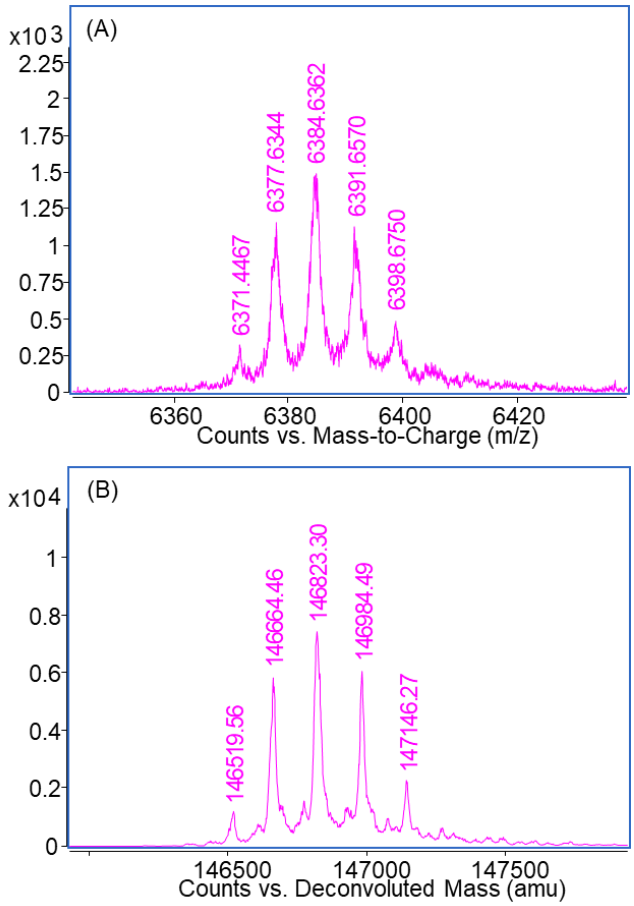

Figure 5.10. (A) A zoom-in mass spectrum of 23+ charge state and (B) deconvolution of SigmaMAb proteoforms observed in Figure 3E. The carbohydrate-coated capillary was used.

Table 5.1. Theoretical and observed masses of the major glyco-proteoforms of SigmaMAb monomer detected in peak 2 (Figure 3E) in native cIEF-assisted CZE-MS.

Glyco-proteoform	Theoretical Mass (Da)	Observed Mass (Da)	Mass Error (Da)	Mass Error (ppm)
G0+G0F	146512.2	146519.6	7.4	50.5
G0F+G0F	146658.4	146664.5	6.1	41.6
G0F+G1F	146820.6	146823.3	2.7	18.4
G1F+G1F	146982.7	146984.5	1.8	12.2
G1F+G2F	147144.8	147146.3	1.5	10.2

We further tested different sample injection volumes using the native cIEF-assisted CZE-MS for the mAb. With the help of cIEF sample stacking, we were able to inject a large sample volume without losing separation resolution significantly. The concentration of Pharmalyte was decreased from 0.25% to 0.1% to reduce the interference of ampholytes to the mass spectrometer. Four different sample injection volumes from 30 nL to 800 nL were evaluated in two aspects: peak intensity and peak width, **Figure 5.11**. The SigmaMAb sample we used was 3 mg/mL. When the injection volume increased from 30 nL to 200 nL, the peak intensity was boosted about 6.5-folds and the peak width was doubled. When we increased the injection volume from 200 nL to 800 nL, the peak intensity was increased slightly, but the peak width was increased by nearly 100%. Although we adopted cIEF to stack the sample in the capillary, the stacking ability of cIEF was limited in native conditions. When too much mAb sample was injected, it would cause peak broadening and even protein precipitation in the capillary, and finally could not provide the expected increment of intensity. Furthermore, excess injection volume also resulted in excess ampholytes in the capillary, which interfered with the ionization of the mAb, enlarged the background noise and decreased the S/N ratio. Therefore, we selected 200 nL as the injection volume for later experiments.

