COUPLING LIQUID CHROMATOGRAPHY TO CAPILLARY ZONE ELECTROPHORESIS TANDEM MASS SPECTROMETRY FOR DEEP TOP-DOWN PROTEOMICS

By

Elijah Neal McCool

A DISSERTATION

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

Chemistry-Doctor of Philosophy

2021

ABSTRACT

COUPLING LIQUID CHROMATOGRAPHY TO CAPILLARY ZONE ELECTROPHORESIS TANDEM MASS SPECTROMETRY FOR DEEP TOP-DOWN PROTEOMICS

By

Elijah Neal McCool

Proteomes are very complex with a large number of unique proteoforms spread across a wide concentration dynamic range. This means that an MS-based platform with highly efficient separation and highly sensitive detection of proteoforms is required. Capillary zone electrophoresis-tandem mass spectrometry (CZE-MS/MS) has been suggested as one such platform. When coupled to offline liquid chromatography-based fractionation, CZE-MS/MS has proven to be invaluable to the TDP community.

In Chapter 2, the first optimization of dynamic pH junction-based sample stacking for TDP is provided along with one of the first comparisons of reversed-phase liquid chromatography coupled to mass spectrometry (RPLC-MS) and CZE-MS/MS. Optimization of dynamic pH junction is performed with a standard protein mixture, and this platform was ultimately applied to an *Eschericia coli* (*E. coli*) whole cell lysate. This resulted in the largest TDP dataset for single-shot CZE-MS/MS. The comparison of RPLC-MS/MS and CZE-MS/MS also included analysis of an *E. coli* cell lysate and resulted in high numbers of identifications and highlighted the various pros and cons of each method.

In Chapter 3, two dimensional LC fractionation (size exclusion chromatography (SEC) and RPLC) was coupled to CZE-MS/MS for deep TDP of *E. coli* cells. This study resulted in the largest TDP dataset, at the time, for *E. coli*, identifying 5700 proteoforms and 850 proteins. We were also able to identify and localize various interesting PTMs and estimate protein abundances using a spectral counting method. From this study it was clear that

our platform was comparable to other RPLC-MS/MS methods for deep TDP in terms of number of proteoform identifications and total instrument time.

In Chapter 4, we applied our TDP platform to two isogenic colorectal cancer (CRC) cell lines, SW480 and SW620, from primary and metastatic tumors. Genetic changes have been known for a long time to affect CRC progression but this was the first proteoform-level deep TDP study of CRC metastasis. In total, we identified over 23000 proteoforms and over 2000 proteins, for the largest TDP dataset of any cell type and was a 400% increase in terms of identifications over previous deep TDP studies. We used a special database searching tool to identify single amino acid variants (SAAVs) for the largest dataset of proteoforms containing SAAVs. Quantitative analysis identified 460 proteoforms with significant differences in abundance between SW480 and SW620. Several of these proteoforms were also phosphorylated which could further impact disease progression and outcome for a specific patient phenotype and could serve as biomarkers for deciding how to treat a patient or for drug development.

In Chapter 5, both activated ion electron transfer dissociation (AI-ETD) and ultraviolet photodissociation (UVPD) at 213 nm were coupled to CZE for deep TDP of *E. coli* and zebrafish brain samples, respectively. Optimized CZE-AI-ETD and CZE-UVPD resulted in large numbers of proteoform identifications, and many important modifications were identified and localized using these effective fragmentation techniques. This included N-terminal acetylation, methylation, S-thiolation, disulfide bonds, and lysine succinylation.

In Chapter 6, a variety of insights into the future of TDP are provided. This includes important applications for TDP, such as personalized medicine, drug development, embryonic development, and pathogen identification. Also, a few advancements to the TDP workflow that may have increased focus on in the future are mentioned.

This thesis is dedicated to Viki, Brian, Ean, and Ema McCool and Radha Patel.

ACKNOWLEDGMENTS

I would like to acknowledge and thank several people who have unwaveringly supported me for as long as I have known them. I want to thank my parents, Brian and Viki McCool, have encouraged me and believed in me, even when I did not have the same confidence. Your willingness to listen and to think critically inspire me and have unquestionably shaped who I am today. I want to thank my siblings, Ean and Ema, for comic relief and for tolerating me as the older sibling, which, as an older sibling, I had to mention. You both are amazing people and I look forward to seeing what you all accomplish on top of what you already have. I want to thank Radha Patel for sharing this journey with me and always cheering me on. You are uniquely compassionate and intelligent and I am very excited for what the future holds.

I want to thank my advisor, Professor Liangliang Sun, for all of his support over the past five years. Countless hours have been spent teaching me how to think about proteomics and how to use various technologies. He has made my time at MSU much more enjoyable and intellectually stimulating. He has also spent a lot of time advocating for me to get various awards and recognitions and oral and poster presentations. Dr. Sun has really been one of the best people to work for and the future holds great things for the Sun group.

I want to thank Professor Leslie Hicks for believing in me as a scientist before I believed in that about myself and for all of the helpful advice. Without the Hicks group I know that I would not be where I am today, and I sincerely thank all of them. I want to thank Professor Dana Spence for believing in me when I applied to MSU for my graduate studies and for your helpful insights and support. I want to thank my committee for reminding me to always stay curious and for teaching me how to be a better analytical chemist. I want to thank my group members, past and present, especially Daoyang Chen, Qianjie Wang, Tian Xu, Rachele Lubeckyj, Zhichang Yang, Qianyi Wang, Xiaojing Shen, and Jorge

Colón-Rosado. I have enjoyed getting to know you all over the past few years and I appreciate all of your insightful discussions, help with my research, and advice for my future.

I would also like to thank my collaborators, especially Professor Jose Cibelli and his student Billy Poulos, Professor Xiaowen Liu and his student Wenrong Chen, Professor Joshua Coon, Professor Yansheng Liu, and Professor Amanda Hummon. They have all provided invaluable insight and help towards a variety of projects and some really interesting parts of my research would not have been possible without them.

I would finally like to thank my friends for providing a much needed respite from the rigors of graduate school despite having very busy lives yourselves. I would not be able to list everyone here but I am sure you all know who you are.

It is impossible to thank everyone as thoroughly as I would like, but I would like everyone mentioned here to know how much they mean to me and how much I appreciate you and I am excited to keep in touch with all of you and continue to learn from you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AI-ETD	Activated ion electron transfer dissociation
BGE	Background electrolyte
BUP	Bottom-up proteomics
CZE	Capillary zone electrophoresis
CID	Collision-induced dissociation
DDA	Data-dependent acquisition
ETD	Electron transfer dissociation
EOF	Electroosmotic flow
μ_{eof}	Electroosmotic mobility
ESI	Electrospray ionization
FDR	False discovery rate
FESS	Field enhanced sample stacking
FT-ICR	Fourier transform-ion cyclotron resonance
GELFrEE	Gel-eluted liquid fraction entrapment electrophoresis
HETP	Height equivalent to a theoretical plate
HPLC	High performance liquid chromatography
HCD	Higher energy collisional dissociation
iTRAQ	Isobaric tagging for relative and absolute quantification
IEF	Isoelectric focusing
ITP	Isotachophoresis
LPA	Linear polyacrylamide
MS	Mass spectrometry
m/z	Mass-to-charge-ratio
MP	Mobile phase
NCE	Normalized collision energy

N	Number of theoretical plates
P_c	Peak capacity
PrSM	Proteoform spectrum match
QEHF	Q Exactive HF mass spectrometer
RPLC	Reversed-phase liquid chromatography
SWATH-MS	Sequential window acquisition of all theoretical mass spectra
SAAVs	Single amino acid variants
SEC	Size exclusion chromatography
MS/MS	Tandem mass spectrometry
TMT	Tandem mass tag
TOF	Time-of-flight
TDP	Top-down proteomics
UHPLC	Ultrahigh-pressure liquid chromatography
UVPD	Ultraviolet photodissociation

Chapter 1

Introduction

1.1 Foundations of proteomics

Proteins and protein complexes either participate in or control every biological process in cells. Proteomics is the large-scale study of the complement of proteins (proteomes) within cells and their dynamic regulation across various conditions [1, 2]. Proteomes are complicated with one gene producing many different protein molecules (proteoforms) due to individual genetic variations, RNA splicing, and post-translational modifications (PTMs), Figure 1.1 [3–8]. Recently, it has been estimated that the human proteome has more than

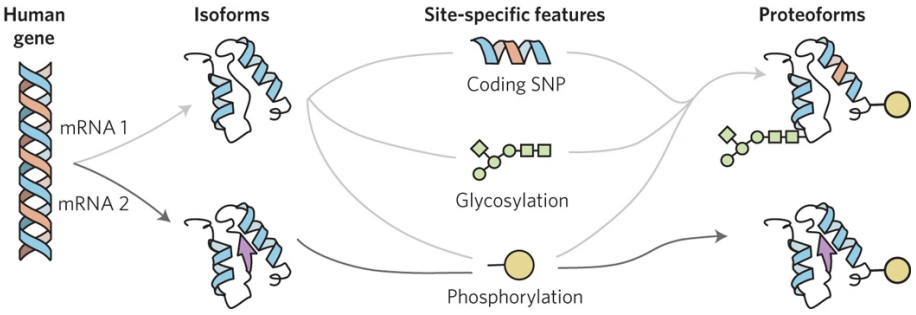


Figure 1.1: Illustration of the genetic and post-translational variability that results in proteoforms. This figure is reprinted with permission from reference [8].

one million unique proteoforms [8]. Proteoforms from the same gene can have very similar sequences as well as contain multiple PTMs, combinations of modifications, or single amino acid variants (SAAVs), making proteoforms difficult to distinguish from one another in complex samples. Proteomes also have a wide proteoform concentration dynamic range,

approaching seven orders of magnitude, adding another level of complexity [8–10]. Differentially modified proteoforms, many of which are low abundant, have been shown to be or are likely to be uniquely integral to various biological processes, including cancer and embryonic development [11–13]. Thus, proteome-scale investigation of proteoform changes during these processes are of great interest to both the scientific and medical communities.

1.2 Electrospray ionization mass spectrometry

Investigation of proteins within cells frequently involves intact and fragment mass analysis of either peptides or proteins by mass spectrometry (MS), the data from which is passed through some sort of database searching protocol to match with theoretical sequences from the genome for protein identification.Introduction of peptides and proteins into the mass spectrometer is made possible through a variety of methods that generate gas phase ions that can then be detected and subsequently or simultaneously fragmented. Simply, ionization techniques can be distinguished from each other based on their relative softness or hardness, or the degree to which the molecular ion is conserved during the ionization process [14]. Electrospray ionization (ESI), seen in Figure 1.2, pioneered for biomolecules by John Fenn in 1984, is a soft ionization technique, producing intact multiply charged ions, with preserved labile modifications, and is the most commonly used ionization technique for peptides and proteins [15–17]. Charge state depends greatly on the size of a particular peptide or protein, with higher mass species having higher charge states, and makes peptides and proteins mass-to-charge ratio (m/z) much more amenable to measurement by a mass spectrometer. Peptides or proteins that pass through an emitter, with an applied potential between the emitter tip and the mass spectrometer, where charge is concentrated at the liquid surface, causing formation of a Taylor cone, from which charged droplets are formed. These charged droplets undergo evaporation and charge concentration at the surface of the droplet. These droplets eventually become smaller, until

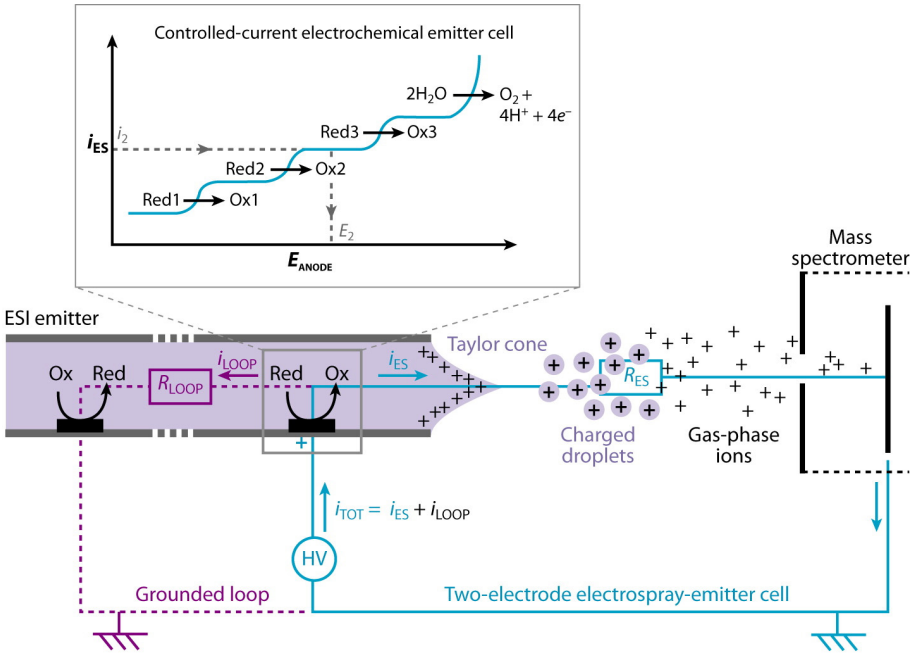


Figure 1.2: Illustration of electrospray ionization using a metal emitter tip. This figure is reprinted with permission from reference [15].

the Rayleigh limit of stability is reached, and undergo a series of Coulomb explosions resulting in smaller droplets. Ultimately, the peptide or protein is either ejected or dried to the point of becoming charged gas phase ions that can then enter the mass spectrometer through a vacuum and series of voltage drops.

MS separates these charged species by their m/z, which can be accomplished through a variety of means. Traditional instruments for mass analysis include time-of-flight (TOF), magnetic sector instruments, linear quadrupoles, linear quadrupole ion traps, quadrupole ion traps, Fourier transform-ion cyclotron resonance instruments (FT-ICR, and Orbitraps (Figure 1.3) [14]. The most common instruments for analysis of intact proteins are FT-ICR, Orbitrap, and TOF. FT-ICR instruments are usually ran at higher resolution ($> 10^6$) and have high mass accuracy, leading to fewer false peptide and protein identifications [3, 14, 19–24]. Simply, FT-ICR instruments trap ions in a fixed magnetic field (Penning trap), followed by excitation by an electric field, and ultimately detection of

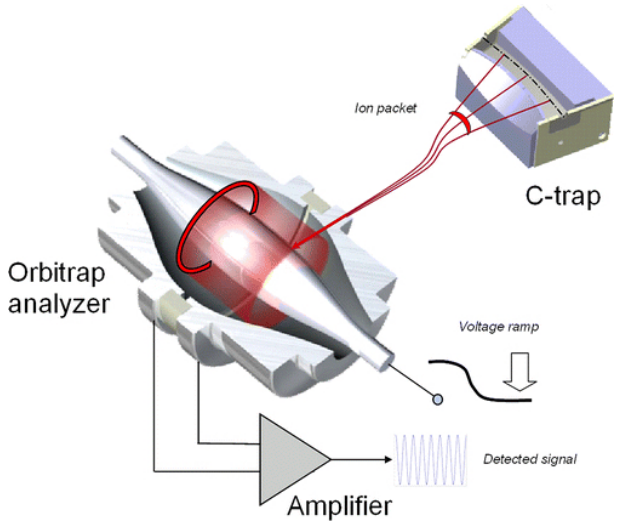


Figure 1.3: Illustration of the cross-section of the C-trap and Orbitrap mass analyzer. This figure is with permission from reference [18].

image current in the time domain. This data can be transformed into the frequency domain (Fourier transformation), and related back to the m/z of the ions.

Like FT-ICR instruments, Orbitraps utilize Fourier transformation to convert an image current into the frequency domain. The Orbitrap was invented by Alexander Makarov in 2000 and first released by Thermo Fisher Scientific in 2005 [25]. Briefly, Orbitraps store ions in the rf-only C-trap followed by an rf down ramp and high voltage pulse that introduces ions into the Orbitrap analyzer for mass analysis, Figure 1.3 [18]. Orbitraps do not require a magnetic field to operate, allowing labs, otherwise limited by price and size constraints, to run at high resolving powers with accurate mass measurements (sub-ppm) [14]. Modification of Orbitrap instruments has extended the mass range significantly, improving the Orbitrap's applicability to high mass species [26].

Although lower resolution, TOF instruments have high mass accuracy, high frequency of spectra acquisition, and a theoretically unlimited m/z range [14]. This makes TOF instruments an important tool in proteomics experiments for time-limited experiments, discovery of high mass species, and analyte quantification [27]. TOF instruments operate

by accelerating ions with an electric field, with the resulting ion velocity dependent upon the m/z of the ion. Various modifications to TOF instruments, including addition of the reflectron and quadrupole mass analyzer to the front-end of TOF instruments, have improved their resolution and mass accuracy [28, 29].

1.3 Bottom-up proteomics

The most widely used proteomics strategy is bottom-up proteomics (BUP). During a typical BUP workflow, proteins are digested into peptides, usually by trypsin which cleaves at lysine and arginine, and these peptides are passed through a reversed-phase liquid chromatography tandem mass spectrometry (RPLC-MS/MS) workflow. Peptide and fragment ion masses from mass analysis are then passed through a database searching protocol where protein identifications are inferred through peptide identifications [30].

BUP is highly sensitive, with demonstrated low zmole limit of detection (LOD) for peptides using capillary zone electrophoresis mass spectrometry (CZE-MS) [31]. However, BUP suffers from the protein inference problem meaning it can only provide limited information about distinct proteoforms, which can be visualized in Figure 1.4 [32]. However, BUP has been and will continue to be useful to the proteomics and medical

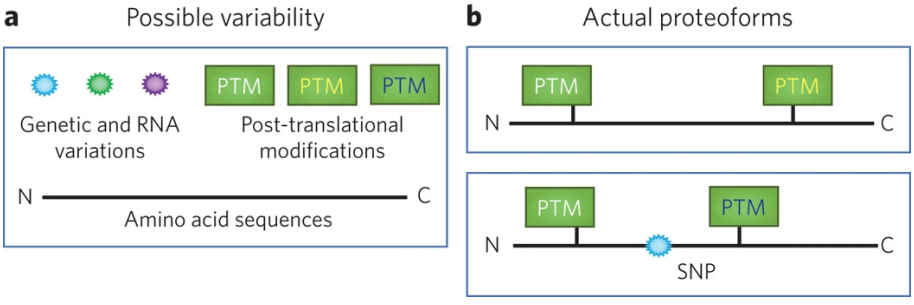


Figure 1.4: Visual representation of distinct proteoforms in cells with (a) sources of variability and (b) resultant proteoforms [8].

community due to the highly developed and extremely high throughput pipelines for

peptide analysis as well as the ability to multiplex through chemical modification of the peptides [30]. Several established workflows exist for selective purification of peptides containing modifications, such as phosphorylation, meaning that low abundant proteins containing these modifications can be detected [30]. Also, the process of creating peptides artificially increases markers for proteins that otherwise may have abundance below the dynamic range of the mass spectrometer. In terms of identifications, BUP is able to identify tens of thousands of peptides and many thousands of proteins [33].

1.4 Top-down proteomics

Top-down proteomics (TDP) directly characterizes proteoforms within cells in their intact form [3, 11, 34]. In a typical TDP workflow proteoforms are usually extracted from cells and fractionated using either LC or electrophoresis, followed by analysis with RPLC-MS/MS, Figure 1.5. Proteoform identification is then accomplished by matching

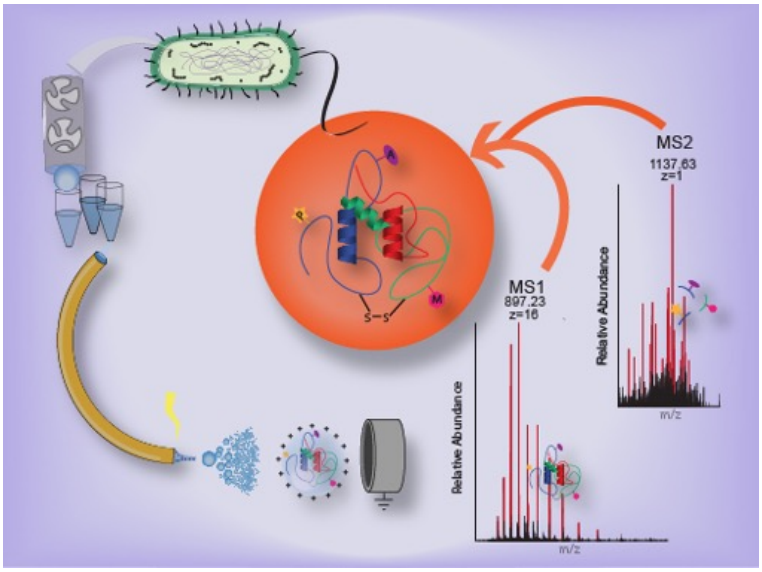


Figure 1.5: A typical TDP workflow.

experimental data to a theoretical protein database based on the genome under investigation. Because TDP analyzes intact proteins, TDP can identify the distinct

proteoforms mentioned above that contain multiple PTMs, combinations of modifications, or SAAVs. Modern deep TDP workflows are able to identify hundreds to low thousands of proteoforms, with notably fewer protein identifications than BUP [33]. Sensitivity for TDP has been improved through the use of CZE-MS with hundreds to thousands of proteoforms identified from tens to hundreds of nanograms of proteins, with 10-30-fold less sample consumption with nanoRPLC-MS [35–37]. However, TDP comes with a host of issues throughout traditional workflows that has kept TDP from becoming as biologically impactful as other -omics methods.

1.5 Challenges of top-down proteomics

1.5.1 Sample preparation

Sample preparation for intact proteins, with a wide range of sizes and diverse properties, is complicated, and membrane protein or extracellular matrix protein samples are notorious for having issues with solubility [38–40]. TDP also starts at a disadvantage without the ability to artificially increase proteoform abundance, making any sample handling and preparation of utmost importance to ensure high protein recovery. Protein bias, reproducibility, and compatibility with downstream analyses of the sample preparation method are also important considerations [38]. For example, chaotropic agents (e.g. urea) and detergents (e.g. sodium dodecyl sulfate, SDS) are commonly used for protein extraction and denaturation, but are incompatible with mass spectrometry, see Figure 1.6 [38, 41, 42]. Protein samples need to be cleaned-up following the use of any incompatible sample buffers. Membrane ultrafiltration, chloroform-methanol precipitation, and SP3 sample clean-up methods have recently been tested for deep TDP, with membrane ultrafiltration showing the most promising results [38]. MS-compatible surfactants, such as Azo, have been developed for TDP and are explored in more detail elsewhere [43].