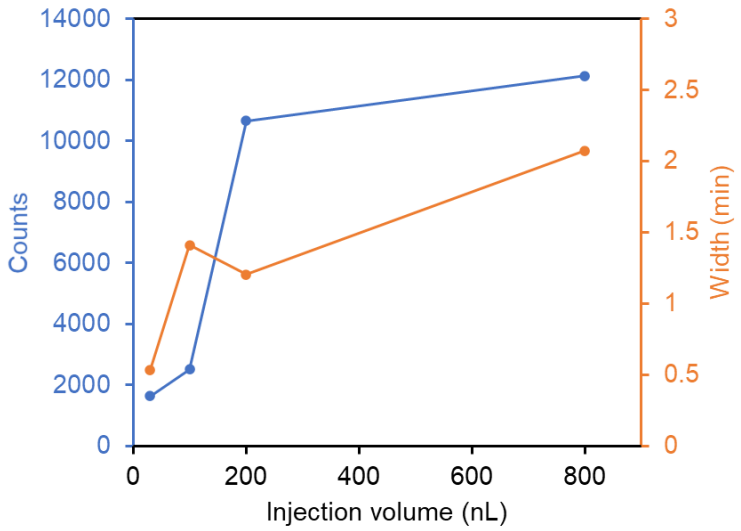

Figure 5.11. Peak intensity and peak width of the mAb as a function of sample injection volume in native cIEF-assisted CZE-MS for SigmaMAb. The most abundant peak in the electropherograms were selected.

5.3.4 Native cIEF-assisted CZE-MS for the NISTmAb

We further applied the native cIEF-assisted CZE-MS in the characterization of another recombinant humanized monoclonal IgG1 antibody, NISTmAb. The NISTmAb was dissolved in a buffer containing 0.1% Pharmalyte and 10 mM AA with a concentration of 1 mg/mL. Because NISTmAb has an isoelectric point as 9.18, we increased the catholyte pH to 9.5 to ensure the mAb could be focused by cIEF. For each run, 200 nL sample was injected into the capillary. **Figure 5.12A** shows the separation of the NISTmAb with three peaks. The main peak (peak 2) reveals the presence of both monomer and homodimer of the mAb, **Figure 5.12B**. Zoom-in mass spectra of the monomer at 24+ and dimer at 38+ are shown in **Figures 5.12C** and **5.12D**. The deconvolution of the monomer signal, **Figure 5.12E**, identified four major and four minor biantennary glyco-proteoforms of the NISTmAb. The minor species included G2F/G2F, the addition of one hexose and the loss of N-acetylglucosamine (GlcNAc). These glyco-proteoforms have been previously

reported for the NISTmAb [45]. The peaks of major glyco-proteoforms in the spectrum had fronting shapes, which were caused by the C-terminal lysine variants that could not be resolved due to the limited mass resolution. We also observed eight corresponding glyco-proteoforms of the homodimeric NISTmAb, **Figures 5.12D** and **5.12F**. The assignments of all proteoforms for monomer and homodimers are listed in **Table 5.2**. The mass spectra and deconvoluted spectra of peak 1 and peak 3 are shown in **Figure 5.13**. Both monomer and dimer signals were detected in the averaged spectrum across peak 1, **Figure 5.13A**. We could not get a clear deconvolution result for glyco-proteoforms from the spectrum due to the low intensity, **Figure 5.13C**. However, it is clear that the deconvolution mass of the mAb proteoform in peak 1 is roughly 1500 Da smaller than that in the major peak 2. The mass shift is close to the mass of one glycan, thus, peak 1 probably represents the hemi-glycosylated mAb. The mass spectrum of peak 3 (**Figure 5.13B**) shows a shift of CSD of monomer to lower charge states compared to the major proteoforms of the NISTmAb, and the dimer is not observed. Several major glyco-proteoforms can be identified by deconvolution, **Figure 5.13D**. The most abundant charge state (23+) in peak 3 is one less than that in peak 2 (24+). One possible explanation is the deamidation of the mAb, which is previously reported as a common post-translational modification for NISTmAb [46]. Unfortunately, we cannot accurately distinguish this 1-Da mass difference in our experimental condition. Nonetheless, the cIEF-assisted CZE shows great potential for the separation of different variants of mAbs in native conditions with large sample loading capacity.