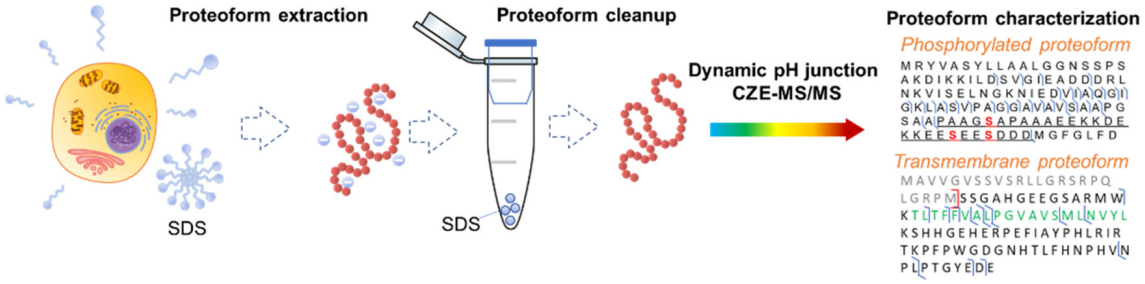


Figure 1.6: An illustration of TDP sample preparation. This figure is reprinted with permission from reference [38].

TDP has proven to be very useful for targeted analysis of proteins, however, this is usually accomplished with antibody-based approaches for affinity purification [44, 45]. At the proteoform-level, antibodies could have more favorable, and therefore more biased, interactions with certain proteoforms within a particular family of proteins.

1.5.2 Separations

As mentioned previously for the typical TDP workflow, Figure 1.5, following sample preparation, protein samples are usually fractionated prior to RPLC-MS/MS. In TDP, efficient and selective separation of intact proteins is vital for downstream identification due to the number and wide concentration dynamic range of unique proteoforms within cells. Separation of intact proteins is extremely difficult compared to smaller molecules. Fundamentally, separation of proteins can be understood by the van Deemter equation, which encapsulates the causes of band broadening during chromatographic experiments, Equation 1.1.

$$HETP = A + \frac{B}{\mu} + C\mu \tag{1.1}$$

In this equation, the height equivalent to a theoretical plate (HETP) is dependent upon multiple flow paths (A term) longitudinal diffusion (B term), and resistance to mass transfer (C term), where μ is mobile phase linear velocity [46] This is the most basic representation of the van Deemter equation, but is all that is necessary for the forthcoming

discussion of intact protein separations. For large biomolecules, such as proteins, the most significant term from Equation 1.1, in terms of separation performance, is the C term, or the resistance to mass transfer. Multiple flow paths and longitudinal diffusion are mostly negated through the large number of particles in the column packing and the low diffusion coefficients of large proteins. Basically, resistance to mass transfer is inversely proportional to the diffusion coefficient, D_s, of a protein, meaning that for large proteins with small D_s, this term becomes more significant and contributes more to band broadening.

The larger the HETP, the larger the band broadening, or the wider the peaks in a chromatogram. Therefore, it is relatively easy to imagine that the wider the peaks in a chromatogram, the more overlap there will be between analytes in the sample. Higher overlap between analytes, especially for proteins, means that there is less time for the mass spectrometer to collect useful mass spectra for all species present in a sample. Also, higher abundant species that overlap with lower abundant species will, sometimes, completely drown out their signal, leading to fewer identifications and poor proteoform characterization in a proteomics experiment.There are several other factors that impact band broadening in chromatography, and other resources go into more detail [46]. For open-tubular columns, as is the case for CZE, there is no stationary phase, therefore the A term and C term are negated, leaving only the B term to contribute to band broadening.

Number of theoretical plates (N) is one of the more common terms for describing the efficiency of a separation, Equation 1.2.

$$N = \frac{L}{H} \tag{1.2}$$

In Equation 1.2, L represents the length of the column and H represents the plate obtained from Equation 1. For high efficiency separations, a small HETP is observed, resulting in high N. Another commonly used term to describe separation efficiency in proteomics studies is peak capacity (P_c). Simply, P_c describes the number of peaks that can be

separated in a given separation window [47]. Traditional proteomics workflows provide P_c values of around 200 using high performance liquid chromatography (HPLC) and around 400 with ultrahigh-pressure liquid chromatography (UHPLC) or with long columns (100-200 cm) optimized for intact proteins [48, 49]. Clearly, this is far from the nearly one million unique proteoforms theoretically present in the human proteome, placing analytical stress on the mass spectrometer in TDP studies.

State-of-the-art separations of intact proteins are performed by 2D-gel electrophoresis which attains peak capacities in the thousands, but low recovery, the presence of SDS, and extensive set-up limit its application within TDP [11]. However, the mechanisms of separation in 2D-gel electrophoresis (charge and size) can be used as a foundation for intact protein separations. In one TDP study, Tran, et al. utilized solution isoelectric focusing (sIEF), gel-eluted liquid fraction entrapment electrophoresis (GELFrEE), and nano-capillary RPLC coupled online to the mass spectrometer [11]. The charge and size sorting steps are sIEF and GELFrEE, while nanocapillary RPLC offers a separation based on hydrophobicity. This platform achieved a peak capacity before mass spectrometric analysis of $\sim 2,500$ and 3,000 proteoform identifications (20-fold increase). However, identifications are still limited by the dynamic range of the mass spectrometer making it clear that orthogonal separations are necessary to expand the coverage of complex proteomes [11, 22].

While sIEF and GELFrEE separated intact proteins efficiently and reduced some of the limitations associated with 2D-gel electrophoresis, the use of SDS required sample cleanup prior to mass spectrometric analysis [11]. Cai, et al. attempted to address SDS limitations by using serial size exclusion chromatography (sSEC) coupled offline to RPLC-MS and RPLC-MS/MS [50]. sSEC-RPLC-MS (10 fractions) boosted the number of identified proteoforms from ~ 900 to $\sim 4,000$ compared to RPLC-MS alone demonstrating the selectivity of the sSEC fractionation method despite low peak capacity ($P_c \sim 10$). Both

studies used RPLC-MS as their final dimension of separation because of its well-known ability to be coupled to mass spectrometry in peptide analysis. However, RPLC has low separation efficiency for intact proteins, and significant sample loss on columns can occur due to the high surface area of the beads.

Shen, et al. recently attempted to optimize RPLC parameters including column length, length of bonded stationary phase, particle physiochemistry, and particle size [49]. Peak capacities > 400 were obtained, achieving a separation efficiency similar to peptide analyses. However, the separation efficiency of RPLC is limited due to the number of trade-offs between particle properties, separation length/time, and maximum pressure limitations of current instruments [49]. This highlights the need for a unique, high resolution final dimension of separation that could be coupled to mass spectrometry.

1.5.3 Capillary zone electrophoresis

CZE-MS is a sensitive and effective method both for the separation of intact proteins (100-fold less sample needed compared to RPLC-MS) and for identifying large proteins in complex mixtures when coupled to mass spectrometry [51–57]. CZE is a separation method based on size and charge, performed in an open tube capillary, Figure 1.7.

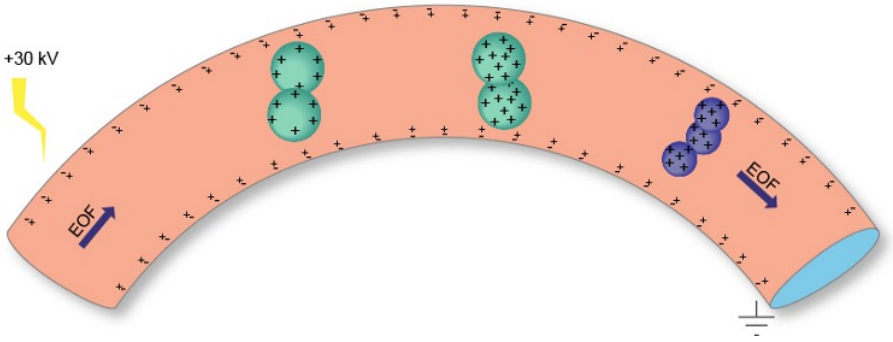


Figure 1.7: Basic illustration of a traditional CZE separation.

Electrophoretic mobility differences between analytes in the capillary drive the

separation in CZE. The speed of analyte movement out of the capillary depends upon (1) analyte mobility and (2) electroosmotic flow (EOF) created by an electric double layer, consisting of the Stern layer and diffuse layer, where immobile charge on the capillary wall and charge from the background electrolyte (BGE) can move in an electric field, Figure 1.8.

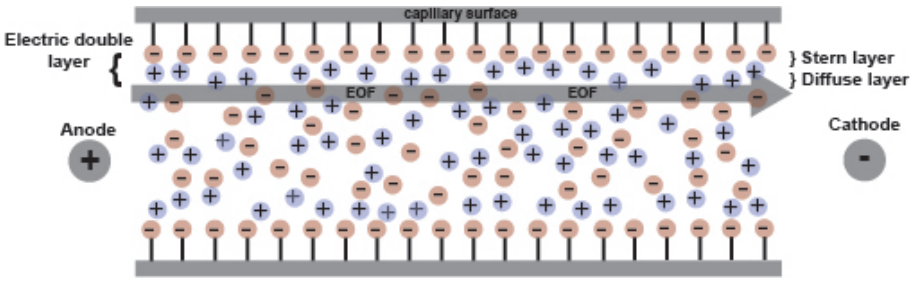


Figure 1.8: Illustration of EOF.

EOF shortens the separation time, but also reduces the separation window, thus reducing the number of proteoform identifications that can be achieved in a complex mixture [58]. EOF can be controlled through the thickness of the electric double layer, shown in Figure 1.8. Zeta potential is directly proportional to the thickness of the double layer. Lower charge density at the capillary wall lowers the zeta potential, Equation 1.3.

$$\zeta = \frac{\delta\sigma}{\varepsilon_0\varepsilon_r} \tag{1.3}$$

ζ is the zeta potential, σ is the charge density on the inner wall, ε_0 is permittivity of a vacuum, and ε_r is the buffer dielectric constant. Lowering the charge density can be achieved with higher electrolyte concentration in the BGE, Equation 1.4.

$$\delta = \left(\frac{\varepsilon_0\varepsilon_r RT}{2cF^2}\right)^{1/2} \tag{1.4}$$

R is the universal gas constant, T is the absolute temperature, c is the molar concentration, and F is the Faraday constant. Lowering the zeta potential ultimately

lowers EOF according to Equation 1.5.

$$\mu_{eof} = \frac{\varepsilon\zeta}{4\pi\eta} \tag{1.5}$$

μ_{eof} is the electroosmotic flow, ε is the buffer dielectric constant, and η is the buffer viscosity. Also, neutral capillary coatings, such as linear polyacrylamide (LPA), reduce both the EOF and any protein analyte adsorption on the capillary wall [55, 58]. The process of coating capillaries with LPA is reproducible and the coating itself is stable [52, 58–60]. The fused silica capillary wall is first derivatized with 3-(trimethoxysilyl)propyl methacrylate, which leaves an acrylic group exposed on the capillary wall surface [58, 61]. A solution containing acrylamide and ammonium persulfate (initiator) is then introduced into the capillary after degassing with N_2. Polymerization, along the capillary wall on the free acrylic groups, is then initiated by heating at 50°C, which decomposes the persulfate and generates radicals, for 30 min.

N in CE is calculated according to Equation 1.6, where μ is equal to electrophoretic mobility (from both the electric field and EOF), V is the applied voltage, and D is the diffusion coefficient of the analyte [56].

$$N = \frac{\mu V}{2D} \tag{1.6}$$

Therefore, the number of theoretical plates is not dependent upon the capillary length or analysis time, and higher voltages should supply higher separation efficiency. Since large molecules have lower diffusion coefficients, they should have higher N, making CE a seemingly perfect application for intact proteins. Another useful equation in CE experiments is the equation for analysis time (t) shown in Equation 1.7.

$$t = \frac{L}{v} = \frac{L^2}{\mu V} \tag{1.7}$$

The v term is equal to the migration velocity of the analyte from both the electric field and EOF, and L is the length of the capillary. Therefore, it follows that higher voltages and shorter capillaries would generate the greatest N with the shortest analysis time. Using high voltage, our group has been able to obtain N values in the hundreds of thousands for intact proteins in a relatively short analysis time [52].

Traditionally, CZE has low loading capacity (tens of nL) and short separation windows ($\sim$ 30 min), which limit identification of low abundance proteoforms and the number of MS/MS spectra that can be obtained [22, 52–54]. Attempts at increasing the loading capacity and separation window of CZE have included various online sample preconcentration methods, including dynamic pH junction, field enhanced sample stacking (FESS), and isotachophoresis [51, 58, 62–66]. In typical dynamic pH junction preconcentration, application of a positive potential across the capillary causes negatively charged analytes in a basic sample buffer (e.g. ammonium bicarbonate) to migrate to the proximal (injection) end of the capillary, where they come in contact with the acidic BGE on both sides of the sample plug, Figure 1.9 [58, 63, 64]. Thus, as analytes are titrated by

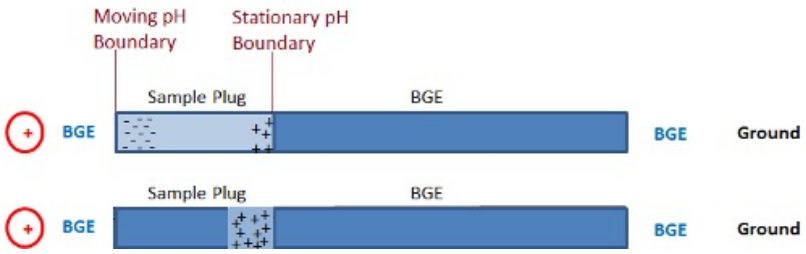


Figure 1.9: Basic illustration of dynamic pH junction sample stacking.

the BGE, two pH boundaries are formed, one stationary and one mobile. The analytes are concentrated into a short sample plug between these two boundaries. Once these boundaries meet, the sample undergoes an isotachophoresis mechanism which further concentrates our samples and is described elsewhere [64]. After concentration, the positively charged sample continues through the capillary, undergoing conventional CZE

separation. Although the dynamic pH junction method was systematically evaluated for online concentration of metabolites and peptides, it had not been systematically investigated for online concentration of intact proteins until recently [52]. FESS is a well-known method where the sample is dissolved in lower conductivity buffer compared to the BGE causing sample stacking due to analyte velocity differences in the sample and BGE zones [51, 65, 66]. Isotachophoresis is a less widely used approach to sample stacking, but, simply, analytes are focused based on their mobility versus the mobility of leading and terminating (fast and slow) electrolytes [67].

Our group mainly utilizes dynamic pH junction which has also been used in various other groups in proteomics experiments. Using 5 mM ammonium bicarbonate sample buffer and 5% acetic acid BGE for dynamic pH junction, Zhao, et al. were able to attain a separation window and peak capacity of around 30 min and 100 respectively while injecting $\sim$ 200 nL of sample [53]. Application of this method to a yeast lysate resulted in 580 proteoform identification from 23 fractions (RPLC) and $<$ 200 proteoforms identified per fraction.

1.5.4 Capillary zone electrophoresis electrospray ionization mass spectrometry

CZE-MS is made possible through a nanospray sheath-flow interface for a stable electrospray with low flowrate (nL/min), resulting in more efficient ionization, Figure 1.10 [31, 68–72]. The end of the separation capillary is etched down to 70-100 μm, for a 50 μm i.d. capillary, with hydrofluoric acid (HF) and fed through the glass emitter, with a 15 to 35 μm orifice diameter, to within a millimeter of the orifice, Figure 1.11 [70]. The use of glass electrospray emitters in this set-up eliminates redox reactions that may affect peptides and proteins and corona discharge that would be present with metal emitters which limits electrospray [68]. Optimization of this interface is reported elsewhere [70].

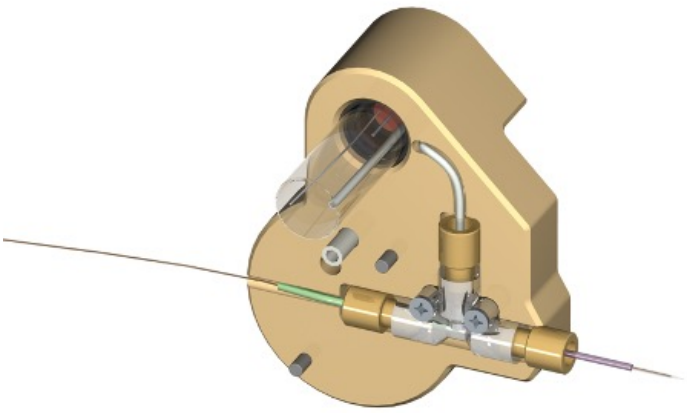


Figure 1.10: Figure of the electrokinetically pumped sheath flow interface from CMP Scientific used in our lab.

As highly complex biological samples are analyzed in TDP, more efficient separations and dynamic range/resolution improvements of the mass spectrometer are necessary. High resolution and mass accuracy during MS diminishes the overlap between co-eluting proteins with similar m/z and lowers the number of potential false positive identifications [22]. However, there is a reasonable limit to resolution in most mass analyzers. Orbitrap mass analyzers, for example, have resolution inversely proportional to the square-root of m/z and directly proportional to acquisition time [22]. The Q Exactive HF (QEHF) mass spectrometer is a potential solution to these limitations and offers high speed, resolution, and sensitivity [22, 73]. For example, BUP analysis of a complex biological sample using CZE-MS (Q Exactive mass spectrometer) has achieved low zmole peptide detection limit and high resolution is routinely used in proteomics experiments [31]. In summary, CZE addresses many of the limitations associated with other prefractionation approaches and can be coupled to mass spectrometry for efficient separation and identification of proteoforms. Combining CZE-MS with orthogonal LC methods could further improve the scale of TDP to provide deep proteome coverage.

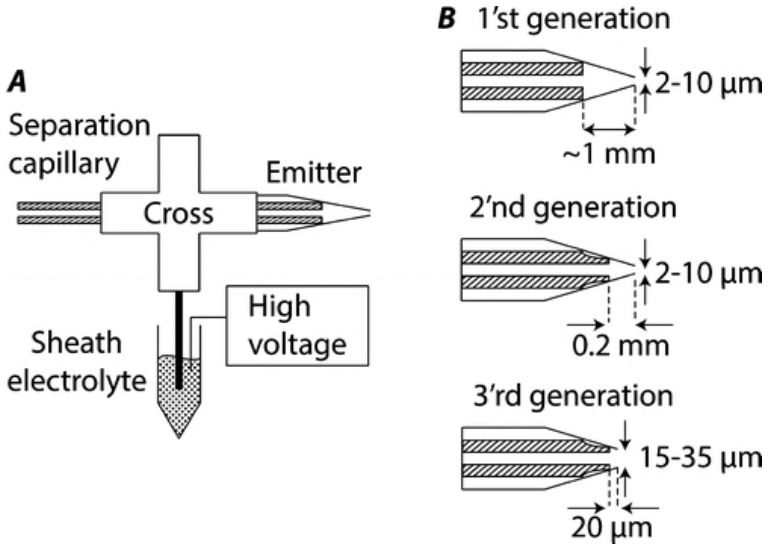


Figure 1.11: Illustration of the electrokinetically pumped sheath flow interface (A) with a zoomed in view of the separation capillary and emitter through various generations of optimization (B). This figure is reproduced with permission from reference [70].

1.5.5 Throughput

Efficient separation of intact proteins inherently takes time, due to many of the restraints previously mentioned, negatively impacting the throughput of TDP experiments. This not only applies to the final dimension of separation coupled directly to the mass spectrometer, but also any other dimensions of separation that are used during fraction collection. TDP also suffers from large amounts of dead time during the final dimension of separation coupled to mass spectrometry. This is demonstrated by Figure 1.12, which shows the number of identified proteoform-spectrum matches per minute as a function of the migration time in a TDP study [62]. Basically, this figure shows when useful data is being produced during the CZE separation. Ultimately, the amount of dead time makes the use of TDP, compared to BUP and other-omics methods, for precise medical approaches tailored to an individual patient's phenotype or biomarker development next to impossible [12, 40, 74, 75]. There are a few approaches that have been used in proteomics to try and

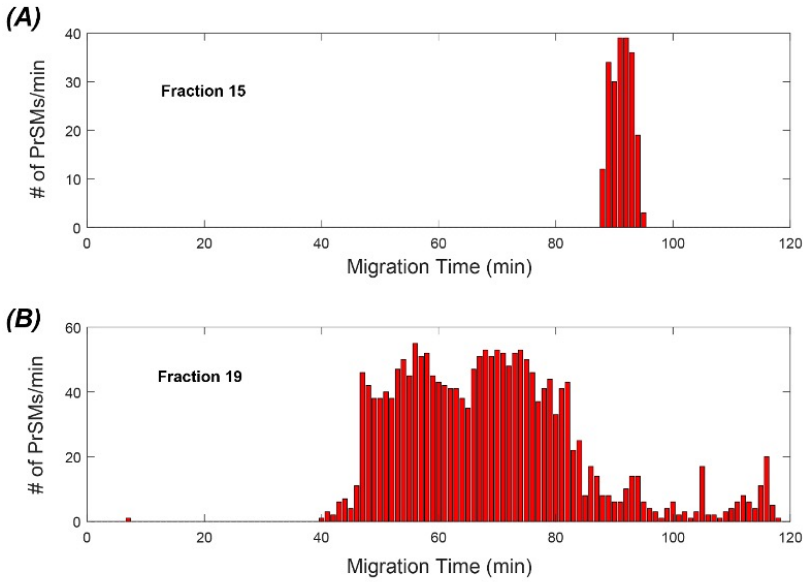


Figure 1.12: The number of identified PrSMs per minute as a function of the migration time for two fractions, (A) Fraction 15 and (B) Fraction 19, from a deep TDP study [62].

increase throughput including sequential injections in CZE and other creative sampling/separations [76–78].

Simply, sequential injection involves the introduction of multiple samples to a capillary followed by mass analysis. Sequential injection has been relatively successful in BUP experiments, but, in our experience, has proven less useful in TDP. There are several possible reasons for this, including a significant voltage drop between the first sample plug and any other samples subsequently introduced into the capillary. This means that separation efficiency and preconcentration of these subsequent samples may be affected. Isobaric tags are very popular in BUP and aids in throughput by being able to run multiple samples at the same time (multiplexing). More details about isobaric tags for TDP is provided in subsection 1.4.6 Proteoform identification and quantification. SampleStream is an example of a creative sampling and separation platform for online immunoprecipitation (IP) coupled directly to MS (IP-SampleStream-MS for TDP [75]. IP-SampleStream-MS showed $>$ 7-fold sample processing rate resulting in faster per-sample run times than

LC-MS, clearly improving throughput and stability. However, this method ultimately sacrifices depth, making it more likely to be useful for validating biomarkers [75].

1.5.6 Proteoform fragmentation

Identification of proteoforms involves a database search using parent and fragment ion data. High speed and range of protein isolation and fragmentation is important for confident and complete proteoform identification and is mostly dependent on the mass spectrometer, with well known trade-offs [22]. Typically, the intact mass of a proteoform is determined using MS data (intact mass), and MS/MS data is searched against a theoretical spectra database created from sequence stretches that match the measured mass of the unknown [3, 62, 79–81]. High resolution mass analysis, necessary for resolution of isotopic peaks and charge determination, of large proteins also takes more time when using FT instruments, resulting in a general hurdle for identifying co-eluting proteoforms in complex mixtures [22]. However, more often than not, the depth of information that can be gleaned using TDP about proteoform-specific modifications depends on both the efficiency and variety of fragmentation.