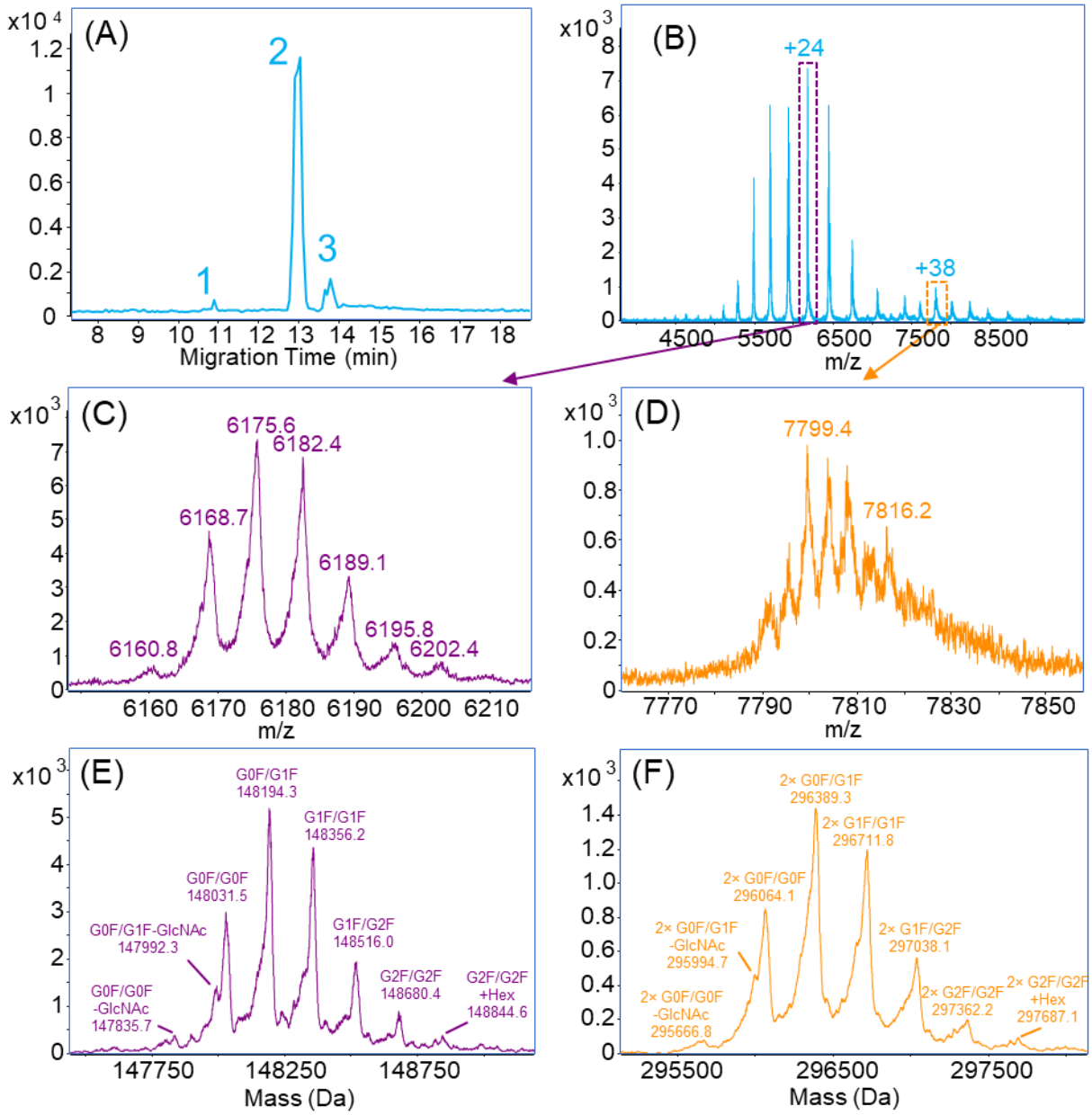

Figure 5.12. (A) Base peak electropherogram of native cIEF-assisted CZE-MS for the NISTmAb. (B) Mass spectrum averaged across the peak 2 in (A). (C, D) Zoom-in mass spectra of +24 and +38 charge states and (E, F) Deconvolution of NISTmAb proteoforms in the main peak (peak 2). Herein the LCP-coated capillary was used.