Fragmentation usually occurs after some sort of internal energy distribution (activation) followed by dissociation or by capture of a near thermal electron [82–87]. Our current method of fragmentation is higher energy collisional dissociation (HCD), where parent ions collide with an inert gas (N_2 in our case). As a high energy event with short interaction times, HCD causes electronic ion excitation followed by vibrational internal energy redistribution, which could result in fragmentation if the activation barrier for bond cleavage is exceeded [86, 88]. When HCD occurs, the resulting fragment ion series are b and y type, Figure 1.13. An example fragmentation of a 30 kDa protein, carbonic anhydrase (CA), is shown in Figure 1.14. These types of fragments are defined by the position of fragmentation along the peptide backbone consisting of Cα-C, C-N and N-Cα bonds. Cα-C, C-N and N-Cα bonds are associated with a and x, b and y, and c and z ion

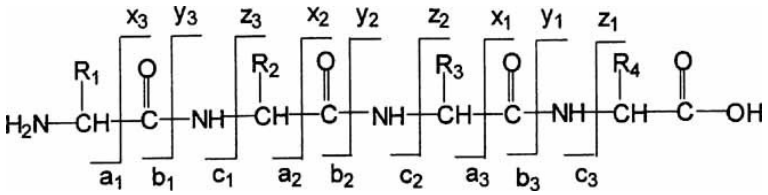


Figure 1.13: Schematic of a and x, b and y, and c and z ion series. This figure is reproduced with permission from reference [89].

Figure 1.14: Example fragmentation pattern of carbonic anhydrase utilizing HCD.

series, respectively, with the particular series distinguished by the side that retains the positive charge [88]. The fragmentation reaction itself can be understood using the "mobile proton model" where fragmentation of a protonated species involves a proton at the cleavage site that withdraws electron density from nearby bonds followed by electron migration and fragmentation [81, 90]. Determination of the amino acid sequence depends on the mass differences between members of either the b or y ion series, which allow amino acid assignment to the extra residue in the larger of the fragments in that series [86, 88].

Electron transfer dissociation (ETD) is another popular technique used for fragmenting intact proteins. ETD transfers an electron from a charged anion and can be used in the RF fields of the most popular mass analyzers [86]. Fragmentation of the protein or peptide using ETD follows a process without vibrational internal energy redistribution and involves

the release of a hydrogen radical [86]. Determination of amino acid sequence using ETD is very similar to HCD, except ETD produces mass differences in mostly c or z type ion series. Recently, an activated ion electron transfer dissociation (AI-ETD) method that combines infrared photoactivation concurrent with ETD has been developed and systematically evaluated for fragmentation of intact proteins [91–94]. AI-ETD showed better performance than HCD and standard ETD, regarding sequence coverage of identified proteoforms and proteoform characterization scores, and other results have demonstrated a good complementarity of HCD and AI-ETD for intact protein fragmentation. An example fragmentation pattern utilizing AI-ETD and HCD for intact CA is shown in Figure 1.15. Fragmentation techniques without vibrational internal energy

Figure 1.15: Example fragmentation pattern of carbonic anhydrase utilizing AI-ETD and HCD. This figure is reprinted with permission from reference [93].

redistribution also result in less biased fragmentation along the protein backbone and, theoretically, better PTM localization. However, the use of these fragmentation techniques has mainly been limited to targeted analysis of intact proteins and not high-throughput TDP workflows, even though the speed of these gas phase reactions occur on a time-scale well within the limits of the mass spectrometer [14].

Ultraviolet photodissociation (UVPD) is fragmentation technique where molecules are irradiated with high-energy photons, mainly 193 and 213 nm, to heat proteins and directly dissociate along the backbone, generating a wide variety of fragments and superior

sequence coverage [95–97]. Close to complete sequence coverage of intact proteins using UVPD (193 nm) has been demonstrated, allowing for in depth characterization of proteins, see Figure 1.16 [95]. However, 213 nm UVPD is the only commercially available method

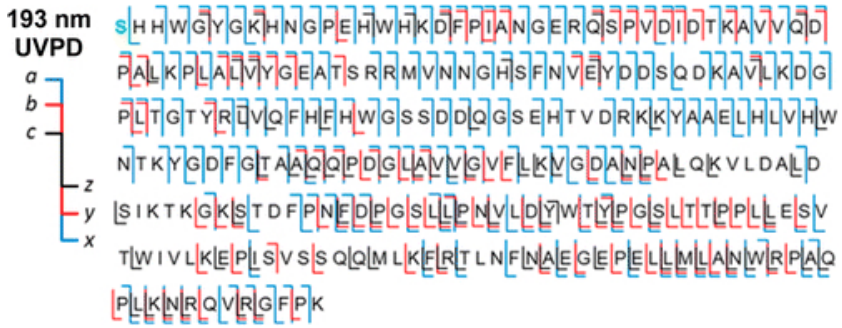


Figure 1.16: Example fragmentation pattern of carbonic anhydrase utilizing UVPD with 193 nm photons.This figure is reprinted with permission from reference [95].

currently. UVPD (193 nm) has been compared to HCD for high-throughput top-down proteomics in a recent study with UVPD resulting in better average proteoform sequence coverage compared to HCD [98].

Compared to BUP, sufficient fragmentation for identification of proteoforms and localization of modifications can be difficult for TDP. With large molecules such as proteins, there are many vibrational modes available for energy redistribution, making the number of collisions required for fragmentation in the hundreds when using collision-based methods, such as collision-induced dissociation (CID) or HCD. Also, spectra collected in TDP, even for relatively well-separated proteoforms, experiments are notoriously complex, Figure 1.17, making interpretation of data from mass analysis difficult, especially for low abundance proteoforms.

1.5.7 Proteoform identification and quantification

Difficulty identifying proteoforms is furthered by the signal of gas phase proteins generated by ESI being distributed over multiple charge states and isotopic peaks, Figure 1.17. Most

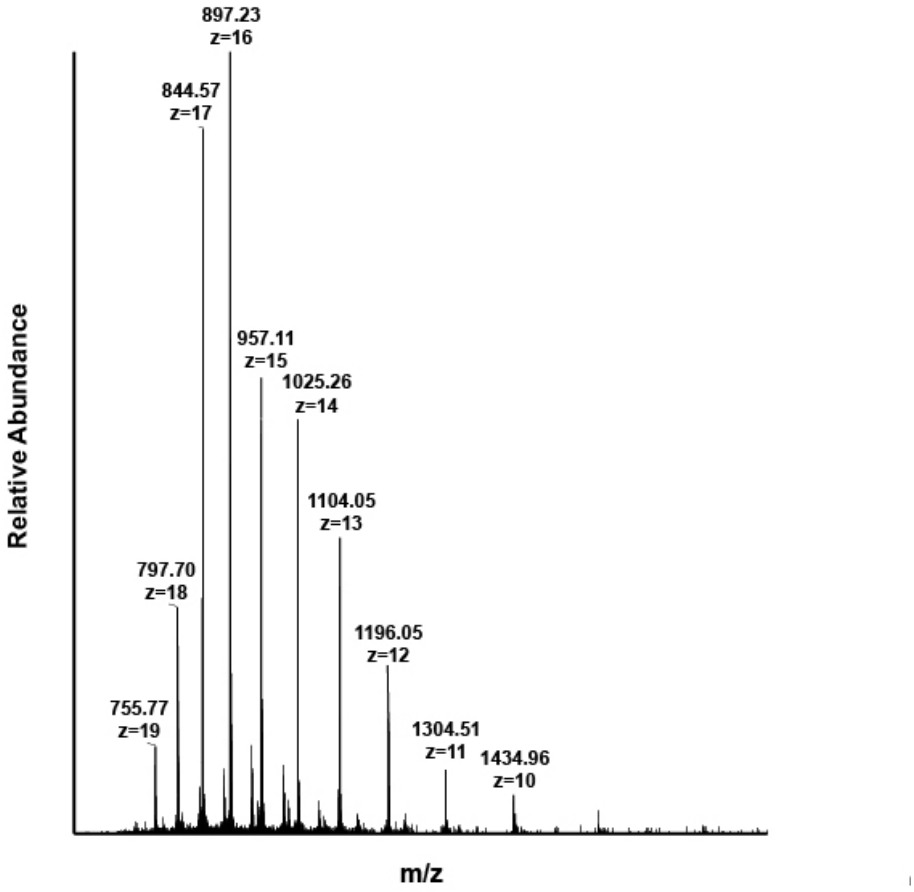


Figure 1.17: An example mass spectrum of an intact protein in a complex sample.

TDP mass spectrometry workflows only select for the top abundant ions in a particular spectrum, usually the top 5 or less most abundant ions, for fragmentation. For a spectrum like the one shown in Figure 1.17, it is highly unlikely that any ions from the low abundant species also present in that spectrum will be chosen for fragmentation and, therefore, will not be identified. Proteoforms containing combinations of or lesser known modifications can also be difficult to identify, as most TDP software allows for only a few unknown mass shifts and a non-exhaustive list of known modifications along the proteoform sequences. Proteoforms containing combinations of unusual mass shifts are, therefore, rarely identified and require manual examination of raw data.

Newer and better software for database searching is constantly being developed to aid in

proteoform identification in TDP [99–104]. One of the most common methods for protein and peptide identification is the target-decoy approach [105, 106]. A basic workflow for this approach begins with deconvolution of MS and MS/MS spectra to get monoisotopic masses of intact parent ions and fragments. This experimental data is then matched to a theoretical protein database, inferred from the genome for the sample in question, to obtain best scoring identifications. Scoring systems depend on which software is being used and therefore, specific numbers will not be provided here. Experimental data is then matched to a reverse (decoy) database for best scoring identifications. One identification is produced for each spectrum (best scoring match from either target or decoy sequences) that scores above a certain score, x. A false discovery rate (FDR) is then used to filter data, Equation 1.8 [107–109].

$$FDR \approx \frac{N_{decoy}(x)}{N_{target}(x)} \tag{1.8}$$

N in this case would stand for the number of identified spectra from the decoy database search (N_{decoy}) and target database search (N_{target}). Several other methods exist for estimating significance of proteoform identifications, and details can be found elsewhere [100, 110, 111].

There are methods that have been developed or are currently being developed to increase proteoform identification confidence, including using BUP data and electrophoretic mobility μ_{ef} predictions [101, 112]. BUP is extremely useful as its methods are much more developed than TDP, and integration of BUP and TDP datasets allows for the use of shared sequence and PTM information for higher confidence and possibly more proteoform identifications [101]. Electrophoretic mobility predictions for peptides in large-scale studies has been relatively successful compared to predictions for retention time during LC, as the size and charge of peptides is relatively easy to calculate [112, 113]. However, for large proteoforms, this is more complex, as is pointed out and addressed in the recent work by Chen, et al [112].

A classification scheme has also been developed at the proteoform-level for addressing the issue of ambiguous proteoform identifications in TDP studies [114]. Importantly, this classification scheme distinguishes between localizing and identifying PTMs, as PTM localization is much more difficult for TDP than BUP because of the relative ease of fragmenting smaller peptides. Following proteoform identification after the database search, various annotation analyses and quantification can be performed that highlight the possible biological impact of proteomics [1, 2, 9].

Limitations associated with the sensitivity and dynamic range of the mass spectrometer, coelution of proteoforms of wide abundance differences, and differences in ionization efficiency of intact proteoforms not only makes identification of a large number of proteoforms difficult but also makes quantitative analysis of these proteoforms extremely difficult [40, 71, 72]. Software tools for quantification in TDP are constantly being developed [103, 115–119]. Label-free formats for quantitation are attractive as they require no additional sample preparation and do not rely on reproducible labeling or chromatographic retention time [120]. Label-free methods widely include "spectral counting" methods in BUP as a rough estimate of true quantitation [22, 120, 121]. It functions off of the assumption that precursor ion selection for fragmentation of higher abundant precursors occurs more often with repetition resulting in higher likelihood of a successful identification [120]. Application to TDP in recent studies has demonstrated that spectral counting can approach other methods of quantitation (e.g. peak areas and intensities) while requiring much less data processing, even though it depends greatly on the speed at which MS/MS spectra can be gathered [22, 120, 122]. However, for the future of TDP in precision medicine, more accurate and precise isotopic labeling methods and label-free methods, such as peak areas and intensities, need to be utilized.

BUP often utilizes multiplexing of samples to aid in throughput and quantification by labeling with isobaric chemical tags. Isobaric tagging for relative and absolute

quantification (iTRAQ) and tandem mass tag (TMT) are the most common methods used in BUP. Stable isotope labeling by amino acids in cell culture (SILAC), neutron coding (NeuCode) SILAC, and protein-level TMT labeling are methods that have been utilized in TDP but have been difficult to develop, Figure 1.18 [119, 123–125]. Sequential window

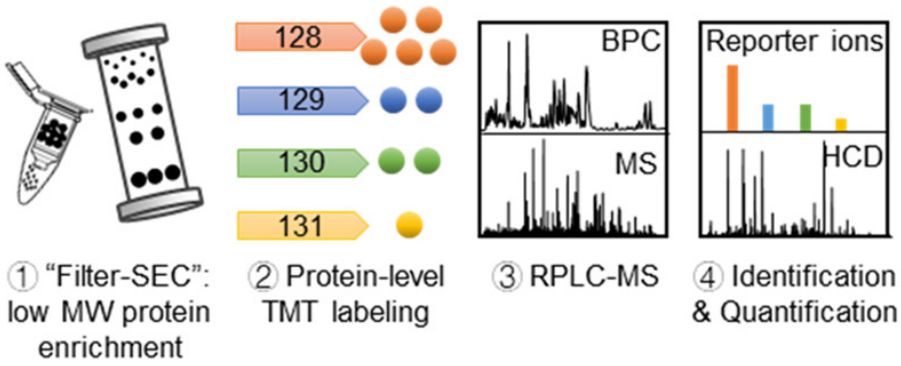


Figure 1.18: TMT workflow in TDP. This figure is reprinted with permission from reference [119].

acquisition of all theoretical mass spectra (SWATH-MS) is a data-independent acquisition method is a method in BUP that is growing in popularity for increasing throughput and reproducibility in proteomics experiments [126, 127]. Simply, all ions in a sample that are within a certain mass range are fragmented instead of choosing more specific m/z values for fragmentation as is the case in data-dependent acquisition (DDA) methods [126]. The recent development of scanning SWATH by Messner, et al. has significantly increased throughput in BUP experiments compared to traditional SWATH [127]. Although DDA methods can be used in TDP experiments, the complicated nature of tandem mass spectra makes software development the most significant bottleneck for regular use [128].

MASH Suite Pro is an example of label-free software that has been developed for proteoform identification, quantification, and characterization of proteoforms, and uses intensity data from MS, not MS/MS, for quantification [40, 129, 130]. TopPIC Suite, developed by Xiaowen Liu at IUPUI, is another example of software for proteoform identification, quantification, and characterization. For quantification, proteoform

isotopomer envelopes are combined across different migration times and charge states, and the sum of the intensities across these peaks are used for proteoform quantification [35].

As is implied earlier, better proteoform separations can aid in proteoform quantification, easing some of the analytical stress on the mass spectrometer. But, ultimately, the better the separations and the faster the mass analysis, the better the quantification. High resolution is important for determining charge and therefore mass to identify proteoforms, however, the accurate mass and speed of analyzers, such as TOF instruments, may provide an important step forward.

1.6 Summary

Due to the various issues associated with TDP, TDP studies have been mainly limited to targeted studies of specific proteoforms. Targeted studies are extremely important for the further development of possible drug targets or treatment regimen for specific diseases, but the initial association of different proteoforms from a certain gene to a specific phenotype can be lost [75]. This includes proteoforms that could be more strongly associated with that phenotype than the whole protein concentration [75]. To take advantage of the information that TDP can provide about these proteoforms, there needs to be significant advancement in almost all aspects of the TDP workflow.

The rest of this dissertation will address portions of the TDP workflow that need improvement and a variety of applications, utilizing CZE-ESI-MS-MS as the final dimension of separation and mass analysis, whose aims are to develop a unique and effective platform for deep TDP. Ultimately, coupling LC to CZE, combined with aforementioned fragmentation techniques will provide a significant step forward for TDP in terms of the numbers of proteoform identifications in large-scale studies and characterization of proteoforms.

BIBLIOGRAPHY

BIBLIOGRAPHY

(1) Lucitt, M. B.; Price, T. S.; Pizzaro, A.; Wu, W.; Yocum, A. K.; Seiler, C.; Pack, M. A.; Fitzgerald, G. A.; Grosser, T. **2008**, *7*, 981–994.

(2) Cravatt, B. F.; Simon, G. M.; III, J. R. Y. *Nature* **2007**, *9*, 991–1000.

(3) Toby, T. K.; Fornelli, L.; Kelleher, N. L. *Annu. Rev. Anal. Chem.* **2016**, *9*, 499–519.

(4) Smith, L. M.; Kelleher, N. L. *Nat. Methods* **2013**, *10*, 186–187.

(5) Wang, X.; Slebos, R. J. C.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B. *J. Proteome Res.* **2012**, *11*, 1009–1017.

(6) Evans, V. C.; Barker, G.; Heesom, K. J.; Fan, J.; Bessant, C.; Matthews, D. A. *Nat. Methods* **2012**, *9*, 1207–1211.

(7) Smith, L. M.; Kelleher, N. L. *Science* **2018**, *359*, 1106–1107.

(8) Aebersold, R. et al. *Nat. Chem. Biol.* **2018**, *14*, 206–214.

(9) Li, C.; Tan, X. F.; Lim, T. K.; Lin, Q.; Gong, Z. *Sci. Rep.* **2016**, *6*, 24329.

(10) Zubarev, R. *Proteomics* **2013**, *13*, 723–726.

(11) Tran, J. C. et al. *Nature* **2011**, *480*, 254–258.

(12) Rodriguez, H.; Zenklusen, J. C.; Staudt, L. M.; Doroshow, J. H.; Lowry, D. R. *Cell* **2021**, *184*, 1661–1670.

(13) Yartseva, V.; Giraldez, A. J. *Curr. Top. Dev. Biol.* **2015**, *113*, 191–232.

(14) Gross, J. H., Mass Spectrometry, 3rd ed.; Springer International Publishing AG: 2017.

(15) Abonnenc, M.; Qiao, L.; Liu, B.; Girault, H. H. *Annu. Rev. Anal. Chem.* **2010**, *3*, 231–254.

(16) Yamashita, M.; Fenn, J. B. *J. Phys. Chem. A* **1984**, *88*, 4451–4459.

(17) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64–71.

(18) Zubarev, R. A.; Makarov, A. *Anal. Chem.* **2013**, *85*, 5288–5296.

(19) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998**, *17*, 1–35.

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

(21) Bowman, A. P.; Blakney, G. T.; Hendrickson, C. L.; Ellis, S. R.; Heeren, R. M. A.; Smith, D. F. *Anal. Chem.* **2020**, *92*, 3133–3142.

(22) Mann, M.; Kelleher, N. L. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18132–18138.

(23) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. *Anal. Chem.* **2018**, *90*, 110–127.

(24) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. *Biochem. Biophys. Res. Commun.* **2014**, *445*, 683–693.

(25) Makarov, A. *Anal. Chem.* **2000**, *72*, 1156–1162.

(26) VanAernum, Z. L.; Gilbert, J. D.; Belov, M. E.; Makarov, A. A.; Horning, S. R.; Wysocki, V. H. *Anal. Chem.* **2019**, *91*, 3611–3618.

(27) Shen, X.; Xu, T.; Hakkila, B.; Hare, M.; Wang, Q.; Wang, Q.; Beckman, J. S.; Sun, L. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 1361–1369.

(28) Cotter, R. J. *Anal. Chem.* **1992**, *64*, 1027A–1039A.

(29) Mamyrin, B. A. *Int. J. Mass Spectrom.* **2001**, *206*, 251–266.

(30) Gillet, L. C.; Leitner, A.; Aebersold, R. *Annu. Rev. Anal. Chem.* **2016**, *9*, 449–472.

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

(32) Nesvizhskii, A. I.; Aebersold, R. *Mol. Cell. Proteom.* **2005**, *4*, 1419–1440.

(33) Chen, D.; McCool, E. N.; Yang, Z.; Shen, X.; Lubeckyj, R. A.; Xu, T.; Wang, Q.; Sun, L. *Mass Spectrom. Rev.* **2021**, *Just Accepted Manuscript*.

(34) Siuti, N.; Kelleher, N. L. *Nat. Methods* **2007**, *4*, 817–821.

(35) Lubeckyj, R. A.; Basharat, A. R.; Shen, X.; Liu, X.; Sun, L. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 1435–1445.

(36) McCool, E. N.; Sun, L. *Se Pu* **2019**, *37*, 878–886.

(37) Chen, D.; Yang, Z.; Shen, X.; Sun, L. *Anal. Chem.* **2021**, *93*, 4417–4424.

(38) Yang, Z.; Shen, X.; Chen, D.; Sun, L. *Anal. Chem.* **2020**, *19*, 3315–3325.

(39) Wiśniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. *Nat. Methods* **2009**, *6*, 359–362.

(40) Melby, J. A.; Roberts, D. S.; Larson, E. J.; Brown, K. A.; Bayne, E. F.; Jin, S.; Ge, Y. *J. Am. Mass Spectrom.* **2021**, *Just Accepted Manuscript*.

(41) Shieh, I. F.; Lee, C. Y.; Shiea, J. *J. Proteome Res.* **2005**, *4*, 606–612.

(42) Botelho, D.; Wall, M. J.; D. B, . V.; Fitzsimmons, S.; Liu, F.; Doucette, A. *J. Proteome Res.* **2010**, *9*, 2863–2870.

(43) Brown, K. A.; Chen, B.; Guardado-Alvarez, T. M.; Lin, Z.; Hwang, L.; Ayaz-Guner, S.; Jin, S.; Ge, Y. *Nat. Methods* **2019**, *16*, 417–420.

(44) Bauer, A.; Kuster, B. *Eur. J. Biochem.* **2003**, *270*, 570–578.

(45) Cai, W.; Tucholski, T. M.; Gregorich, Z. R.; Ge, Y. *Expert Rev. Proteomics* **2016**, *13*, 717–730.

(46) Skoog, D. A.; Holler, F. J.; Crouch, S. R., Principles of Instrumental Analysis, 6th ed.; Cengage Learning: 2007.

(47) Neue, U. D. *J. Chromatogr. A* **2005**, *1079*, 153–161.

(48) Jorgenson, J. W. *Annu. Rev. Anal. Chem.* **2010**, *3*, 129–150.

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

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

(51) Sun, L.; Knierman, M. D.; Zhu, G.; Dovichi, N. J. *Anal. Chem.* **2013**, *85*, 5989–5995.