Table 5.2. Theoretical and observed masses of the glyco-proteoforms of NISTmAb monomer and homodimer detected in the main peak (peak 2) with the native cIEF-assisted CZE-MS.

Structure	Glyco-proteoform	Theoretical Mass (Da)[a]	Observed Mass (Da)	Mass Error (Da)	Mass Error (ppm)
Monomer	G0F/G0F – GlcNAc	147834.0	147835.7	1.7	11.5
	G0F/G1F – GlcNAc	147996.1	147992.3	3.8	25.7
	G0F/G0F	148037.2	148031.5	5.7	38.5
	G0F/G1F	148199.3	148194.3	5.0	33.7
	G1F/G1F	148361.4	148356.2	5.2	35.1
	G1F/G2F	148523.6	148516.0	7.6	51.2
	G2F/G2F	148685.7	148680.4	5.3	35.6
	G2F/G2F + Hex	148847.7	148844.6	3.1	20.8
Homodimer	2× G0F/G0F – GlcNAc	295668.0	295666.8	1.2	4.1
	2× G0F/G1F – GlcNAc	295992.2	295994.7	2.5	8.4
	2× G0F/G0F	296074.4	296064.1	10.3	34.8
	2× G0F/G1F	296398.6	296389.3	9.3	31.4
	2× G1F/G1F	296722.8	296711.8	11.0	37.1
	2× G1F/G2F	297047.2	297038.1	9.1	30.6
	2× G2F/G2F	297371.4	297362.2	9.2	31
	2× G2F/G2F + Hex	297695.4	297687.1	8.3	28

a: The theoretical masses are from reference 45.