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

(53) Zhao, Y.; Sun, L.; Zhu, G.; Dovichi, N. J. *J. Proteome Res.* **2016**, *15*, 3679–3685.

(54) Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; III, J. R. Y. *J. Proteome Res.* **2014**, *13*, 6078–6086.

(55) Haselberg, R.; de Jong, G. J.; Somsen, G. W. *Anal. Chem.* **2013**, *85*, 2289–2296.

(56) Jorgenson, J. W.; Lukacs, K. D. *Science* **1983**, *222*, 266–272.

(57) Zhao, Y.; Sun, L.; Champion, M. M.; Knierman, M. D.; Dovichi, N. J. *Anal. Chem.* **2014**, *86*, 4873–4878.

(58) Zhu, G.; Sun, L.; Dovichi, N. J. *Talanta* **2016**, *146*, 839–843.

(59) Makham, M.; Vakhshouri, L. *Int. J. Mol. Sci.* **2010**, *11*, 1546–1556.

(60) Hjertén, S. J. *J. Chromatogr.* **1985**, *347*, 191–198.

(61) Gao, L.; Liu, S. *Anal. Chem.* **2004**, *76*, 7179–7186.

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

(63) Imami, K.; Monton, M. R.; Ishihama, Y.; Terabe, S. *J. Chromatogr. A* **2007**, *1148*, 250–255.

(64) Wang, L.; MacDonald, D.; Huang, X.; Chen, D. D. *Electrophoresis* **2016**, *37*, 1143–1150.

(65) Sun, L.; Hebert, A. S.; Yan, X.; Zhao, Y.; Westphall, M. S.; Rush, M. J.; Zhu, G.; Champion, M. M.; Coon, J. J.; Dovichi, N. J. *Angew Chem. Int. Ed.* **2014**, *53*, 13931–13933.

(66) Jr., S. L. S.; Quirino, J. P.; Terabe, S. *J. Chromatogr. A* **2008**, *1184*, 504–541.

(67) Altria, K. D.; Elder, D. *J. Chromatogr. A* **2004**, *1023*, 1–14.

(68) Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2554–2560.

(69) Maxwell, E. J.; Chen, D. D. Y. *Anal. Chim. Acta* **2008**, *627*, 25–33.

(70) Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J. *J. Proteome Res.* **2015**, *14*, 2312–2321.

(71) Smith, R. D.; Barinaga, C. J.; Udseth, H. R. *Anal. Chem.* **1988**, *60*, 1948–1952.

(72) Smith, R. D.; Loo, J. A.; Loo, R. R. O.; Busman, M.; Udseth, H. R. *Mass Spectrom. Rev.* **1991**, *10*, 359–451.

(73) Scheltema, R. A.; Hauschild, J.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M. *Mol. Cell. Proteomics* **2014**, *13*, 3698–3708.

(74) Slavov, N. *Nat. Biotechnol.* **2021**, *39*, 809–810.

(75) Seckler, H. D. S.; Park, H.; Lloyd-Jones, C. M.; Melani, R. D.; Camarillo, J. M.; Wilkins, J. T.; Compton, P. D.; Kelleher, N. L. *J. Am. Mass Spectrom.* **2021**, *32*, 1659–1670.

(76) Faserl, K.; Sarg, B.; Sola, L.; Linder, H. H. *Proteomics* **2017**, *17*, doi.org/10.1002/pmic.201700310.

(77) Boley, D. A.; Zhang, Z.; Dovichi, N. J. *J. Chromatogr. A* **2017**, *1523*, 123–126.

(78) Garza, S.; Moini, M. *Anal. Chem.* **2006**, *78*, 7309–7316.

(79) Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W. *J. Am. Chem. Soc.* **1999**, *121*, 806–812.

(80) Ge, Y.; Lawhorn, B. G.; ElNaggar, M.; Strauss, E.; Park, J. H.; Begley, T. P.; McLafferty, F. W. *J. Am. Chem. Soc.* **2002**, *124*, 672–678.

(81) Wysocki, V. H.; Tsapralis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* **2000**, *35*, 1399–1406.

(82) Hunt, D. F.; III, J. R. Y.; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 6233–6237.

(83) Wysocki, V. H.; Kenttämaa, H. I.; Cooks, R. G. *Int. J. Mass Spectrom. Ion Processes* **1987**, *75*, 181–208.

(84) Biemann, K.; Martin, S. A. *Mass Spectrom. Rev.* **1987**, *6*, 1–76.

(85) Biemann, K. *Meth. Enzymol.* **1990**, *193*, 455–479.

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

(87) Coon, J. J.; Ueberheide, B.; Syka, J. E. P.; Dryhurst, D. D.; Ausio, J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 9463–9468.

(88) De Hoffman, E.; Stroobant, V., Mass Spectrometry: Principles and Applications, 3rd ed.; Wiley: New York: 2007.

(89) Harrison, A. G. *Mass Spectrom. Rev.* **2009**, *28*, 640–654.

(90) Dongre, A. R.; Jones, J. L.; Somogyi, Á.; Wysocki, V. H. *J. Am. Chem. Soc.* **1996**, *118*, 8365–8374.

(91) Rush, M. J. P.; Riley, N. M.; Westphall, M. S.; Coon, J. J. *Anal. Chem.* **2018**, *90*, 8946–8953.

(92) Riley, N. M.; Westphall, M. S.; Coon, J. J. *J. Proteome Res.* **2017**, *16*, 2653–2659.

(93) Riley, N. M.; Westphall, M. S.; Coon, J. J. *J. Am. Soc. Mass Spectrom.* **2018**, *29*, 140–149.

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

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

(96) McCool, E. N.; Chen, D.; Li, W.; Liu, Y.; Sun, L. *Anal. Methods* **2019**, *11*, 2855–2861.

(97) Fornelli, L. et al. *Mol. Cell. Proteom.* **2020**, *19*, 405–420.

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

(99) Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. *Mol. Cell. Proteom.* **2010**, *9*, 2772–2782.

(100) Kou, Q.; Wang, Z.; Lubeckyj, R. A.; Wu, S.; Sun, L.; Liu, X. *J. Proteome Res.* **2019**, *18*, 878–889.

(101) Schaffer, L. V.; Millikin, R. J.; Shortreed, M. R.; Scalf, M.; Smith, L. M. *J. Proteome Res.* **2020**, *19*, 3510–3517.

(102) Schaffer, L. V.; Shortreed, M. R.; Cesnik, A. J.; Frey, B. L.; Solntsev, S. K.; Scalf, M.; Smith, L. M. *Anal. Chem.* **2018**, *90*, 1325–1333.

(103) Durbin, K. R.; Fornelli, L.; Fellers, R. T.; Doubleday, P. F.; Narita, M.; Kelleher, N. L. *J. Proteome Res.* **2016**, *15*, 976–982.

(104) Chen, W.; Liu, X. *J. Proteome Res.* **2021**, *20*, 261–269.

(105) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207–214.

(106) Elias, J. E.; Gygi, S. P., Proteome Bioinformatics Methods in Molecular *Biology*^{TM} (Methods and Protocols); Hubbard, S., Jones, A., Eds.; Humana Press: 2010; Vol. 604; Chapter Target-Decoy Search Strategy for Mass Spectrometry-based Proteomics.

(107) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *J. Proteome Res.* **2008**, *7*, 40–44.

(108) Tang, W. H.; Shilov, I. V.; Seymour, S. L. *J. Proteome Res.* **2008**, *7*, 3661–3667.

(109) Kou, Q.; Xun, L.; Liu, X. *Bioinformatics* **2016**, *32*, 3495–3497.

(110) Kou, Q.; Wu, S.; Tolić, N.; Paša-Tolić, L.; Liu, Y.; Liu, X. *Bioinformatics* **2017**, *33*, 1309–1316.

(111) Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. *Nucleic Acids Res.* **2007**, *35*, W701–W706.

(112) Chen, D.; Lubeckyj, R. A.; Yang, Z.; McCool, E. N.; Shen, X.; Wang, Q.; Xu, T.; Sun, L. *Anal. Chem.* **2020**, *92*, 3503–3507.

(113) Krokhin, O. V.; Anderson, G.; Spicer, V.; Sun, L.; Dovichi, N. J. *Anal. Chem.* **2017**, *89*, 2000–2008.

(114) Smith, L. M. et al. *Nat. Methods* **2019**, *16*, 939–940.

(115) Schaffer, L. V. et al. *Proteomics* **2019**, *19*, 1800361.

(116) Millikin, R. J.; Shortreed, M. R.; Scalf, M.; Smith, L. M. *J. Proteome Res.* **2020**, *19*, 1975–1981.

(117) Mehta, S. et al. *Proteomes* **2020**, *8*, 15.

(118) Wu, Z. et al. *J. Proteome Res.* **2020**, *19*, 3867–3876.

(119) Yu, D.; Wang, Z.; Cupp-Sutton, K. A.; Guo, Y.; Kou, Q.; Smith, K.; Liu, X.; Wu, S. *J. Am. Mass Spectrom.* **2021**, *32*, 1336–1344.

(120) Geis-Asteggiante, L.; Ostrand-Rosenberg, S.; Genselau, C.; Edwards, N. J. *Anal. Chem.* **2016**, *88*, 10900–10907.

(121) Liu, H.; Sadygov, R. G.; III, J. R. Y. *Anal. Chem.* **2004**, *76*, 4193–4201.

(122) Domon, B.; Aebersold, R. *Science* **2006**, *312*, 212–217.

(123) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. *Nat. Methods* **2013**, *10*, 332–334.

(124) Rhoads, T. W. et al. *Anal. Chem.* **2014**, *86*, 2314–2319.

(125) Collier, T. S.; Hawkridge, A. M.; Georgianna, D. R.; Payne, G. A.; Muddiman, D. C. *Anal. Chem.* **2008**, *80*, 4994–5001.

(126) Ludwig, C.; Gillet, L.; Rosenberger, G.; Amon, S.; Collins, B. C.; Aebersold, R. *Mol. Syst. Biol.* **2018**, *14*, e8126.

(127) Messner, C. B. et al. *Nat. Biotechnol.* **2021**, *39*, 846–854.

(128) Cupp-Sutton, K. A.; Wu, S. *Mol. Omics* **2020**, *16*, 91–99.

(129) Guner, H.; Close, P. L.; Cai, W.; Zhang, H.; Peng, Y.; Gregorich, Z. R.; Ge, Y. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 464–470.

(130) Cai, W.; Guner, H.; Gregorich, Z. R.; Chen, A. J.; Ayaz-Guner, S.; Peng, Y.; Valeja, S. G.; Liu, X.; Ge, Y. *Mol. Cell. Proteomics* **2016**, *15*, 703–714.

Chapter 2

Optimization of capillary zone electrophoresis mass spectrometry for top-down proteomics

2.1 Introduction

Efficient separations coupled to MS are a vital aspect to any proteomics study characterizing complex biological mixtures [1, 2]. [1] Highly efficient and high capacity separations are especially difficult to achieve when characterizing intact proteoforms, as in TDP studies [5–10]. A large variety of separations have been utilized in TDP studies with trade-offs between separation efficiency, protein solubility, and compatibility with MS. The most common methods for separating intact proteins include separations based on hydrophobicity, such as RPLC, charge-based separations including CZE, and isoelectric focusing (IEF), and size-based separations including size exclusion chromatography (SEC), gel-electrophoresis, and GELFrEE [3, 7, 8, 11–13].

CZE-ESI-MS has been well recognized for characterization of intact proteins due to its high separation efficiency [14–17]. CZE-ESI-MS/MS has been suggested as an alternative to widely used RPLC-ESI-MS/MS for TDP. [5, 15, 18–26].

[1] This chapter was adapted with permission from references [3, 4]

CZE-MS/MS has been evaluated for top-down characterization of intact proteins for over 20 years ago. In 1996, Valaskovic et al. developed a CZE-ESI-MS/MS platform for characterization of attomole amounts of intact proteins, and identified carbonic anhydrase in crude extract of human red blood cells by sequence-specific fragment ions [18]. However, the CZE-MS interface used in that work had limited lifetime and robustness, which impeded the wide application of the platform for TDP. An electrokinetically pumped sheath flow CE-MS interface with good sensitivity and robustness was developed by Dovichi group in 2010 [27]. Sun et al. demonstrated fast, reproducible and sensitive characterization of intact proteins with the electrokinetically pumped sheath flow interface based CZE-MS/MS [19]. Later, Zhao et al. further applied the CZE-MS/MS system for TDP of *Mycobacterium marinum* secretome and yeast proteome [20, 21]. Coupling offline RPLC fractionation to CZE-MS/MS identified 580 proteoforms from a yeast lysate. In total, 23 RPLC fractions were analyzed by CZE-MS/MS and up to 180 proteoforms could be identified with single-shot CZE-MS/MS [21]. Li et al. developed a CZE-MS system based on the electrokinetically pumped sheath flow interface and applied the system to a complex proteome sample for characterization of large proteins, resulting in identification of 30 proteins in the mass range of 30-80 kDa [22].

A sheathless CE-MS interface using a porous tip for ESI was developed by the Moini group in 2007 and showed great sensitivity and robustness [28]. Han et al. employed the sheathless interface-based CZE-MS/MS for TDP of a *Pyrococcus furiosus* lysate, resulting in identification of 291 proteoforms with RPLC fractionation and CZE-MS/MS [23]. Han et al. also characterized the Dam1 protein complex using the sheathless interface-based CZE-MS. Their results showed that CZE-MS approached complete characterization of the protein complex with 100-times less sample consumption compared to RPLC-MS [5]. Sensitive and comprehensive characterization of intact pharmaceutical proteins via the sheathless interface based CZE-MS has been demonstrated recently, thus leading to detection of over 250 different isoforms of recombinant human erythropoietin and 138

proteoforms from recombinant human interferon-β1 [15, 24]. The sheathless interface based CZE-MS has also been applied for characterization of intact histones by the Lindner group [25, 26].

The current CZE-MS interfaces are robust and sensitive, enabling CZE-MS/MS to be used for TDP. However, two issues remain for CZE-MS/MS-based TDP. First, the largest sample loading capacity of CZE-MS/MS systems reported in the literature for TDP is only about 200 nL [21, 23]. The low sample loading capacity impedes identification of low abundant proteoforms from complex proteome samples. Second, the reported separation window of CZE-MS/MS systems for TDP is roughly 30 min [21, 23]. The narrow separation window limits the number of MS/MS spectra acquired during one experiment, which restricts the number of proteoform identifications (IDs) from CZE-MS/MS. Capillary isoelectric focusing (cIEF)-MS is a promising technique for large-scale TDP due to its large sample loading capacity and high resolution for separation of intact proteins. The Smith group evaluated cIEF-MS for top-down characterization of complex proteomes over one decade ago [29, 30]. However, coupling cIEF to MS is still not straightforward, which hinders its wide application for TDP.

In order to improve the sample loading capacity and separation window of CZE-MS, our group recently systematically evaluated a dynamic pH junction-based CZE-MS/MS system for BUP. We observed a 140-min separation window and a μL-scale sample loading capacity using the CZE-MS/MS system for analysis of complex proteome digests [31]. Dynamic pH junction is a simple method for sample stacking in CZE [32, 33]. For instance, sample is dissolved in a basic buffer (e.g., ammonium bicarbonate, pH 8) and the BGE is acidic (e.g., 0.1% (v/v) formic acid, pH 2.8). The capillary is first filled with BGE, and then a long plug of sample is injected into the separation capillary via applying pressure. After that, both ends of the separation capillary are immersed in the BGE vials. Two pH boundaries exist at the two junctions of the BGE and the sample, one at the injection end

(pH boundary I) and the other one inside the capillary (pH boundary II). When a positive high voltage is applied at the injection end of the separation capillary, the hydrogen positive ions in the BGE vial will migrate into the capillary and titrate the sample zone, which makes the pH boundary I slowly move toward the pH boundary II. In the meantime, the negatively charged analytes in the sample zone migrate toward the injection end of the capillary, and they are focused at the moving pH boundary I [34–37]. After those two pH boundaries meet, isotachophoresis (ITP) plays a role for stacking the analytes with NH4+ as the leading ion, followed by the typical CZE [34].

Although dynamic pH junction has been widely used for concentration of small molecules and peptides, it has not been thoroughly investigated for concentration of intact proteins for TDP. To our best knowledge, there is only one published paper in the literature about using dynamic pH junction-based CZE-MS/MS for large-scale TDP. Zhao et al. performed TDP of a yeast lysate using dynamic pH junction-based CZE-MS/MS [21]. They used 5 mM ammonium bicarbonate (pH 8) as the sample buffer and 5% (v/v) acetic acid (pH 2.4) as the BGE. 100-240 nL of the sample was injected for CZE-MS/MS analysis. The separation window and peak capacity of the dynamic pH junction-based CZE-MS/MS system was roughly 30 min and less than 100, respectively [21]. In this work, for the first time dynamic pH junction-based CZE-MS was systematically evaluated for concentration and separation of proteins. We applied the optimized CZE-MS/MS system for TDP of *E. coli*, thus leading to μL-scale loading capacity, 90-min separation window, high peak capacity ($\sim$ 280), and nearly 600 proteoform IDs with single-shot CZE-MS/MS.

Advancements in RPLC have recently been made with the use of shorter bonded stationary phases (C4 and shorter), a variety of particle types and column packing procedures, the use of monoliths, and longer columns [8, 38, 39]. The main advantages of RPLC over CZE is the loading capacity, which is consistently in the μg-range, and control over the separation window which is the window of time in which proteoforms are coming

out of the capillary [3, 34]. This is an important point when considering the large dynamic range of protein concentrations in the proteome, which can approach 7 orders of magnitude [40, 41]. Also, it is estimated that ~ 1 million unique proteoforms exist in the human proteome [42]. Shen, et al. demonstrated that high peak capacity ($P_c \sim 400$) was possible for intact protein separation with an optimized long-column RPLC system, which, when coupled to MS, resulted in the identification of ~ 900 proteoforms from an *S. oneidensis* lysate in a single run [8]. Specifically, long columns (120 cm long x 100 μm i.d.) were used with superficially porous particles (3.6 μm, 200 Å pores) in that study. The loading amount was 2.5 μg, and the separation window approached 800 minutes while still retaining an efficient separation.

Without the use of particles and with a complementary separation mechanism, CZE can offer additional insight into complex proteomes [15, 19, 21, 43]. The lack of particles is especially important for the separation of intact proteins to limit zone broadening and sample loss [16]. Also, with current interfaces, CZE-MS outperforms LC-MS platforms in terms of sensitivity [5]. An improved CZE-MS platform for the separation and identification of proteoforms has resulted in ~ 600 proteoform and 200 protein identifications in a single run of an *E. coli* sample [3]. This CZE-MS platform, utilizing dynamic pH junction-based sample stacking, increased the loading capacity to $\sim 1\mu$g and the separation window to 90 minutes [3, 34, 36, 44, 45]. Combining offline SEC-RPLC fractionation to this optimized CZE-MS platform resulted in the identification of 5705 proteoforms and 850 proteins from an *E. coli* lysate [46]. Fractionating before CZE-MS analysis alleviates the inherent limitations of CZE and using RPLC combines the complementary nature of these two techniques to reach unrivaled proteome coverage. Li et al. showed that, on the protein level, the overlap was less than 35% using a C18 column (100 μm i.d. x 100 mm long, 1.7 μm particles) for RPLC separation and an uncoated capillary (30 cm in length) for CZE separation under high voltage (5.5 kV) [43].

Herein, we offer the first optimization of CZE parameters for increased loading capacity and separation window using dynamic pH junction-based sample stacking and the first direct comparisons of RPLC-MS and CZE-MS for top-down MS characterization of a standard protein mixture and an *E. coli* proteome sample [3, 4].

2.2 Experimental

2.2.1 Materials and reagents

Acrylamide was purchased from Acros Organics (NJ, USA). Standard proteins, ammonium bicarbonate (NH4HCO3), urea, dithiothreitol (DTT), iodoacetamide (IAA) and 3-(trimethoxysilyl)propyl methacrylate were purchased from Sigma-Aldrich (St. Louis, MO). LC/MS grade water, acetonitrile (ACN), methanol, formic acid (FA and HPLC-grade acetic acid (AA) were purchased from Fisher Scientific (Pittsburgh, PA). Aqueous mixtures were filtered with Nalgene Rapid-Flow Filter units (Thermo Scientific) with 0.2 μm CN membrane and 50 mm diameter. Fused silica capillaries (50 μm i.d./360 μm o.d.) were obtained from Polymicro Technologies (Phoenix, AZ). Hydrofluoric acid (HF, 48-51% solution in water) and acrylamide were purchased from Acros Organics (NJ, USA). Complete, mini protease inhibitor cocktail and PhosSTOP (provided in EASYpacks) were bought from Roche (Indianapolis, IN). Capillary columns for RPLC were bought from CoAnn Technologies, LLC (Richland, WA).

2.2.2 Sample preparation

E. coli (strain K-12 substrain MG1655) was cultured in LB medium at 37°C with 225 rpm shaking until OD600 reached 0.7. *E.coli* cells were harvested by centrifuge at 4,000 rpm for 10 min. Then the *E.coli* cells were washed with PBS three times. The *E.coli* cells were then lysed in a lysis buffer containing 8 M urea, 100 mM Tris-HCl (pH 8.0) and protease

inhibitors. The cell lysis was assisted by sonication with a Branson Sonifier 250 (VWR Scientific, Batavia, IL) on ice for 10 minutes. After centrifugation (18,000 x g for 10 min), the supernatant containing the extracted proteins was collected. A small aliquot of the extracted proteins was used for bicinchoninic acid (BCA) assay to determine the protein concentration. The leftover protein extracts were stored at -80°C before use. *E. coli* samples were prepared the same way for comparing RPLC and CZE except for the final solvent for analysis. The *E. coli* protein sample was redissolved in either MP A for RPLC-ESI-MS or in an NH4HCO3 buffer (50 mM, pH 8) for CZE-ESI-MS for a final protein concentration of $\sim$ 2 mg/mL.

E.coli proteins in 8 M urea and 100 mM Tris-HCl (pH 8.0) were denatured at 37°C, reduced with DTT and alkylated with IAA. Then, the proteins were desalted with a C4 trap column (Bio C4, 3 μm, 300 Å, 4.0 mm i.d., 10 mm long) from Sepax Technologies, Inc. (Newark, DE). A HPLC system (Agilent Technologies, 1260 Infinity II) was used. The HPLC eluate from the trap column was collected and further lyophilized with a vacuum concentrator (Thermo Fisher Scientific). The dried protein sample was redissolved in 50 mM NH4HCO3 (pH 8.0) to get about 2 mg/mL protein concentration (theoretical concentration based on 100% recovery from the whole sample preparation process) for CZE-MS/MS analysis.