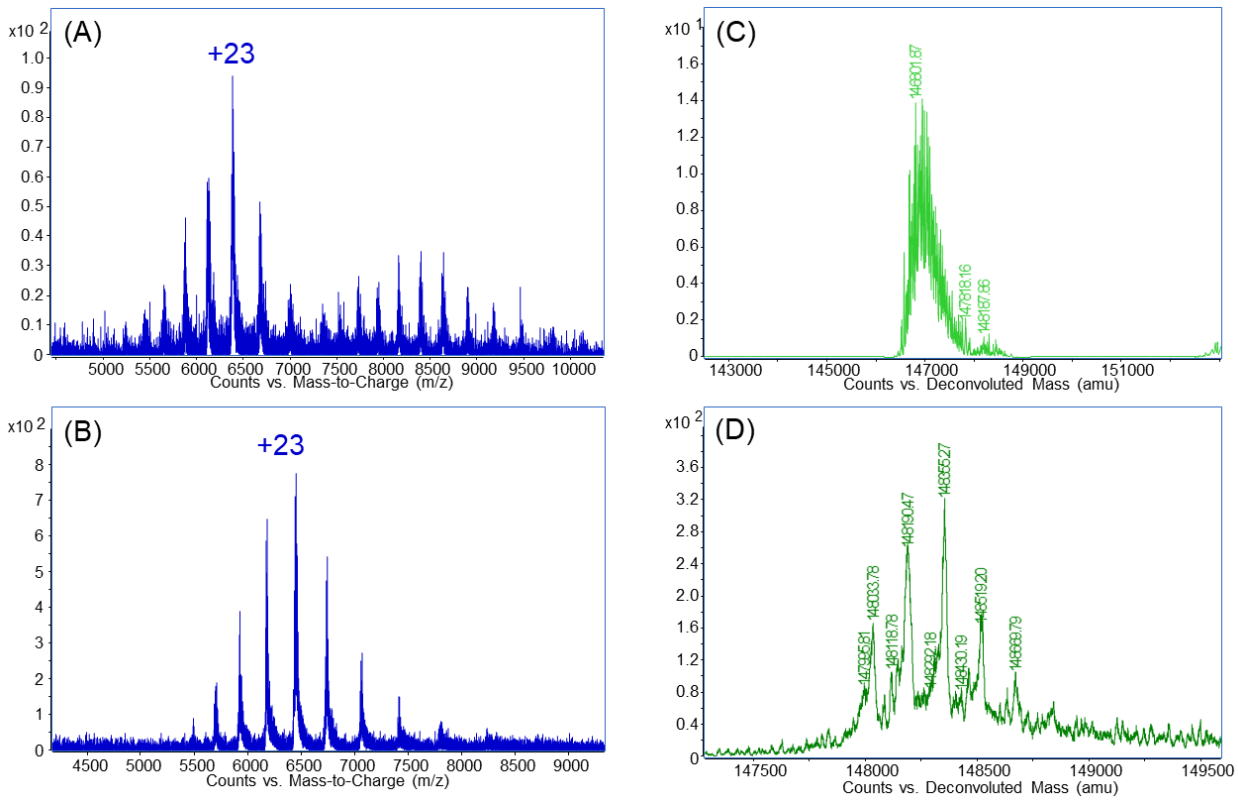

Figure 5.13. Mass spectra averaged across peak 1 (A) and peak 3 (B) in Figure 4A. Their deconvoluted mass spectra are shown in (C) and (D), respectively.

5.4 Conclusion

We developed a novel cIEF-assisted CZE-MS platform for the analysis of mAbs under native conditions with large sample loading capacity. The optimizations of the Q-TOF parameters and CZE conditions provide a reference guide to the community for the characterization of native mAbs with a Q-TOF mass spectrometer and CZE separation. With the new capillary coating and the online cIEF sample stacking, this platform achieved high-quality characterization of glyco-proteoforms, variants and aggregates of two mAbs. Using cIEF in a narrow pH range for sample stacking is the first attempt and a proof of concept to preconcentrate the analytes and increase the

loading capacity in native CZE conditions. We expect our novel platform could be a useful analytical tool for the characterization of various mAbs and large protein complexes.

Although we used Pharmalyte 3-10 in the experiments, we believe the ampholytes in a narrow pI range (*e.g.,* 6-9) would improve the stacking performance in the native conditions further and will be investigated in future studies. Another direction of improvement is to integrate gas-phase fragmentation in the platform (*i.e.* electron capture dissociation), which could offer more precise information of PTMs on mAbs and help us to understand the formation of different variants and aggregates during production processes. We also need to note that the separation performance of our native CZE-MS system for mAb charge variants needs to be boosted further via investigating different additives to the separation buffer of CZE and evaluating different sugar monomers for the LCP coatings.

5.5 Acknowledgment

We thank Agilent and CMP Scientific for their help for this project. In particular, we thank John Sausen (Director of Strategic Initiatives-Mass Spectrometry), Dr. David Wong, Dr. Caroline S. Chu, Dr. Christian Klein, Dr. Christopher Colangelo at Agilent and Dr. James Xia at CMP Scientific for their useful discussions about the data. We thank the support from the National Institute of General Medical Sciences (NIGMS) through Grant R01GM125991 and the National Science Foundation through Grant DBI1846913 (CAREER Award).

REFERENCES

REFERENCES

[1] P. J. Carter, G. A. Lazar, Nat. Rev. Drug Discov. 17 (2017) 197-223.

[2] P. Chames, M. Van Regenmortel, E. Weiss, D. Baty, Br. J. Pharmacol. 157 (2009) 220-233.

[3] D. R. Goulet, W. M. Atkins, J. Pharm. Sci.109 (2020) 74-103.

[4] H. Kaplon, J. M. Reichert, mAbs 11 (2019) 219-238.