For comparing RPLC and CZE, the stock standard protein mixture consisted of ubiquitin ($\sim$ 8.5 kDa, 0.1 mg/mL), cytochrome c (cyto.c, $\sim$ 12 kDa, 0.1 mg/mL), bovine serum albumin (BSA, $\sim$ 66 kDa, 1 mg/mL), myoglobin (myo, $\sim$ 17 kDa, 0.2 mg/mL), carbonic anhydrase (CA, $\sim$ 29 kDa, 0.2 mg/mL), and β-casein ($\sim$ 24 kDa, 0.4 mg/mL). Mixtures were dissolved in either mobile phase (MP) A, NH4HCO3 (50 mM) for RPLC or CZE, respectively. The stock mixture was diluted by a factor of 10 with MP A or NH4HCO3 (50 mM) before analysis.

2.2.3 CZE-ESI

An automated CZE-ESI-MS system was used in the experiments. The system contained an ECE-001 CE autosampler and a commercialized electrokinetically pumped sheath flow CE-MS interface from CMP Scientific (Brooklyn, NY) [27, 47]. The CE system was coupled to a LTQ-XL or a QEHF (Thermo Fisher Scientific). A fused silica capillary (50 μm i.d., 360 μm o.d., 1 meter long) was used for CZE separation. The inner wall of the capillary was coated with LPA based on previous studies [31, 48]. One end of the capillary was etched with HF acid based on reference to reduce the outer diameter of the capillary to $\sim$ 70-80 μm [49]. (Caution: use appropriate safety procedures while handling hydrofluoric acid solutions.) Different BGEs were used for CZE, including 5-10% (v/v) acetic acid and 0.1-0.5% (v/v) formic acid. The sheath buffer was 0.2% (v/v) formic acid containing 10% (v/v) methanol. Sample injection was carried out by applying pressure (5-10 psi) at the sample injection end and the injection periods were calculated based on the Poiseuille's law for different sample loading volume. High voltage (30 or 20 kV) was applied at the injection end of the separation capillary for separation and 2-2.2 kV was applied for ESI. At the end of each CZE-MS run, we flushed the capillary with BGE by applying 5 psi for 10 min. The ESI emitters were pulled from borosilicate glass capillaries (1.0 mm o.d., 0.75 mm i.d., and 10 cm length) with a Sutter P-1000 flaming/brown micropipet puller. The opening size of the ESI emitters was 20-40 μm.

Samples were injected into the capillary by applying pressure for a specified amount of time to achieve the necessary volume and sample amount based on Poiseuille's law. Loading volume was 200 nL for the 40 ng standard protein analysis and 500 nL for the 1 μg *E. coli* protein analysis. A separation voltage of 30 kV and 20 kV was used for the standard protein sample and *E. coli* sample, respectively. The capillary was flushed between runs with BGE with a pressure of 10 psi for 10 min.

2.2.4 RPLC-ESI

NanoRPLC was performed with an analytical column (C2, 90 cm long x 100 μm i.d., 3 μm beads, 300 Å pores) connected to an EASY nanoLC-1200 (Thermo Fisher Scientific). The RPLC system was connected directly to an ESI emitter with a metal union. A Spellman CZE1000R (Hauppage, NY) power supply was used to provide voltage for ESI through the metal union. A 240-min linear gradient from 100% MP A (10% ACN, 0.1% FA) to 75% MP B (70% ACN, 30% IPA, 0.1% FA) was used with a flow rate of 300 nL/min. The sample was loaded onto the column in MP A under a pressure of 800 bar. Blanks were performed between runs using a 120-min linear gradient (10% B-90% B) at 300 nL/min. Loading volumes were 0.5 μL for the 1 μg *E. coli* run and 4 μL for the 8 μg *E. coli* run.

2.2.5 MS and MS/MS

For all of the LTQ-XL experiments, only MS1 spectra were acquired using positive ion mode, and no protein fragmentation was performed. The scan range was m/z 600-2,000 using three microscans. The maximum injection time was 50 ms and the AGC target value was 3.0E4.

For all of the standard protein mixture experiments on the QEHF mass spectrometer, only MS1 spectra were acquired and no protein fragmentation was performed. "Intact protein mode" was used for all experiments with a trapping pressure of 0.2. The temperature of the ion transfer capillary was 320°C and the s-lens RF level was 55. Full MS scans were acquired with the number of microscans as three, the resolution as 240,000 (at m/z 200), the AGC target value as 1E6, the maximum injection time as 50 ms and the scan range as m/z 600-2000.

DDA methods were used for analysis of the *E.coli* sample on the QEHF mass spectrometer. The MS/MS spectra were acquired with the number of microscans as one, the resolution as 120,000 (at m/z 200), the AGC target value as 1E5 and the maximum

injection time as 200 ms. The three or eight most intense ions (Top 3 or Top 8 DDA) in the full MS spectrum were sequentially isolated with 4 m/z isolation window and further sequentially fragmented in the HCD fragmentation cell with NCE as 20%. The intensity threshold for triggering fragmentation was 1.0E5. Charge exclusion and exclude isotopes were turned on. Only protein ions with charge state higher than five can be isolated for fragmentation. The dynamic exclusion was turned on, and the setting was 30 s. The other parameters were the same as those used for the standard protein mixture experiments.

For comparing RPLC and CZE, a QEHF mass spectrometer was used for the experiments. The "intact protein mode" was turned on and a trapping pressure of 0.2 was used. The same MS and MS/MS settings were used for the RPLC and CZE experiments. Ion transfer capillary temperature was set to 320°C and the s-lens RF level was 55. For full MS, the number of microscans was 3, resolution was 120,000 (at m/z 200), AGC target value was 1E6, maximum injection time was 100 ms and the scan range was m/z 600-2000. For MS/MS, the number of microscans was 3, resolution was 120,000 (at m/z 200), AGC target value was 1E5, and maximum injection time was 200 ms. The top 5 most intense ions, for data dependent acquisition (Top 5 DDA), in full MS spectra were isolated with a 4 m/z window and sequentially fragmented at normalized collision energy (NCE) of 20%. The intensity threshold for triggering fragmentation was 1E5. Charge exclusion and exclude isotopes settings were turned on with proteins with charge state higher than 5 able to be fragmented. Dynamic exclusion was used with a setting of 30 s.

2.2.6 Measurement of electroosmotic flow

The protocol used here for measuring the EOF in the LPA coated capillary was based on previous works [50, 51]. Benzyl alcohol (neutral marker) was dissolved in the BGE, and used as the sample. The LPA coated capillary (50 μm i.d., 360 μm o.d., 1-meter-long) was flushed and filled with the BGE. First, the neutral marker (N1) was injected by applying 5 psi for t_{inj} (2s). Then, a plug of BGE was injected into the separation capillary by

applying 5 psi for time t_r (40s). After that, a second short plug of neutral marker (N2) was injected into the capillary for t_{inj} (2 s). Subsequently, another plug of BGE was injected into the capillary by applying pressure for t_r. The separation voltage (30 kV) was then applied at the injection end of the capillary for t_{mig} (50 min). During this period, the two neutral markers (N1 and N2) moved toward the cathode end with mobilities that were equal to the electroosmotic mobility (μ_{eof}). After t_{mig} has been completed, a third short plug of neutral marker (N3) was injected into the capillary for t_{inj} (2s). Finally, 5 psi was applied at the injection end of the capillary, which was immersed in the BGE, to push the three plugs of neutral marker out of the separation capillary, and the MS data acquisition was simultaneously started to record the signal of the neutral marker. The μ_{eof} was calculated by:

$$\mu_{eof} = \frac{[(t_{N3} - t_{N1}) - (t_{N2} - t_{N1})]L^2}{V_{separation} t_{mig}(t_{N3} + \frac{t_{inj}}{2})} \tag{2.1}$$

Where t_{N1}, t_{N2}, t_{N3} are the observed migration time for neutral marker N1, N2, and N3. L corresponds to the length of the capillary, and $V_{separation}$ is the separation voltage applied.

2.2.7 Data analysis

The standard protein data was analyzed using Xcalibur software (Thermo Fisher Scientific) to get intensity and migration time of proteins. The electropherograms were exported from Xcalibur and were further formatted using Adobe Illustrator to make the final figures.

All of the *E.coli* RAW files were analyzed with the TopFD (TOP-Down Mass Spectrometry Feature Detection) and TopPIC (TOP-Down Mass Spectrometry Based Proteoform Identification and Characterization) pipeline [52, 53]. TopFD is an improved version of MS-Deconv [54]. It converts precursor and fragment isotope clusters into monoisotopic masses and finds possible proteoform features in CZE-MS data by combining precursor isotope clusters with similar monoisotopic masses and close migration times (the isotopic clusters may have different charge states). The RAW files were first transferred

into mzXML files with Msconvert tool [55]. Then, spectral deconvolution was performed with TopFD to generate msalign files. Finally, TopPIC (version 1.1.3 and version 1.2.3) was used for database searching with msalign files as input. *E. coli* (strain K12) UniProt database was used for the first study (UP000000625, 4307 entries, version June 7, 2017) and second study (UP000000625, 4313 entries, version June 28, 2018). The spectrum-level false discovery rate (FDR) was estimated using the target-decoy approach [56, 57]. Cysteine carbamidomethylation was set as a fixed modification, and the maximum number of unexpected modifications was 2. The precursor and fragment mass error tolerances were 15 ppm. The maximum mass shift of unknown modifications was 500 Da, and the identified PrSMs were filtered with a 1% FDR at the spectrum level. For the second study, proteoform identifications were filtered with a 5% proteoform-level FDR. In order to reduce the redundancy of proteoform identifications, we considered the proteoforms identified by multiple spectra as one proteoform ID if those spectra correspond to the same proteoform feature reported by TopFD or those proteoforms are from the same protein and have similar precursor masses (within 1.2 Da).

2.3 Results and discussion

In order to approach large-scale TDP using CZE-MS/MS, the sample loading capacity and the separation window of CZE need to be improved for characterization of complex proteomes. We recently showed that dynamic pH junction based CZE-MS could reach μL-scale sample loading capacity and 140-min separation window simultaneously for analysis of complex peptide mixtures [31]. We speculated that the dynamic pH junction based CZE-MS should also work for characterization of complex mixtures of intact proteins. To test our speculation, we first investigated the performance of dynamic pH junction-based CZE-MS using a mixture of six standard proteins across a wide range of sample injection volumes (50-500 nL). The dynamic pH junction method was compared

with the field enhanced sample stacking (FESS) method, which is another widely used sample stacking mechanism of CZE [19, 45, 58]. We then optimized the dynamic pH junction-based CZE-MS, and applied the optimized system for TDP of an *E. coli* proteome.

2.3.1 Comparison of dynamic pH junction and FESS methods

We compared the performance of dynamic pH junction method and FESS method for concentrating intact proteins during CZE-MS across four different sample injection volumes that were 50 nL (2.5% of the total capillary volume), 100 nL (5% of the total capillary volume), 200 nL (10% of the total capillary volume), and 500 nL (25% of the total capillary volume). The stock solution of the standard protein mixture was diluted by a factor of two for the experiments. The sample was finally dissolved in 5% (v/v) AA (BGE) for control, 2.5% (v/v) AA in water containing 35% (v/v) ACN (lower conductivity than BGE) for FESS, and 10 mM NH4HCO3 (pH 8.0) for dynamic pH junction. The protein sample contained six different standard proteins with varied molecular weights (12-66 kDa) and isoelectric points (pI 4.5-11). The BGE was 5% (v/v) AA (pH 2.4). Choices for the BGE and the sample buffer for dynamic pH junction method was based on previous work [21, 31].

Figure 2.1 summarizes the results of the comparison experiments. Figure 2.1A-2.1C shows the change of protein intensity as a function of sample injection volume (50, 100, 200, and 500 nL) for control (A), FESS (B) and dynamic pH junction (C) methods. All of the protein intensity were obtained from the extracted ion electropherograms (EIEs) of the protein mixture. Error bars represent the standard deviations of protein intensity from triplicate CZE-MS analyses. The four proteins (lysozyme, cyto.c, myoglobin and CA) were extracted with m/z 1590.33, 765.33, 808.20, and 880.55, respectively. For (A)-(C), β-casein was extracted with m/z 1043.76. For (D), three different m/z (m/z 1043.76, 1045.45 and 1048.5) corresponding to three different forms of β-casein separated by CZE using dynamic pH junction method (e3, e2 and e1) were used for extraction. Figure 2.1D shows the EIEs

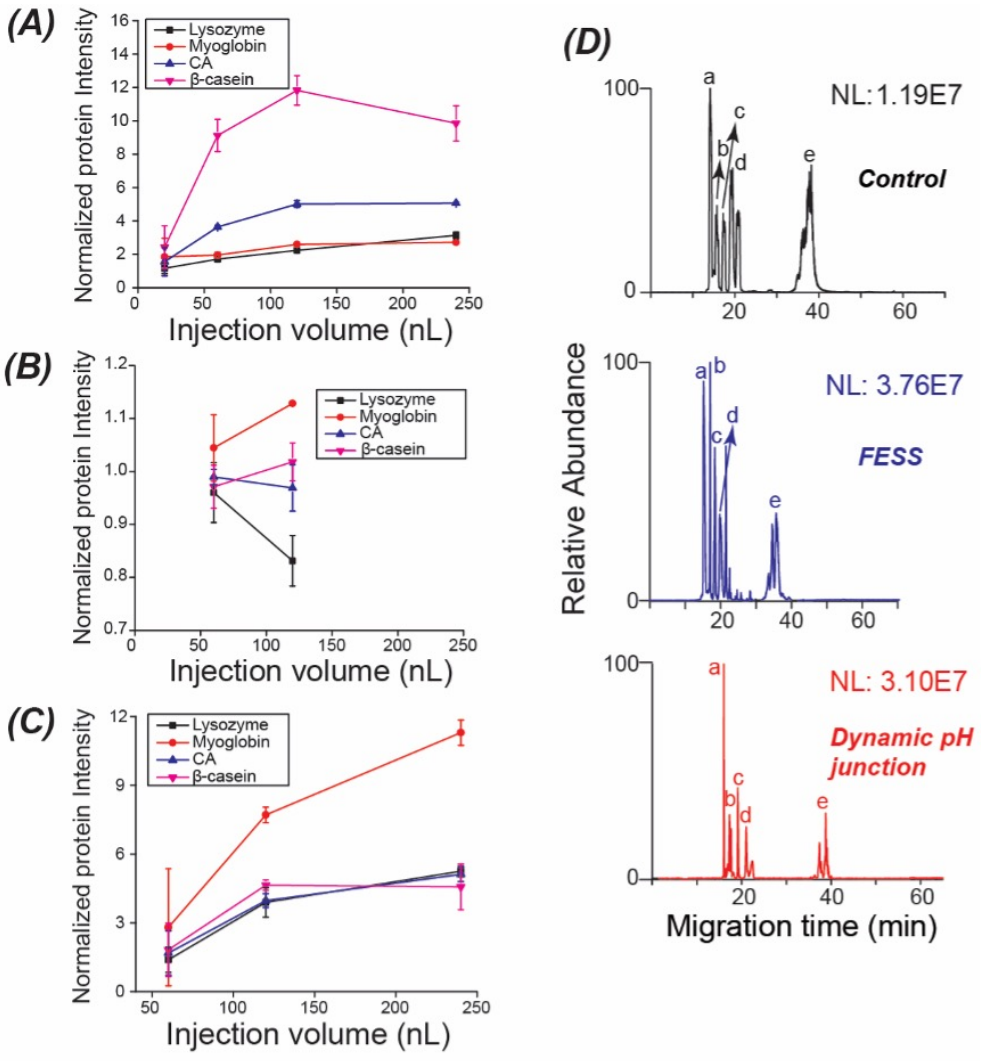


Figure 2.1: Protein intensity change across different sample injection volumes for control (A), FESS (B) and dynamic pH junction (C). (D) EIEs of the mixture of standard proteins from CZE-MS under the three different conditions. Proteins shown: lysozyme (a), cyto.c (b), myoglobin (c), CA (d) and β-casein (e). This figure is reprinted with permission from reference [3].

of the mixture of standard proteins from control, FESS and dynamic pH junction experiments with 500 nL sample injection. We detected BSA in all of the experiments, however, its signal-to-noise ratio was low due to its large molecular weight. Therefore, we did not extract the peak of BSA for comparison.

As shown in Figure 2.1A (control), the intensity of proteins (except lysozyme)

reasonably increased as the injection volume increased from 50 nL to 100 nL. On average, the increase of intensity of the five proteins was about two times. We noted that the intensity change of lysozyme was negligible. The intensity of proteins (except cyto.c) were reasonably consistent when the injection volume increased from 100 nL to 500 nL. On average, the change of intensity of the five proteins was less than 10%. We noted that the intensity of cyto.c from 500-nL sample injection was significantly lower than that from 100-nL sample injection, which was most likely due to the electrospray ionization suppression from BSA. BSA and cyto.c were partially separated by CZE with 100-nL sample injection volume, but they co-migrated out of the separation capillary when the sample injection volume increased to 500 nL.

As shown in Figure 2.1B (FESS), on average, the protein intensity increased roughly by two times when the injection volume increased from 50 nL to 100 nL. We observed reasonably steady intensity of proteins (except cyto.c) when the sample injection volume increased from 100 nL to 500 nL. On average, the increase of protein intensity was only around 20%. The data indicated that FESS method could not efficiently concentrate protein molecules when the sample injection volume was higher than 100 nL, corresponding to 5% of the total capillary volume. We noted that the intensity of cyto.c declined significantly, which is also due to the ionization suppression from BSA mentioned in the previous paragraph. We also noted that the protein intensity from FESS method was, on average, 2-3 times higher than that from control with the same sample injection volume, which is due to the stacking performance of FESS. As shown in Figure 2.1D, the protein intensity from FESS was much higher than that from the control. In addition, FESS method yielded much better separation of proteins than the control, which is also due to its concentration performance. Lysozyme, cyto.c, myoglobin and CA showed poor separated in control experiments using the 500-nL sample injection volume. In contrast, the FESS experiments demonstrated reasonable separation, and produced much higher separation efficiency than control. For example, the number of theoretical plates of myoglobin was less

than 400 for control and around 6,600 for FESS with 500-nL sample injection.

As shown in Figure 2.1C (dynamic pH junction), we observed significant increase in intensity for all five proteins when the injection volume increased from 50 nL to 100 nL. On average, the protein intensity increase was about 2 times. We still observed significant protein intensity increase when the sample injection volume changed from 100 nL to 500 nL. On average, the intensity of proteins from 500 nL sample injection were about 2 times higher than that from 100-nL sample injection. The result demonstrated that the dynamic pH junction method could efficiently concentrate protein molecules with even 500-nL sample injection volume, which corresponded to 25% of the total capillary volume. On average, the intensity from the five proteins using the dynamic pH junction method showed a comparable intensity to that of the FESS method when evaluating the 50-, 100-, and 200-nL sample injection volumes. During the 500-nL injection volume, the intensity showed an 80% improvement when comparing the dynamic pH junction method to the FESS method, as shown in Figure 2.1D. Dynamic pH junction method also produced better separation of proteins than FESS method. Myoglobin and CA could only be partially separated with FESS method (R=1); they could be baseline separated with dynamic pH junction method (R=1.6). In addition, three forms of β-casein [19, 59]. having different masses were well separated with dynamic pH junction method; they could not be separated from each other with FESS method. The mass of those three β-casein forms were 23,983 Da (e3), 24,022 Da (e2) and 24,092 Da (e1), which were manually calculated based on the most abundant isotope peaks of those forms at charge +23. We noted that the calculated mass of those three forms were different from those reported in reference, which were the monoisotopic masses of those three forms exported from MS-Deconv software [19, 54]. Finally, dynamic pH junction method generated better separation efficiency than FESS method. For example, the number of theoretical plates of myoglobin was 6,600 for FESS and 23,000 for dynamic pH junction. Overall, dynamic pH junction method outperformed the FESS method for characterization of proteins with 500-nL

sample injection, and it was used for the following experiments.

We performed a calibration curve experiment with the dynamic pH junction-based CZE-MS shown in Figure 2.2. Each sample was analyzed by CZE-MS in duplicate runs

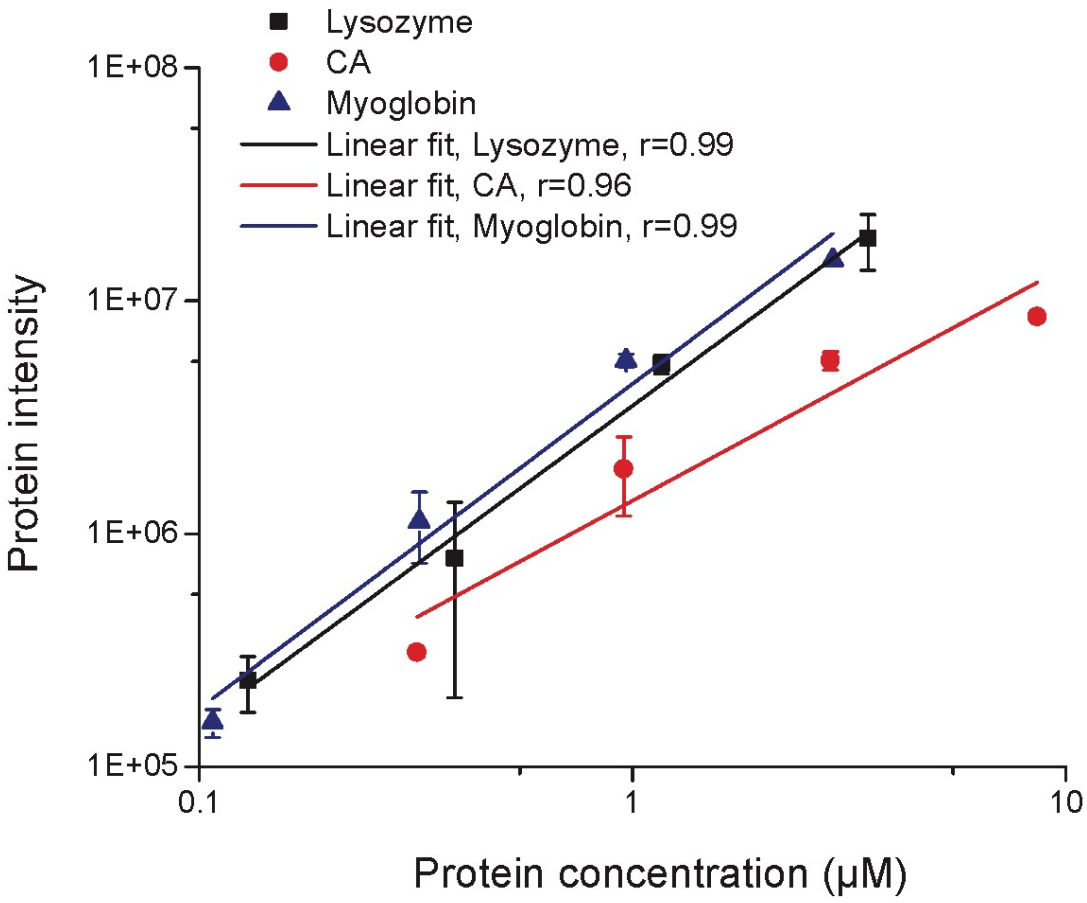


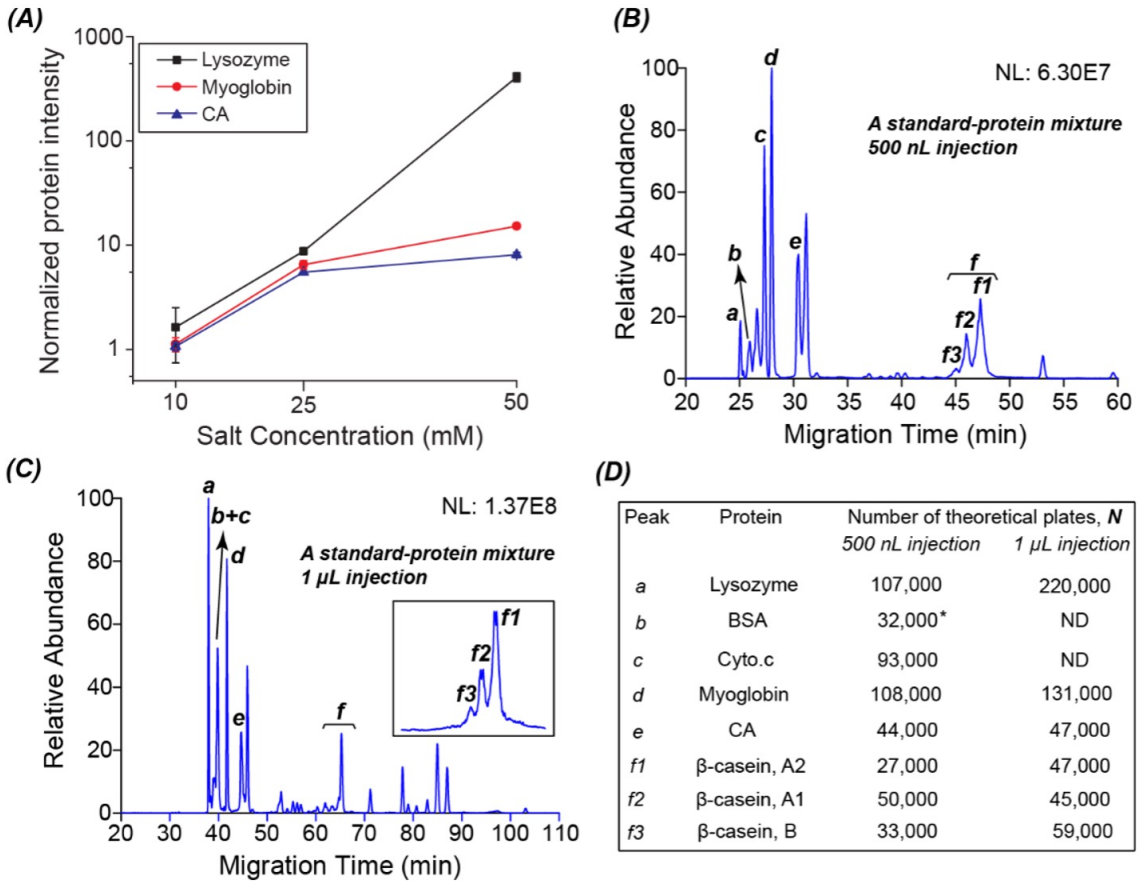
Figure 2.2: Correlations between protein concentration and protein intensity for lysozyme, CA and myoglobin. This figure is reprinted with permission from reference [3].

and error bars represent the standard deviations of protein intensity from the duplicate runs. The stock solution of the standard protein mixture was diluted with NH4HCO3 buffers by four different dilution factors that were 2, 6, 18 and 54, respectively. All of the dilute samples were dissolved in 10 mM NH4HCO3 (pH 8.0). The sample injection volume was 500 nL per CZE-MS run. We chose three proteins (lysozyme, CA and myoglobin) for

the calibration curve and those proteins were detected and well separated in all the CZE-MS runs. Good linear correlations (r=0.96-0.99) were observed between protein concentration and protein intensity for all of the three proteins across nearly 30-times concentration range. The results indicate that the dynamic pH junction-based CZE-MS is quantitative and has the potential for quantitative TDP.

2.3.2 Optimization of the dynamic pH junction-based CZE-MS

We chose 10 mM NH4HCO3 (pH 8.0) as the sample buffer for dynamic pH junction based CZE-MS at the beginning based on previous works [31, 48]. Imami et al. systematically investigated the effect of the concentration of NH4HCO3 in the sample buffer on the concentration performance of dynamic pH junction method using a peptide mixture [36]. They increased the concentration of NH4HCO3 from 20 mM to 200 mM, and observed steady increase of the peptide intensity until 100 mM, which was consistent with an ITP mechanism. We also recognized a similar phenomenon in our recent work [31]. When we increased the concentration of NH4HCO3 in the sample buffer from 5 mM to 20 mM, we observed increase of peptide intensity. Those results motivated us to try higher concentration of NH4HCO3 in the sample buffer. We recognized that when ITP was coupled with CZE-MS for biomolecule analysis, the salt concentration in the sample buffer was typically 50 mM [60–62]. Therefore, we tested 50 mM NH4HCO3 (pH 8.0) as the sample buffer for dynamic pH junction based CZE-MS using the mixture of the six standard proteins, Figure 2.3. The m/z used for protein peak extraction were the same as those in Figure 2.1. The mass tolerance was 100 ppm for peak extraction. The N of each protein was calculated based on the peak width and migration time of each protein in the EIEs. BSA was not extracted in the figures due to its low signal-to-noise ratio. The BGE was 5% (v/v) AA. The stock solution of the standard protein mixture was diluted with a NH4HCO3 buffer (pH 8.0) by a factor of 10, and the concentration of NH4HCO3 in the dilute sample was 50 mM. The dilute sample was used for all of the following experiments.



Peak	Protein	Number of theoretical plates, *N*	
		500 nL injection	*1 µL injection*
a	Lysozyme	107,000	220,000
b	BSA	32,000*	ND
c	Cyto.c	93,000	ND
d	Myoglobin	108,000	131,000
e	CA	44,000	47,000
f1	β-casein, A2	27,000	47,000
f2	β-casein, A1	50,000	45,000
f3	β-casein, B	33,000	59,000

Figure 2.3: EIEs of the standard protein mixture dissolved in 50 mM NH4HCO3 (pH 8.0) analyzed by the dynamic pH junction based CZE-MS with 500-nL sample injection (A) and 1 μL sample injection (B). The number of theoretical plates (N) of different proteins in (A) and (B) are summarized in (C). This figure is reprinted with permission from reference [3].

Figure 2.3 shows the corresponding EIEs of the standard protein mixture using 500-nL sample injection (Figure 2.3A) and 1-μL sample injection (Figure 2.3B). The CZE-MS system using 50 mM NH4HCO3 as the sample buffer produced high separation efficiency for proteins with both 500-nL and 1-μL sample injection. As shown in Figure 2.3C, the N of proteins ranged from 21,000 (β-casein, peak e2) to 206,000 (lysozyme) for 500-nL sample injection, and ranged from 30,000 (β-casein, peak e3) to 292,000 (lysozyme) for 1-μL sample injection. On average, the intensity of proteins from 1-μL sample injection were about 2.5 times higher than those from 500-nL sample injection based on the EIEs. The results indicated that the CZE-MS system using 50 mM NH4HCO3 as the sample buffer

could efficiently concentrate proteins even when 50% of the capillary was filled with the sample.

We also compared the intensity of proteins observed using 10 mM NH4HCO3 and 50 mM NH4HCO3 as the sample buffers based on the EIEs in Figure 2.1D and Figure 2.3A. 50 mM NH4HCO3 sample buffer generated, on average, comparable intensity of proteins with 5-times lower protein concentration compared with 10 mM NH4HCO3 sample buffer. Therefore, we chose 50 mM NH4HCO3 (pH 8.0) as the sample buffer in all of the following experiments.

Next, we screened different BGEs including 0.1-0.5% (v/v) FA and 5-10% (v/v) AA. The sample injection volume was 500 nL per CZE-MS run. We observed that the overall performance of 0.1% (v/v) FA BGE (pH 2.8) was better than that of 0.3% and 0.5% (v/v) FA (pH 2.3 and 2.1) in terms of protein intensity, Figure 2.4. We also observed comparable protein intensity from 0.1% (v/v) FA BGE and AA BGEs (5% and 10% (v/v)), Figure 2.4B and Figure 2.4C. However, 5% and 10% (v/v) AA BGEs (pH $\sim$ 2.4 and $\sim$ 2.2) produced significantly wider separation window than 0.1% (v/v) FA BGE for the standard protein mixture. In addition, the migration time of protein analytes in the capillary in 5% and 10% (v/v) AA BGEs was significantly longer than that in 0.1% (v/v) FA BGE (e.g., 10 minutes longer for lysozyme). There are two potential reasons for the phenomenon. First, 5-10% (v/v) AA has much lower pH than 0.1% (v/v) FA (2.4-2.2 vs. 2.8), which further reduces the remaining EOF in the LPA-coated separation capillary. We measured the EOF in the LPA-coated capillary based on the method reported in literature [50, 51]. The EOF in 10% (v/v) AA BGE was lower than that in 0.1% (v/v) FA BGE (6.8×10^{-6}cm^{2}V^{-1}s^{-1} vs. 1.1×10^{-5}cm^{2}V^{-1}s^{-1}). Second, the more acidic a BGE is (5-10% (v/v) AA) typically leads to more severe protein unfolding and an increase in hydrodynamic radii of the protein analytes, resulting in slower migration of proteins in the separation capillary.

Based on the results discussed above, we chose 5-10% (v/v) AA and 50 mM NH4HCO3

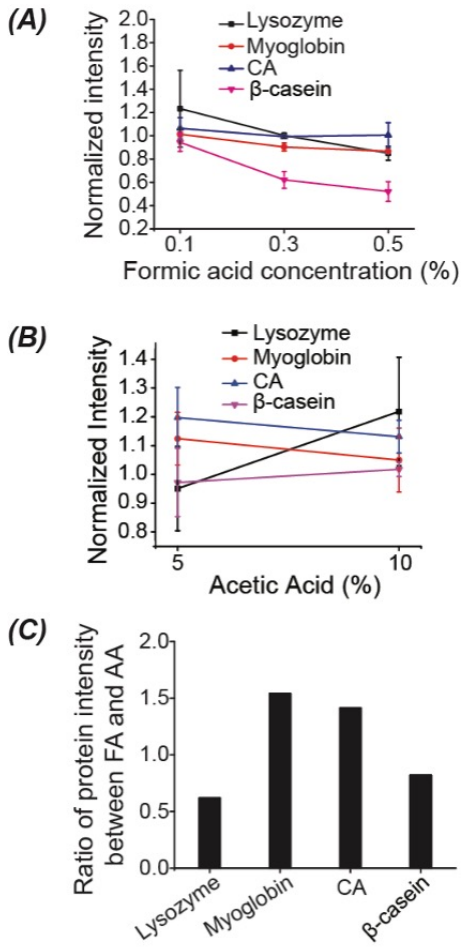


Figure 2.4: Base peak electropherograms of the standard protein mixture dissolved in 50 mM NH4HCO3 (pH 8.0) analyzed by the dynamic pH junction-based CZE-MS with 500-nL sample injection (A) and 1-μL sample injection (B). This figure is reprinted with permission from reference [3].

(pH 8.0) as the optimized BGE and sample buffer for the following experiments. We then evaluated the reproducibility of the optimized dynamic pH junction-based CZE-MS system for intact protein analysis; using 5% AA as the BGE, and 500-nL sample injection volume. The system produced reproducible separation and detection of proteins during 16 hours of continuous analysis (11 CZE-MS runs) with the relative standard deviations (RSDs) of migration time and intensity of proteins less than 7% and 16%, respectively, Table 2.1. One LPA-coated capillary can typically be used for continuous analysis of complex samples for at least one week without significant loss of separation performance based on our

Table 2.1: Summary of the reproducibility data, relative standard deviations (%), from the 11 CZE-MS runs.

Protein	Migration time	Intensity
Lysozyme	6.6	15.4
BSA	6.0	9.2
Cyto.c	5.9	15.4
CA	4.4	8.7
β-casein	5.4	12.2

experience, signifying that the inner wall of the LPA coating of the separation capillary is stable.

2.3.3 Single-shot TDP with CZE-MS/MS

We further applied the optimized CZE-MS/MS system for TDP of *E.coli*. An *E.coli* protein sample (2 mg/mL) in 50 mM NH4HCO3 (pH 8.0) was used for the experiments. AA (10% v/v) was used as the BGE. A QEHF mass spectrometer was used.

First, we evaluated the effect of sample loading volume on the number of proteoform IDs and the number of PrSMs, Figure 2.5. A top 3 DDA method was used for data acquisition. CZE-MS/MS with 500-nL sample loading volume produced the highest number of proteoform IDs (407) after filtered with 1% spectrum-level FDR. When the sample loading volume increased from 100 nL to 500 nL, the number of PrSMs increased, leading to identification of over 2,100 PrSMs with 500-nL sample injection after filtering with 1% spectrum-level FDR. The number of PrSMs remained reasonably consistent when the sample loading volume changed from 500 nL to 1 μL. We further tried to decrease the voltage applied at the injection end of the separation capillary from 30 kV to 20 kV, resulting in slower migration of analytes in the capillary and wider separation window. 468 proteoforms were identified with 20 kV voltage and 500-nL sample loading volume. The number of proteoform IDs was 15% higher than that from the 30 kV voltage (468 vs. 407).

We then performed CZE-MS/MS analysis of the *E.coli* sample in duplicate with 20 kV

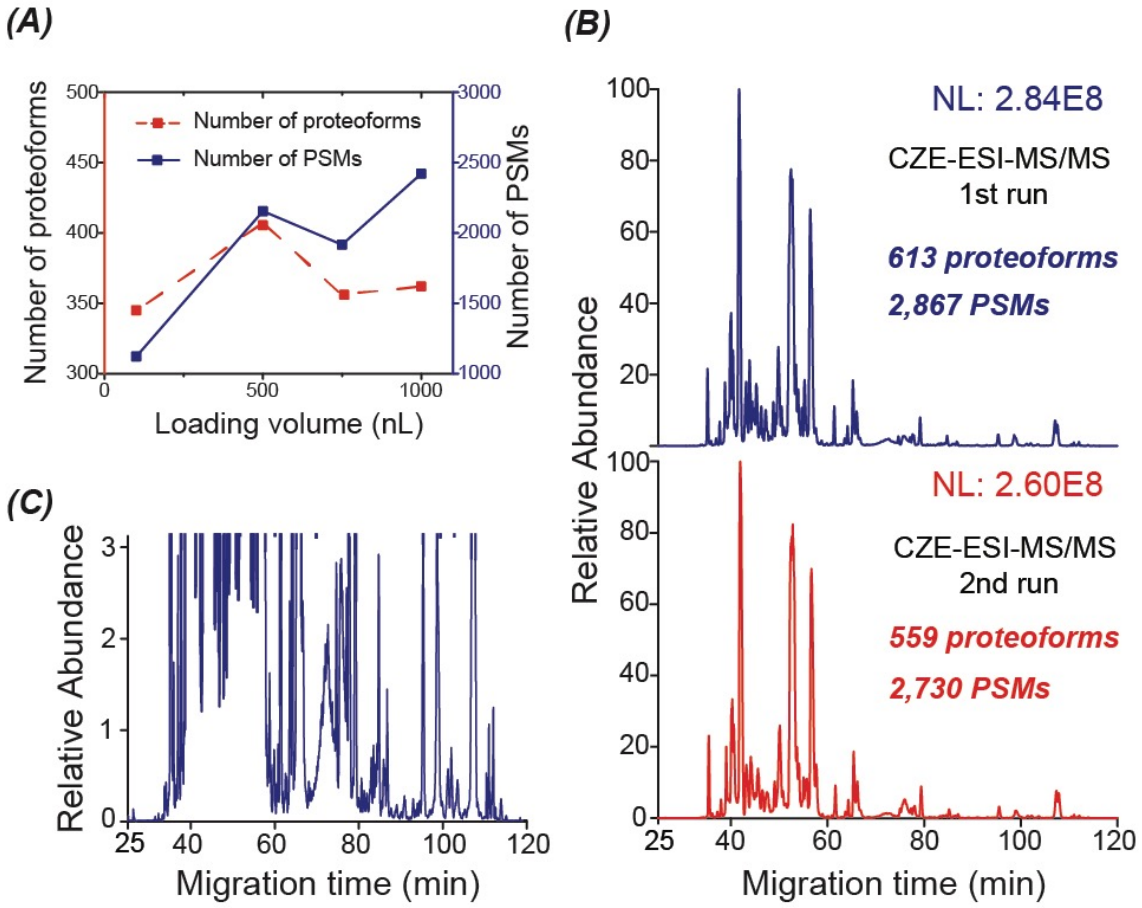


Figure 2.5: Data about TDP of *E. coli* using CZE-MS/MS. (A) Effect of *E. coli* sample loading volume on the number of proteoform IDs and the number of proteoform-spectrum matches (PrSMs). (B) Electropherograms of the *E. coli* protein sample analyzed by CZE-MS/MS in duplicate runs. (C) The zoom-in electropherogram of the *E.coli* protein sample from the 1st run CZE-MS/MS in (B). This figure is reprinted with permission from reference [3].

voltage and 500 nL sample loading, Figures 2.5B and 2.5C. We applied a top 8 DDA method instead of the top 3 DDA method. We identified 586±38 proteoforms (n=2) and 2,798±97 PrSMs (n=2) with single shot CZE-MS/MS after filtered with 1% spectrum-level FDR. The lists of identified proteoforms from those duplicate CZE-MS/MS runs can be found elsewhere [3]. The corresponding raw files have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD007273 [63]. The number of proteoform IDs is over three times higher than

that reported in the literature using single-shot CZE-MS/MS (586 vs. 140-180) [21–23]. Results clearly demonstrate the capability of CZE-MS/MS for large-scale TDP. If 0% spectrum-level FDR was used to filter the data, 419±25 proteoforms still could be identified, which is still over 2 times higher than the data that has been presented in the literature. We also analyzed the molecular weight distribution of the identified proteoforms from the single-shot CZE-MS/MS, Figure 2.6. The molecular weight of identified

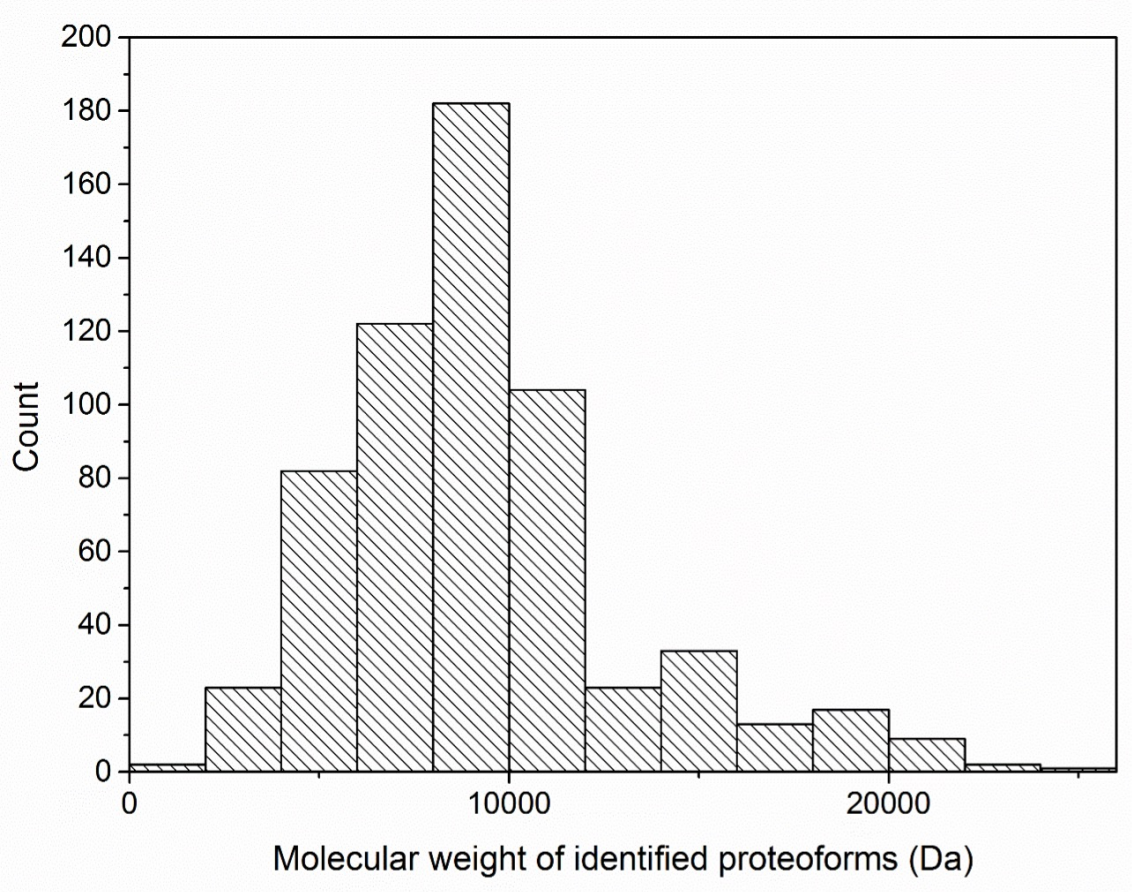


Figure 2.6: Distribution of the molecular weight of identified proteoforms from single-shot CZE-MS/MS. This figure is reprinted with permission from reference [3].

proteoforms ranged from $\sim 2,000$ Da to $\sim 24,000$ Da. About 33% of the identified proteoforms had molecular weight higher than 10 kDa. 1% spectrum-level FDR was used to filter the data.

We attributed the significant improvement of the number of proteoform IDs from

single-shot CZE-MS/MS to three reasons. First, the large sample loading capacity of the CZE-MS/MS system (0.5 μL, 1 μg of *E.coli* proteins) and the optimized dynamic pH junction method guaranteed the identification of large numbers of proteoforms. Second, the optimized dynamic pH junction-based CZE-MS/MS system produced 90-min separation window for the *E.coli* proteome (Figures 3B and 3C), providing enough time for acquisition of tandem mass spectra. The separation window is about three times wider than those in previous reports [21–23]. Third, the dynamic pH junction-based CZE produced higher P_c for separation of the *E.coli* proteome. Based on the electropherograms in Figure 3B, the P_c of the system was ~ 280 (using the average peak width at 50% peak height) for the *E.coli* proteome sample, which is 2-3 times higher than those in the previous reports [21–23].