[5] K. Groves, A. Cryar, S. Cowen, A. E. Ashcroft, M. Quaglia, J. Am. Soc. Mass Spectrom. 31 (2020) 553-564.

[6] K. Srzentić, L. Fornelli, Y. O. Tsybin, J. A. Loo, H. Seckler, J. N. Agar, L. C. Anderson, D. L. Bai, A. Beck, J. S. Brodbelt, Y. E. M. van der Burgt, J. Chamot-Rooke, S. Chatterjee, Y. Chen, D. J. Clarke, P. O. Danis, J. K. Diedrich, R. A. D'Ippolito, M. Dupré, N. Gasilova, Y. Ge, Y. A. Goo, D. R. Goodlett, S. Greer, K. F. Haselmann, L. He, C. L. Hendrickson, J. D. Hinkle, M. V. Holt, S. Hughes, D. F. Hunt, N. L. Kelleher, A. N. Kozhinov, Z. Lin, C. Malosse, A. G. Marshall, L. Menin, R. J. Millikin, K. O. Nagornov, S. Nicolardi, L. Paša-Tolić, S. Pengelley, N. R. Quebbemann, A. Resemann, W. Sandoval, R. Sarin, N. D. Schmitt, J. Shabanowitz, J. B. Shaw, M. R. Shortreed, L. M. Smith, F. Sobott, D. Suckau, T. Toby, C. R. Weisbrod, N. C. Wildburger, J. R. Yates, S. H. Yoon, N. L. Young, M. Zhou, J. Am. Soc. Mass Spectrom. 31 (2020) 1783-1802.

[7] A. Beck, E. Wagner-Rousset, D. Ayoub, A. Van Dorsselaer, S. Sanglier-Cianférani, Anal. Chem. 85 (2012) 715-736.

[8] R. Gahoual, A. Beck, E. Leize-Wagner, Y. -N. François, J. Chromatogr. B. 1032 (2016) 61-78.

[9] C. D. Whitmore, L. A. Gennaro, Electrophoresis 33 (2012) 1550-1556.

[10] R. Gahoual, A. Burr, J. -M. Busnel, L. Kuhn, P. Hammann, A. Beck, Y. -N. François, E. Leize-Wagner, mAbs 5 (2013) 479-490.

[11] O.O. Dada, Y. Zhao, N. Jaya, O. Salas-Solano, Anal. Chem. 89 (2017) 11236–11242.

[12] C. Lew, J. L. Gallegos-Perez, B. Fonslow, M. Lies, A. Guttman, J. Chromatogr. Sci. 53 (2015) 443-449.

[13] G. Chevreux, N. Tilly, N. Bihoreau, Analytical Biochemistry 415 (2011) 212-214.

[14] E. A. Redman, N. G. Batz, J. S. Mellors, J. M. Ramsey, Anal. Chem. 87 (2015) 2264-2272.

[15] Y. Zhao, L. Sun, M. D. Knierman, N. J. Dovichi, Talanta 148 (2016) 529-533.

[16] M. Han, B. M. Rock, J. T. Pearson, D. A. Rock, J. Chromatogr. B. 1011 (2016) 24-32.

[17] K. Jooß, J. Hühner, S. Kiessig, B. Moritz, C. Neusüß, Anal. Bioanal. Chem. 409 (2017) 6057-6067.

[18] A. M. Belov, L. Zang, R. Sebastiano, M. R. Santos, D. R. Bush, B. L. Karger, A. R. Ivanov, Electrophoresis 39 (2018) 2069-2082.

[19] J. Cheng, L. Wang, C.M. Rive, R. A. Holt, G. B. Morin, D. D. Y. Chen, J Proteome Res. 19 (2020) 2700–2707.

[20] X. Shen, Q. Kou, R. Guo, Z. Yang, D. Chen, X. Liu, H. Hong, L. Sun, Anal. Chem. 90 (2018) 10095–10099.

[21] X. Shen, Z. Yang, E. N. McCool, R. A. Lubeckyj, D. Chen, L. Sun, TrAC Trends in Anal. Chem. 120 (2019) 115644.