We further analyzed the identified proteoforms from the *E.coli* proteome with single-shot CZE-MS/MS. The nearly 600 proteoforms from single-shot CZE-MS/MS corresponded to about 200 *E.coli* genes. On average, we identified about three proteoforms from each gene. Distribution of the number of proteoform IDs from each gene is shown in Figure 2.7. The single-shot *E. coli* data in Figure 3B was used for these analyses. We identified one proteoform/gene for about 100 *E.coli* genes, 2-5 proteoforms/gene for about 80 genes, and 6-44 proteoforms/gene for about 20 genes. We identified 44, 30 and 21 proteoforms for *E.coli* genes hdeA, acpP and ybgS, respectively. The proteins corresponding to those three genes are the most abundant proteins in *E.coli* (top 5%) based on the information in PaxDb (Protein Abundance Database, http://pax-db.org/). About 80% and 65% of the identified proteoforms from single-shot CZE-MS/MS had significant mass errors with and without consideration of cysteine carbamidomethylation, respectively. Figure 2.7B shows the distribution of detected mass errors of identified proteoforms. In total, we detected 870 mass error events corresponding to different modifications of the proteins including cysteine carbamidomethylation (57 Da), oxidation (16 Da) and acetylation (42 Da). In addition, truncations, N-terminal methionine excision and signal peptide removal of proteins were also detected. Figures 2.7C and 2.7D show sequences and observed fragmentation patterns

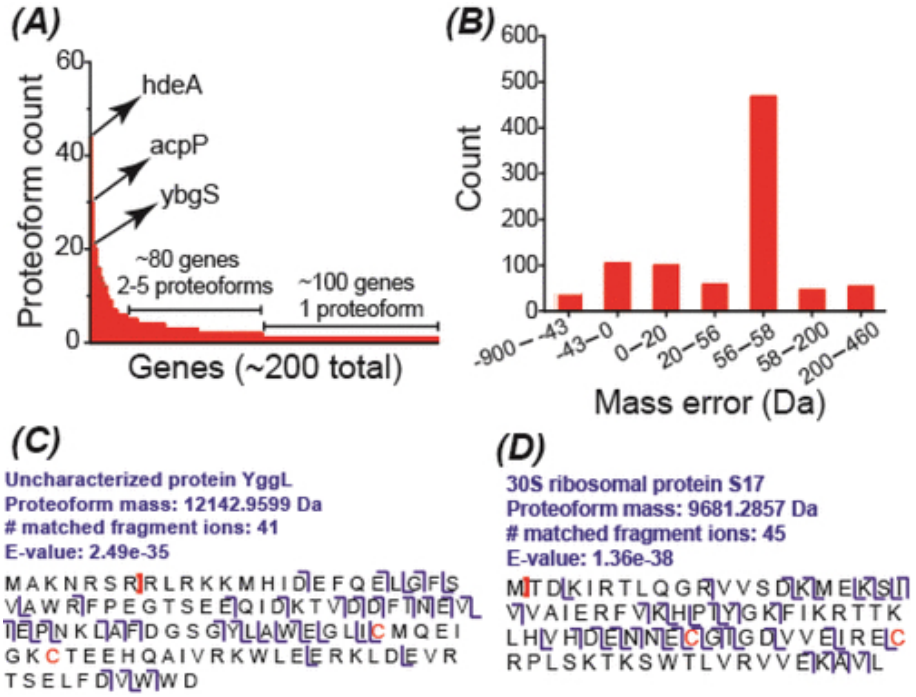


Figure 2.7: (A) Distribution of the number of identified proteoforms from each *E. coli* gene. (B) Distribution of the detected mass errors from the identified proteoforms. (C and D) Sequences of two identified proteins with carbamidomethylation sites (cysteines) marked in red and the fragmentation patterns observed. This figure is reprinted with permission from reference [3].

of two proteins. The fragmentation covered the termini and middle parts of those two proteins, leading to identification of over 40 fragment ions. N-terminal truncation was detected for uncharacterized protein YggL (Figure 2.7C), while there was N-terminal methionine excision that was detected for 30S ribosomal protein S17 (Figure 2.7D).

2.3.4 Comparing RPLC-MS and CZE-MS for analysis of a standard protein mixture

We first employed a standard protein mixture containing proteins with molecular weight in a range of 8.5-66 kDa for the comparison of RPLC-MS and CZE-MS in terms of separation, signal intensity, and charge state distributions of proteins. For CZE-MS, we employed an LPA-coated capillary to reduce the protein adsorption on the inner wall of the capillary

and to reach a wider separation window [3]. We used a dynamic pH junction-based sample stacking method for online concentration of the proteins based on our recent work [3]. For RPLC-MS, we employed a 90-cm-long nanoRPLC column (100-μm i.d.) packed with 3 μm C2 porous beads (300 Å pores). We got the nanoRPLC column from Dr. Yufeng Shen at CoAnn Technologies, LLC (Richland, WA). Based on previous work, the column represents one of the state-of-the-art nanoRPLC columns for the separation of complex intact protein mixtures [8]. We employed a QEHF mass spectrometer in the experiment. The MS parameters for CZE-MS and RPLC-MS were the same.

Figure 2.8 shows the separation profiles and charge state distributions of proteins using CZE-MS and RPLC-MS. First, CZE and RPLC have different mechanisms for protein separation, size-to-charge ratio vs. hydrophobicity, leading to drastically different migration or elution orders, as shown in Figures 2.8A and 2.8B. 40 and 400 ng of proteins were injected in Figures 2.8A and 2.8B, respectively. For example, BSA migrated fastest in CZE and had stronger retention than ubiquitin and cyto.c. It suggests that a combination of RPLC and CZE can produce orthogonal and high capacity separation of complex mixtures of intact proteins. We recently demonstrated the power of nanoRPLC-CZE-MS/MS for orthogonal and high capacity separation of an MCF7 cancer cell proteome digest, leading to the identifications of nearly 8000 proteins and 60000 peptides starting from only 5-μg peptides [64]. We noted that RPLC-MS produced a very broad peak of cyto.c (over 10-mins wide), and CZE-MS yielded a reasonably sharp peak of cyto.c (less than 1-min wide). The broad peak of cyto.c in RPLC-MS might be due to its weak retention on the C2 column, leading to its inefficient trapping at the front end of the column during sample loading. Second, CZE-MS had much higher sensitivity than RPLC-MS. CZE-MS with 40 ng protein injected produced comparable protein intensity to RPLC-MS with 400 ng protein injected, as shown in Figures 2.8A, 2.8B, and Table 2.2. Cyto.c and BSA are not listed in the table because they were not separated well in CZE, and because cyto.c had a very broad peak in RPLC runs. S/N ratios were estimated by

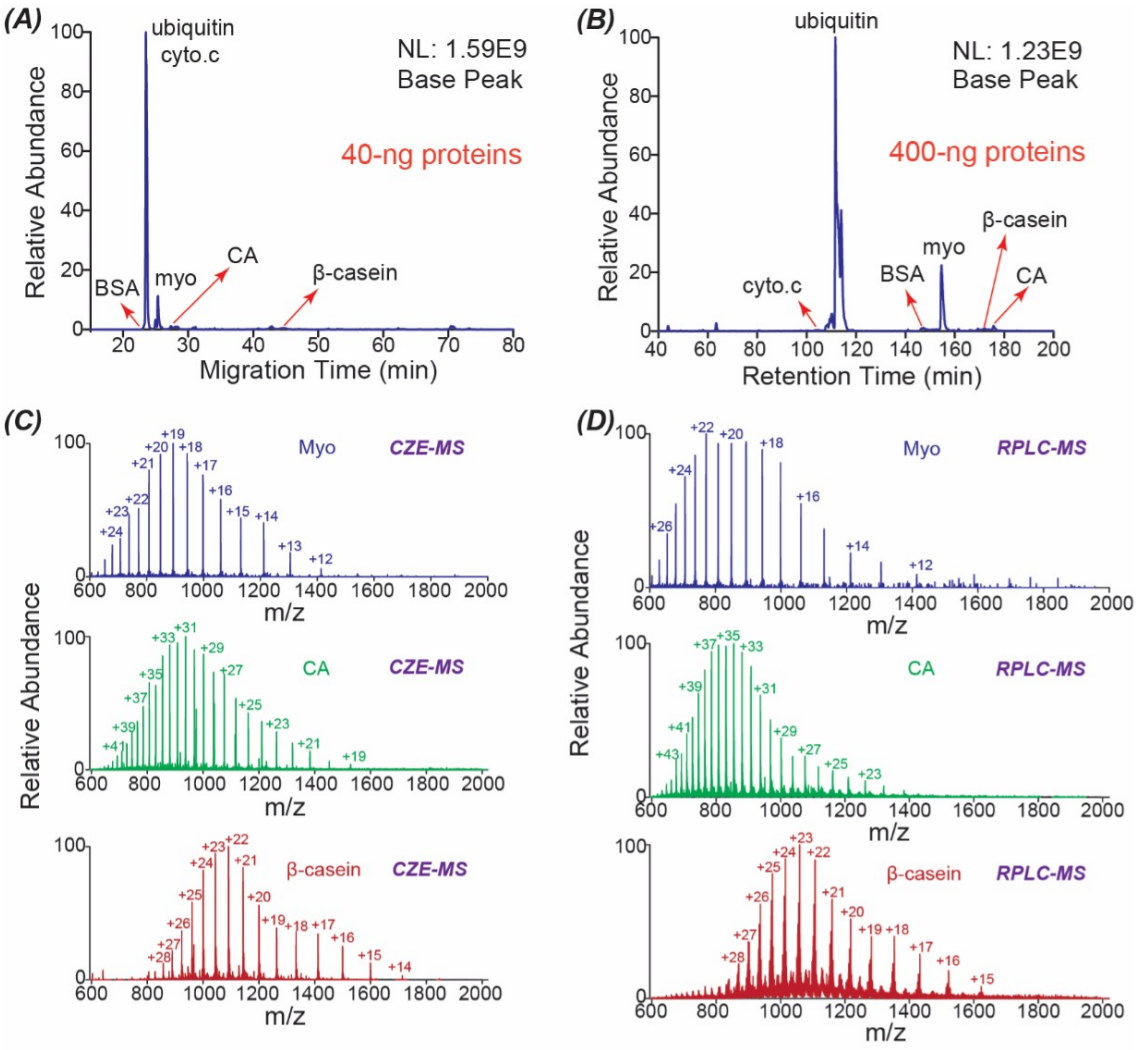


Figure 2.8: The standard protein mixture data from CZE-MS and RPLC-MS. (A) Base peak electropherogram of the protein mixture using CZE-MS. (B) Base peak chromatogram of the protein mixture using RPLC-MS. (C) Charge state distributions of myo, CA, and β-casein using CZE-MS. (D) Charge state distributions of myo, CA, and β-casein using RPLC-MS. This figure is reprinted with permission from reference [4].

dividing base peak intensity of the protein by the base peak intensity at the base of that protein's peak. The data agreed well with that from Yates group recently [5]. The high sensitivity of CZE-MS makes it extremely useful for TDP of mass-limited samples. Very recently, we showed that thousands of proteoforms were identified from zebrafish brains using advanced CZE-MS/MS with only 500 ng protein material [65]. Third, interestingly, the protein in RPLC-MS tended to have higher charge states than that in CZE-MS, as

Table 2.2: Base peak intensity of proteins in the standard protein mixture from CZE-MS (40 ng protein) and RPLC-MS (400 ng protein) and their S/N ratios.

Protein	CZE-MS (40 ng proteins)	RPLC-MS (400 ng proteins)	S/N (CZE-MS)	S/N (RPLC-MS)
Ubiquitin	1.6E9	1.2E9	1.3E4	3.0E3
Myo	1.8E8	2.7E8	90	1.7E2
CA	1.8E7	2.3E7	11	64
β-casein	7.4E6	1.2E7	37	17

shown in Figures 2.8C and 2.8D. For example, the most abundant charge states of CA in CZE-MS and RPLC-MS are +31 and +35, respectively. This phenomenon is most likely due to the high acetonitrile concentration in RPLC mobile phase, leading to the more extensive unfolding of proteins in RPLC compared to that in CZE. Higher charge states can potentially benefit gas-phase fragmentation of protein for identifications due to their more unfolded structures. Fourth, RPLC-MS produced a much wider separation window than CZE-MS for the standard protein mixture, 80 min vs. 20 min, as shown in Figures 2.8A and 2.8B. Because we employed a 90-cm-long RPLC column, we used a 240-min gradient for separation, leading to a wide separation window. This feature becomes one very important advantage of RPLC-MS compared to CZE-MS for TDP. For large-scale TDP, a wide separation window is vital because more MS/MS spectra can be acquired during a run for more extensive characterization of complex protein mixtures.

2.3.5 Comparing RPLC-MS/MS and CZE-MS/MS for TDP of *E. coli* cells

We then applied both RPLC-MS/MS and CZE-MS/MS for top-down MS characterization of an *E. coli* proteome sample. Two kinds of sample loading, 1-μg and 8-μg *E. coli* protein, were tested for RPLC-MS/MS. For CZE-MS/MS, 1-μg *E. coli* protein was injected for analysis. The same MS and MS/MS parameters were used for CZE-MS/MS and RPLC-MS/MS. A 240-min gradient was used for RPLC-MS/MS analyses because the RPLC column was long (90 cm). The CZE-MS/MS analysis was completed in 120 min, as

shown in Figure 2.9A. CZE-MS/MS with 1-μg protein produced only 27% lower total ion

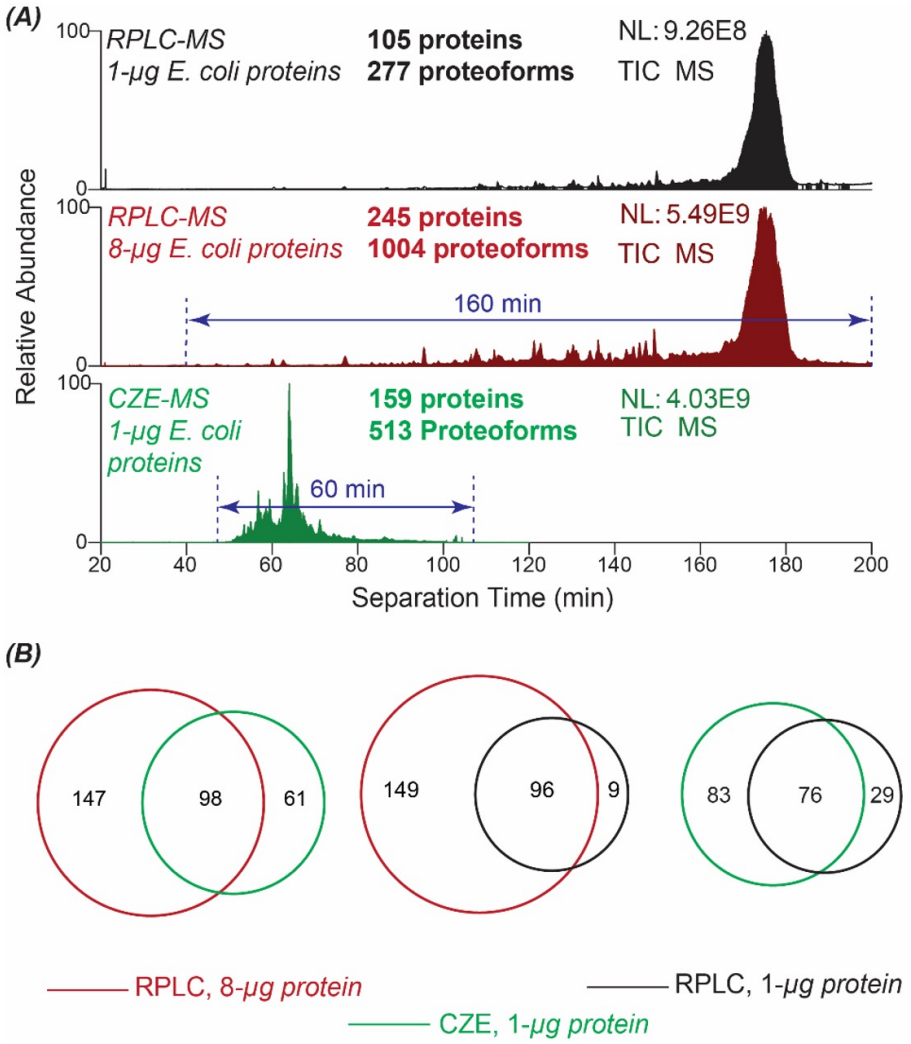


Figure 2.9: Summary of the *E. coli* data from RPLC-MS/MS and CZE-MS/MS. (A) Total ion current (TIC) chromatograms of *E. coli* proteins from RPLC-MS/MS with 1-μg and 8-μg protein injected, and the TIC electropherogram of *E. coli* protein from CZE-MS/MS with 1-μg protein injected. (B) Protein-level overlaps between RPLC-MS/MS (8-μg protein), CZE-MS/MS (1-μg protein), and RPLC-MS/MS (1-μg protein). This figure is reprinted with permission from reference [4].

chromatogram (TIC) signal than RPLC-MS/MS with 8-μg protein (4.03E9 vs. 5.49E9), and it generated over 4-fold higher TIC signal than RPLC-MS/MS with 1-μg protein (4.03E9 vs. 9.26E8). RPLC-MS/MS produced a much wider separation window than CZE-MS/MS (160 min vs. 60 min). Overall, RPLC-MS/MS identified 245 proteins and

1,004 proteoforms using 8-μg protein and 105 proteins and 277 proteoforms using 1-μg protein. CZE-MS/MS identified 159 proteins and 513 proteoforms using 1-μg protein. A 5% proteoform-level FDR was used to filter the proteoform identifications.

With the same loading amount of 1-μg *E. coli* protein, CZE-MS/MS was able to identify 51% more proteins and 85% more proteoforms. However, RPLC-MS/MS with 8-μg *E. coli* protein identified 54% more proteins and 96% more proteoforms compared to CZE-MS/MS with 1-μg protein. This data highlights the advantages and limitations of CZE compared to RPLC. The better sensitivity of CZE results in more protein and proteoform identifications than RPLC-MS/MS with the 1-μg protein material, but the increased loading capacity of RPLC (8-μg protein) and wider separation window allowed for more protein and proteoform identifications. In order to improve the CZE-MS/MS for more proteoform identifications, we recently developed a CZE-MS/MS system using a much longer LPA-coated capillary compared to this work (1.5 m vs. 1 m) [65]. The novel CZE-MS/MS system produced a 180-min separation window, leading to the identifications of 800 proteoforms and 260 proteins from an *E. coli* sample [65].

We then evaluated the protein-level overlaps between RPLC-MS/MS and CZE-MS/MS, as shown in Figure 2B. RPLC-MS/MS (8-μg protein) and CZE-MS/MS (1-μg protein) identified 306 unique proteins in total with 98 shared between the two methods for a total overlap of 32%. 96 of the 105 proteins identified in the 1-μg RPLC-MS/MS run were also found in the 8-μg RPLC-MS/MS run; 76 of the 105 proteins identified in the 1-μg RPLC-MS/MS run were identified by CZE-MS/MS.

We further compared the RPLC-MS/MS (8-μg protein) and CZE-MS/MS (1-μg protein) data regarding the molecular weight of identified proteoforms. The RPLC-MS/MS and CZE-MS/MS runs identified proteoforms with molecular weight in a range of 1-22 kDa and 3-23 kDa, respectively. The molecular weight distributions of proteoforms from these two methods have no drastic differences, as shown in Figure 2.10. We need to note that because

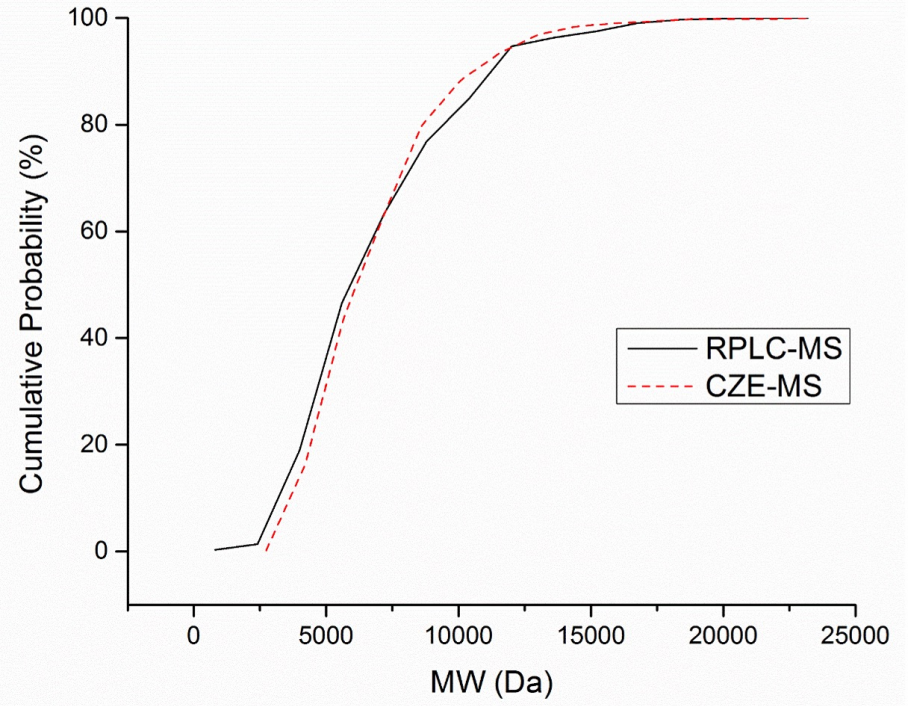


Figure 2.10: Molecular weight distributions of identified proteoforms from the *E. coli* sample using RPLC-MS/MS (8-μg protein) and CZE-MS/MS (1-μg protein). This figure is reprinted with permission from reference [4].

the proteoforms identified in this work are relatively small (less than 23 kDa), there may be drastic differences between RPLC-MS/MS and CZE-MS/MS for identification of large proteoforms (larger than 30 kDa), which will be tested in our future work.

We identified various modifications on the protein sequences, including N-terminal methionine removal, signal peptide cleavage, truncations, N-terminal acetylation, succinylation, disulfide bond, and S-thiolations. For example, we detected a -2 Da mass shift on one proteoform of Thioredoxin 1, and the mass shift was localized in a small region of the proteoform sequence (WCGPCK), as shown in Figure 2.11A, suggesting a disulfide bond between the two cysteine residues. The data agree well with the information in the UniProt database. As another example, we detected two kinds of S-thiolations for Chaperone protein DnaK and they are S-glutathionylation (+305-Da mass shift) and S-cysteinylation (+119-Da mass shift), as shown in Figures 2.11B and 2.11C. For the

Figure 2.11: Sequences and fragmentation patterns of thioredoxin 1 with a disulfide bond (A), Chaperone protein DnaK with an S-glutathionylation (B), and Chaperone protein DnaK with an S-cysteinylation (C). The cysteine residues marked with circles have the modifications. This figure is reprinted with permission from reference [4].