[22] A.M. Belov, R. Viner, M.R. Santos, D.M. Horn, M. Bern, B.L. Karger, A.R. Ivanov, J. Am. Soc. Mass Spectrom. 28 (2017) 2614–2634.

[23] M. Moini, Anal. Chem. 79 (2007) 4241–4246.

[24] H. J. Maple, O. Scheibner, M. Baumert, M. Allen, R. J. Taylor, R. A. Garlish, M. Bromirski, R. J. Burnley, Rapid Commun Mass Spectrom. 28 (2014) 1561–1568.

[25] V. Le-Minh, N. T. Tran, A. Makky, V. Rosilio, M. Taverna, C. Smadja, J. Chromatogr. A. 1601 (2019) 375-384.

[26] R. D. Smith, C. J. Barinaga, H. R. Udseth, Anal. Chem. 60 (1988) 1948–1952.

[27] P. Britz-McKibbin, D.D.Y. Chen, Anal. Chem. 72 (2000) 1242–1252.

[28] R. A. Lubeckyj, E. N. McCool, X. Shen, Q. Kou, X. Liu, L. Sun, Anal. Chem. 89 (2017) 12059–12067.

[29] R. Wojcik, O. O. Dada, M. Sadilek, N. J. Dovichi, Rapid Commun. Mass Spectrom. 24 (2010) 2554–2560.

[30] L. Sun, G. Zhu, Z. Zhang, S. Mou, N. J. Dovichi, J. Proteome Res. 14 (2015) 2312–2321.

[31] G. Zhu, L. Sun, N. J. Dovichi, Talanta 146 (2016) 839-843.

[32] E. N. McCool, R. Lubeckyj, X. Shen, Q. Kou, X. Liu, L. Sun, J. Vis. (2018) e58644.

[33] L. Sun, G. Zhu, Y. Zhao, X. Yan, S. Mou, N. J. Dovichi, Angew. Chem. Int. 52 (2013) 13661-13664.

[34] I. K. Ventouri, D. B. A. Malheiro, R. L. C. Voeten, S. Kok, M. Honing, G. W. Somsen, R. Haselberg, Anal. Chem. 92 (2020) 4292-4300.

[35] L. Konermann, J. Am. Soc. Mass Spectrom. 28 (2017) 1827-1835.

[36] M. Metzke, Z. Guan, Biomacromolecules 9 (2008) 208–215.

[37] M. Metzke, J. Z. Bai, Z. Guan, J. Am. Chem. Soc. 125 (2003) 7760–7761.

[38] S. Maiti, S. Manna, J. Shen, A.P. Esser-Kahn, W. Du, J. Am. Chem. Soc. 141 (2019) 4510–4514.

[39] S. Martinović, L. Paša-Tolić, C. Masselon, P. K. Jensen, C. L. Stone, R. D. Smith, Electrophoresis 21 (2000) 2368-2375.

[40] C. Przybylski, M. Mokaddem, M. Prull-Janssen, E. Saesen, H. Lortat-Jacob, F. Gonnet, A. Varenne, R. Daniel, Analyst 140 (2015) 543-550.

[41] X. Z. Wu, S. Asai, Y. Yamaguchi, Electrophoresis 30 (2009) 1552-1557.

[42] J. M. Cunliffe, Z. Liu, J. Pawliszyn, R. T. Kennedy, Electrophoresis 25 (2004) 2319-2325.

[43] V. M. Okun, Electrophoresis 19 (1998) 427-432.

[44] T. Formolo, M. Ly, M. Levy, L. Kilpatrick, S. Lute, K. Phinney, L. Marzilli, K. Brorson, M. Boyne, D. Davis, J. Schiel, Determination of the NISTmAb Primary Structure, American Chemical Society, 2015.

[45] Q. Dong, Y. Liang, X. Yan, S. P. Markey, Y. A. Mirokhin, D. V. Tchekhovskoi, T. H. Bukhari, S. E. Stein, mAbs 10 (2018) 354-369.