Chaperone protein DnaK proteoforms in Figure 2.11, the N-terminal methionine was removed, and the C-terminal was truncated. Examples of the Thioredoxin 1 proteoform with a disulfide bond and the S-thiolation proteoforms of dnaK in Figure 2.11 are from the CZE-MS/MS study. The disulfide bond and S-glutathionylation was identified using all methods and S-cysteinylation was identified using CZE-MS/MS and 8-μg RPLC-MS/MS.

2.4 Conclusion

We presented a CZE-MS/MS system with μL-scale sample loading capacity, 90-min separation window and high peak capacity ($\sim$ 280) for large-scale TDP, thus leading to nearly 600 proteoform IDs from an *E.coli* proteome using single-shot CZE-MS/MS. The number of proteoform IDs is over three times higher than that from previous single-shot CZE-MS/MS studies. The 600 proteoform IDs from single-shot CZE-MS/MS is roughly equivalent to the data from single-shot RPLC-MS/MS using a 21T FT-ICR mass spectrometer [66]. This CZE-MS/MS system established the foundation for large-scale TDP using CZE-MS/MS.

RPLC-MS/MS is typically used for large-scale TDP, and separates proteins based on their hydrophobicity. CZE separates proteins based on their size-to-charge ratios. CZE and RPLC can provide orthogonal separation of intact proteins. It has been reported that CZE-MS approached better characterization of Dam1 complex subunits in terms of separation efficiency and resolution with 100-times less sample consumption compared to RPLC-MS [5]. In addition, CZE can separate protein(s)/protein complexes in native condition [67, 68]. Very recently, Belov et al. characterized a ribosomal isolate from *E. coli* using CZE-MS/MS in native condition, leading to the identification of 42 ribosomal proteins and 137 proteoforms in a single experiment [68]. The results demonstrate the potential of CZE-MS/MS for TDP of complex proteomes in native conditions.

However, CZE-MS/MS based large-scale TDP is still at the early stage. The 600 proteoform IDs in this work represents the largest TDP dataset using CZE-MS/MS. The number of proteoform IDs from CZE-MS/MS, at the time of this work, was still far away from the state of the art of LC-MS/MS-based TDP, which had reached thousands of proteoform IDs from mammalian cell lines [6, 7, 66, 69–71]. Around 1,000 proteoform IDs from complex proteome samples has been approached using one-dimension high-resolution

RPLC-MS/MS [8, 72]. To improve the scale of CZE-MS/MS-based TDP, we need to further improve the CZE-MS/MS system in terms of the separation window and sample loading capacity. One solution could be to use longer separation capillary (e.g., 1.5 meters) and higher separation voltage (e.g., 60 kV or higher) [65].

This work also represents the first comparison of the state-of-the-art nanoRPLC-MS/MS and CZE-MS/MS for TDP. Overall CZE-MS/MS has drastically better sensitivity than nanoRPLC-MS/MS for characterization of a simple intact protein mixture and an *E. coli* proteome. CZE-MS/MS can be very useful for top-down MS characterization of mass-limited proteome samples. RPLC has large sample loading capacity and wide separation windows for fraction collection and CZE-MS/MS can characterize proteoforms with high sensitivity. More importantly, RPLC and CZE are orthogonal for separation of proteoforms. Combining RPLC and prefractionation (using SEC or RPLC) to CZE-MS/MS should be an ideal platform for large-scale TDP.

2.5 Acknowledgments

We thank Prof. Heedeok Hong's group at Department of Chemistry, Michigan State University for kindly providing the *Escherichia coli* cells for our experiments. This research was funded by Michigan State University and the National Institute of General Medical Sciences, National Institutes of Health (USA) through Grant R01GM125991.

BIBLIOGRAPHY

BIBLIOGRAPHY

(1) Toby, T. K.; Fornelli, L.; Kelleher, N. L. *Annu. Rev. Anal. Chem.* **2016**, *9*, 499–519.

(2) Cravatt, B. F.; Simon, G. M.; III, J. R. Y. *Nature* **2007**, *9*, 991–1000.

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

(4) McCool, E. N.; Sun, L. *Se Pu* **2019**, *37*, 878–886.

(5) Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; III, J. R. Y. *J. Proteome Res.* **2014**, *13*, 6078–6086.

(6) Tran, J. C. et al. *Nature* **2011**, *480*, 254–258.

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

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

(9) Fornelli, L.; Toby, T. K.; Schachner, L. F.; Doubleday, P. F.; Srzentić, K.; DeHart, C. J.; Kelleher, N. L. *J. Proteomics* **2018**, *175*, 3–4.

(10) Smith, L. M.; Kelleher, N. L. *Science* **2018**, *359*, 1106–1107.

(11) Tran, J. C.; Doucette, A. A. *J. Proteome Res.* **2008**, *7*, 1761–1766.

(12) Tran, J. C.; Doucette, A. A. *Anal. Chem.* **2008**, *80*, 1568–1573.

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

(14) Domínguez, E.; Haselberg, R.; Somsen, G. W. *Methods Mol. Biol.* **2016**, *1466*, 25–41.

(15) Haselberg, R.; de Jong, G. J.; Somsen, G. W. *Anal. Chem.* **2013**, *85*, 2289–2296.

(16) Jorgenson, J. W.; Lukacs, K. D. *Science* **1983**, *222*, 266–272.

(17) Harstad, R. K.; Johnson, A. C.; Weisenberger, M. M.; Bowser, M. T. *Anal. Chem.* **2016**, *88*, 299–319.

(18) Valaskovic, G. A.; Kelleher, N. L.; McLafferty, F. W. *Science* **1996**, *273*, 1199–1202.

(19) Sun, L.; Knierman, M. D.; Zhu, G.; Dovichi, N. J. *Anal. Chem.* **2013**, *85*, 5989–5995.

(20) Zhao, Y.; Sun, L.; Champion, M. M.; Knierman, M. D.; Dovichi, N. J. *Anal. Chem.* **2014**, *86*, 4873–4878.

(21) Zhao, Y.; Sun, L.; Zhu, G.; Dovichi, N. J. *J. Proteome Res.* **2016**, *15*, 3679–3685.

(22) Li, Y.; Tran, J. C.; Ntai, I.; Kelleher, N. L. *Proteomics* **2014**, *14*, 1158–1164.

(23) Han, X.; Wang, Y.; Aslanian, A.; Bern, M.; Lavellée-Adam, M.; III, J. R. Y. *Anal. Chem.* **2014**, *86*, 11006–11012.

(24) Bush, D. R.; Zang, L.; Belov, A. M.; Ivanov, A. R.; Karger, B. L. *J. Chromatogr. A* **2017**, *1498*, 1138–1146.

(25) Faserl, K.; Sarg, B.; Sola, L.; Linder, H. H. *Proteomics* **2017**, *17*, doi.org/10.1002/pmic.201700310.

(26) Sarg, B.; Faserl, K.; Kremser, L.; Halfinger, B.; Sebastiano, R.; Lindner, H. H. *Mol. Cell Proteomics* **2013**, *12*, 2640–2656.

(27) Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2554–2560.

(28) Moini, M. *Anal. Chem.* **2007**, *79*, 4241–4246.

(29) Yang, L.; Lee, C. S.; Hofstadler, S. A.; Paša-Tolić, L.; Smith, R. D. *Anal. Chem.* **1998**, *70*, 3235–3241.

(30) Jensen, P. K.; Paša-Tolić, L.; Anderson, G. A.; Horner, J. A.; Lipton, M. S.; Bruce, J. E.; Smith, R. D. *Anal. Chem.* **1999**, *71*, 2076–2084.

(31) Chen, D.; Shen, X.; Sun, L. *Analyst* **2017**, *142*, 2118–2127.

(32) Aebersold, R.; Morrison, H. D. *J. Chromatogr.* **1990**, *516*, 79–88.

(33) Britz-Mckibbin, P.; Chen, D. D. Y. *Anal. Chem.* **2000**, *72*, 1242–1252.

(34) Wang, L.; MacDonald, D.; Huang, X.; Chen, D. D. *Electrophoresis* **2016**, *37*, 1143–1150.

(35) Cao, C. X.; Fan, L. Y.; Zhang, W. *Analyst* **2008**, *133*, 1139–1157.

(36) Imami, K.; Monton, M. R.; Ishihama, Y.; Terabe, S. *J. Chromatogr. A* **2007**, *1148*, 250–255.

(37) Ptolemy, A. S.; Britz-McKibbin, P. *Analyst* **2008**, *133*, 1643–1648.

(38) Shishkova, E.; Hebert, A. S.; Westphall, M. S.; Coon, J. J. *Anal. Chem.* **2018**, *90*, 11503–11508.

(39) Liang, Y.; Jin, Y.; Wu, Z.; Tucholski, T.; brown, K. A.; Zhang, L.; Zhang, Y.; Ge, Y. *Anal. Chem.* **2019**, *91*, 1743–1747.

(40) Zubarev, R. *Proteomics* **2013**, *13*, 723–726.

(41) Li, C.; Tan, X. F.; Lim, T. K.; Lin, Q.; Gong, Z. *Sci. Rep.* **2016**, *6*, 24329.

(42) Aebersold, R. et al. *Nat. Chem. Biol.* **2018**, *14*, 206–214.

(43) Li, Y.; Champion, M. M.; Sun, L.; Champion, P. A. D.; Wojcik, R.; Dovichi, N. J. *Anal. Chem.* **2012**, *84*, 1617–1622.

(44) Zhu, G.; Sun, L.; Dovichi, N. J. *Analyst* **2016**, *141*, 5216–5220.

(45) Jr., S. L. S.; Quirino, J. P.; Terabe, S. *J. Chromatogr. A* **2008**, *1184*, 504–541.

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

(47) Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J. *J. Proteome Res.* **2015**, *14*, 2312–2321.

(48) Zhu, G.; Sun, L.; Dovichi, N. J. *Talanta* **2016**, *146*, 839–843.

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

(50) Williams, B. A.; Vigh, G. *Anal. Chem.* **1996**, *68*, 1174–1180.

(51) Zhang, Z.; Peuchen, E. H.; Dovichi, N. J. *Anal. Chem.* **2017**, *89*, 6774–6780.

(52) Kou, Q.; Xun, L.; Liu, X. *Bioinformatics* **2016**, *32*, 3495–3497.

(53) TopFD `http://proteomics.informatics.iupui.edu/software/topfd/`.

(54) Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. *Mol. Cell. Proteom.* **2010**, *9*, 2772–2782.

(55) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534–2536.

(56) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207–214.

(57) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383–5392.

(58) Sun, L.; Hebert, A. S.; Yan, X.; Zhao, Y.; Westphall, M. S.; Rush, M. J.; Zhu, G.; Champion, M. M.; Coon, J. J.; Dovichi, N. J. *Angew Chem. Int. Ed.* **2014**, *53*, 13931–13933.

(59) Wu, S.; Lourette, N. M.; Tolić, N.; Zhao, R.; Robinson, E. W.; Tolmachev, A. V.; Smith, R. D.; Paša-Tolić, L. *J. Proteome Res.* **2009**, *8*, 1347–1357.

(60) Busnel, J. M.; Schoenmaker, B.; Ramautar, R.; Carrasco-Pancorbo, A.; Ratnayake, C.; Feitelson, J. S.; Chapman, J. D.; Deelder, A. M.; Mayboroda, O. A. *Anal. Chem.* **2010**, *82*, 9476–9483.

(61) Faserl, K.; Kremser, L.; Müller, M.; Teis, D.; Lindner, H. H. *Anal. Chem.* **2015**, *87*, 4633–4640.

(62) Wang, X.; Slebos, R. J. C.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B. *J. Proteome Res.* **2012**, *11*, 1009–1017.

(63) Vizcaíno, J. A. et al. *Nucleic Acids Res.* **2016**, *44*, D447–D456.

(64) Yang, Z.; Shen, X.; Chen, D.; Sun, L. *Anal. Chem.* **2018**, *90*, 10479–10486.

(65) Lubeckyj, R. A.; Basharat, A. R.; Shen, X.; Liu, X.; Sun, L. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 1435–1445.

(66) Anderson, L. C. et al. *J. Proteome Res.* **2017**, *16*, 1087–1096.

(67) Nguyen, A.; Moini, M. *Anal. Chem.* **2008**, *80*, 7169–7173.

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

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

(70) Durbin, K. R.; Fornelli, L.; Fellers, R. T.; Doubleday, P. F.; Narita, M.; Kelleher, N. L. *J. Proteome Res.* **2016**, *15*, 976–982.

(71) Valeja, S. G.; Xiu, L.; Gregorich, Z. R.; Guner, H.; Jin, S.; Ge, Y. *Anal. Chem.* **2015**, *87*, 5363–5371.

(72) Ansong, C. et al. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 10153–10158.

Chapter 3

Large-scale top-down proteomics of a model system

3.1 Introduction

In TDP, intact proteins extracted from cells are typically fractionated by LC or electrophoresis, followed by RPLC-MS/MS analysis. [1] The resulting MS/MS spectra are compared with a protein database derived from the genome sequence for proteoform identifications (IDs) [2–4]. The state-of-the-art RPLC-MS/MS based workflows have approached 3000-5000 proteoform IDs corresponding to around 1000 proteins [5–8].

CZE-MS/MS has been recognized as a useful tool for TDP due to the high resolution of CZE for separation of intact proteins and the high sensitivity of CZE-MS/MS for detection of intact proteins [9–15]. However, the performance of CZE-MS/MS based platforms is still far below that of RPLC-MS/MS based platforms in terms of the number of proteoform IDs. Several groups have made some effort to improve CZE-MS/MS for TDP [16–20]. Li et al. identified 30 large proteins (30-80 kDa) from *P. aeruginosa* PA01 cell lysate using CZE-MS/MS, indicating the potential of CZE-MS/MS for top-down identification of large proteins from a complex proteome [16]. Han et al. coupled RPLC fractionation to CZE-MS/MS for TDP of *Pyrococcus furiosus* and identified nearly 300 proteoforms corresponding to 134 proteins, demonstrating the capability of CZE-MS/MS for large-scale

[1] This chapter was adapted with permission from reference [1]

TDP [17]. Zhao et al. combined high-resolution RPLC fractionation and CZE-MS/MS for large-scale TDP of yeast and observed nearly 600 proteoform and 200 protein IDs [19]. The data represents the state-of-the-art of CZE-MS/MS for top-down proteomics.

Two major issues have limited the number of proteoform IDs from complex proteomes using CZE-MS/MS. One issue is the low sample loading capacity of CZE. The other one is the low P_c of CZE for separation of intact proteins. The sample loading capacity and P_c of CZE was 200 nL or lower and less than 100, respectively, in the reports mentioned in the previous paragraph. Recently, we boosted the sample loading capacity and P_c of CZE-MS/MS to 1 μL and 280, respectively, using dynamic pH junction-based sample stacking for analysis of complex mixtures of intact proteins [20–23]. Duplicate CZE-MS/MS analyses of an *Escherichia coli* (*E. coli*) proteome generated 586±38 proteoform IDs with a 1% spectrum-level FDR [20]. We compared the identified proteoforms from the duplicate CZE-MS/MS analyses and revealed that, on average, about 76% of the proteoform IDs were the same in each CZE-MS/MS run, suggesting the good reproducibility of the CZE-MS/ MS system. The remainder of this chapter deals with applying a multidimensional proteoform separation and identification platform with optimized parameters to a model system, *E. coli*.

3.2 Experimental

3.2.1 Materials and reagents

MS-grade water, ACN, MeOH, FA and HPLC-grade AA were purchased from Fisher Scientific (Pittsburgh, PA). NH4HCO3, urea, DTT, IAA and 3-(trimethoxysilyl)propyl methacrylate were from Sigma-Aldrich (St. Louis, MO). HF (48-51% solution in water) and acrylamide were purchased from Acros Organics (NJ, USA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ). Complete,

mini protease inhibitor cocktail (EASYpacks) was from Roche (Indianapolis, IN).

3.2.2 Sample preparation

E. coli (K-12 MG1655) was cultured in LB medium (37°C) while shaking (225 rpm) until the OD600 reached 0.7. *E. coli* cells were harvested through centrifugation (4000 rpm, 10 min.) and washed three times with PBS. The *E. coli* cells were lysed in a lysis buffer containing 8 M urea, 100 mM Tris HCl (pH 8.0) and protease inhibitor cocktail with the assistance of sonication on ice for 10 min with a Branson Sonifier 250 from VWR Scientific (Batavia, IL). Following centrifugation (18,000 x g, 10 min), the supernatant containing extracted proteins was collected. A BCA assay was performed to determine the protein concentration using a small aliquot of the extracted proteins while leftover proteins were stored at -80°C for later use. Prior to fractionation, the *E. coli* sample (1 mg of proteins) was denatured at 37 °C for 30 min, reduced (1 μL, 2 M DTT) for 30 minutes at 37°C, alkylated (3 μL, 2 M IAA) for 20 minutes at room temperature in the dark and then quenched with 1 μL DTT (2 M) for 5-10 minutes. Then the sample was acidified with FA to get a pH $\sim$ 3.

3.2.3 HPLC

All separations were performed on a 1260 Infinity II HPLC system from Agilent (Santa Clara, CA). Detection was performed using a UV-visible detector at a wavelength of 254 nm. Data was collected and analyzed using OpenLAB software.

3.2.4 SEC separation

The SEC column (4.6×300 mm, 3 μm particles, 300 Å pores) from Agilent was used for separation of proteins. The mobile phase was 0.1% (v/v) FA, and the flow rate was 0.15 mL/min. The column temperature was kept at 40°C. One mg of *E. coli* proteins were

loaded onto the column for separation. We collected 5 fractions from 12-22 min (2 min for each fraction).

3.2.5 RPLC separation

RPLC was performed using a column with C4-bonded stationary phase (2.1 x 250 mm, 3.0 μm particles) and porous particles (300 Å) from Sepax Technologies (Newark, Delaware). Mobile phase A (2.0% ACN, 0.1% FA) and mobile phase B (0.1% FA in ACN) were used to generate gradient separation. The flow rate was 0.25 mL/min. A 90-min gradient was used for protein separation: 100% A to 80% B. The *E. coli* proteins in each SEC fraction were loaded onto the RPLC column for separation. Two different methods were used in the collection of RPLC fractions from these SEC fractions. For the first SEC fraction, we collected 40 fractions from 15 to 75 min with 1.5 min per fraction. Then we combined those 40 fractions to 20 fractions by combining adjacent fractions. Based on the UV-visible data from the first SEC fraction, we changed the way for fraction collection for SEC fractions 2-5. Twenty fractions were collected between 36 and 66 minutes with 1.5 minute per fraction. All of the RPLC fractions (100 total, 5×20) were dried down with a vacuum concentrator and the proteins in each RPLC fraction were redissolved in 5 μL of 50 mM NH4HCO3 (pH 8.0) for CZE-ESI-MS/MS analysis.

3.2.6 CZE-ESI-MS/MS

An automated CZE-ESI-MS system containing an ECE-001 CE autosampler and a commercialized electrokinetically pumped sheath flow interface from CMP Scientific (Brooklyn, NY) was used in all *E. coli* experiments [24, 25]. This online system was coupled to a QEHF mass spectrometer (Thermo Fisher Scientific, Waltham, MA). A 1-meter-long fused silica capillary (50 μm i.d., 360 μm o.d.) coated with LPA was used for CZE separation. The inner wall of the capillary was coated with LPA based on the previous works [23, 26]. One end of the capillary was etched with HF to reduce the outer

diameter of the capillary to about 70-80 μm based on the procedure described in reference [27]. (Caution: use appropriate safety procedures while handling HF solutions) The sample was injected into the capillary via applying 5-psi pressure for 95 s. The sample loading volume was about 500 nL based on the calculation using Poiseuille's law. The voltage applied at the injection end was 20 kV for separation and the ESI voltage was 2-2.3 kV. The ESI spray emitter was pulled from a glass capillary (1.0 mm o.d., 0.75 mm i.d., 10 cm long) with a Sutter P-1000 flaming/brown micropipette puller. The size of the emitter orifice was 20-40 μm. The background electrolyte (BGE) was 10% AA and the sheath buffer consisted of 0.2% FA (v/v) and 10% MeOH (v/v).

The "intact protein mode" was turned on and a trapping pressure of 0.2 was used in the QEHF mass spectrometer. DDA was used with NCE of 20% for protein fragmentation. Ion transfer capillary temperature was set to 320°C and the s-lens RF level was 55. For full MS, the number of microscans was 3, resolution was 240,000 (at m/z 200), AGC target value was 1E6, maximum injection time was 50 ms and the scan range was m/z 600-2000. For MS/MS, the number of microscans was 1 or 3, resolution was 120,000 (at m/z 200), AGC target value was 1E5, and maximum injection time was 200 ms. The top 5 or 8 most intense ions (Top 5 or Top 8 DDA) in full MS spectra were isolated with a 4 m/z window and sequentially fragmented (NCE = 20%) and an intensity threshold for triggering fragmentation of 1E5. Charge exclusion and exclude isotopes settings were turned on with proteins with charge state higher than 5 able to be fragmented. Dynamic exclusion was used with a setting of 30 s.

3.2.7 Data analysis

Electropherograms were exported using Xcalibur software from Thermo Fisher Scientific and formatted using Adobe Illustrator to make final figures. The 43 RAW files from the 43 CZE-MS/MS runs were analyzed with the TopFD and TopPIC pipeline [28]. TopFD is an improved version of MS-Deconv [29]. It converts precursor and fragment isotope clusters

into monoisotopic masses and finds possible proteoform features in CZE-MS data by combining precursor isotope clusters with similar monoisotopic masses and close migration times (the isotopic clusters may have different charge states). The 43 RAW files were converted into 43 mzXML files with msconvert [30]. Then, the spectral deconvolution was performed with TopFD to generate msalign files for database search using TopPIC (version 1.1.0). The *E. coli* (strain K12) UniProt database (UP000000625, 4307 entries, version June 7, 2017) was used for database search. The database search parameters were as follows. Cysteine carbamidomethylation was set as a fixed modification, and the maximum number of unexpected modifications was 2. The precursor and fragment mass error tolerances were 15 ppm. The spectrum-level FDR was estimated using the target-decoy approach [31]. In order to reduce the redundancy of proteoform identifications, we considered the proteoforms identified by multiple spectra as one proteoform ID if those spectra correspond to the same proteoform feature reported by TopFD or those proteoforms are from the same protein and have similar precursor masses (within 1.2 Da).

We used a two-step approach for data analysis. In the first step, we employed TopPIC to search each msalign data file against the *E. coli* proteome database separately, and no FDR filter was used in this step. In the second step, we combined all the PrSMs identified from the 43 data files and used 1% spectrum-level FDR to filter out the identified PrSMs.

3.3 Results and discussion

On the basis of the previous work, we report a multidimensional platform with high peak capacity for separation of intact proteins in complex proteomes, Figure 3.1. The proteins in an *E. coli* lysate were first fractionated with SEC into five fractions based on their size, Figure 3.1A. The proteins in each SEC fraction were further fractionated with RPLC into 20 fractions based on their hydrophobicity, resulting in 100 RPLC fractions (5×20) in total, Figure 3.1B. The proteins in those fractions were separated by the dynamic pH

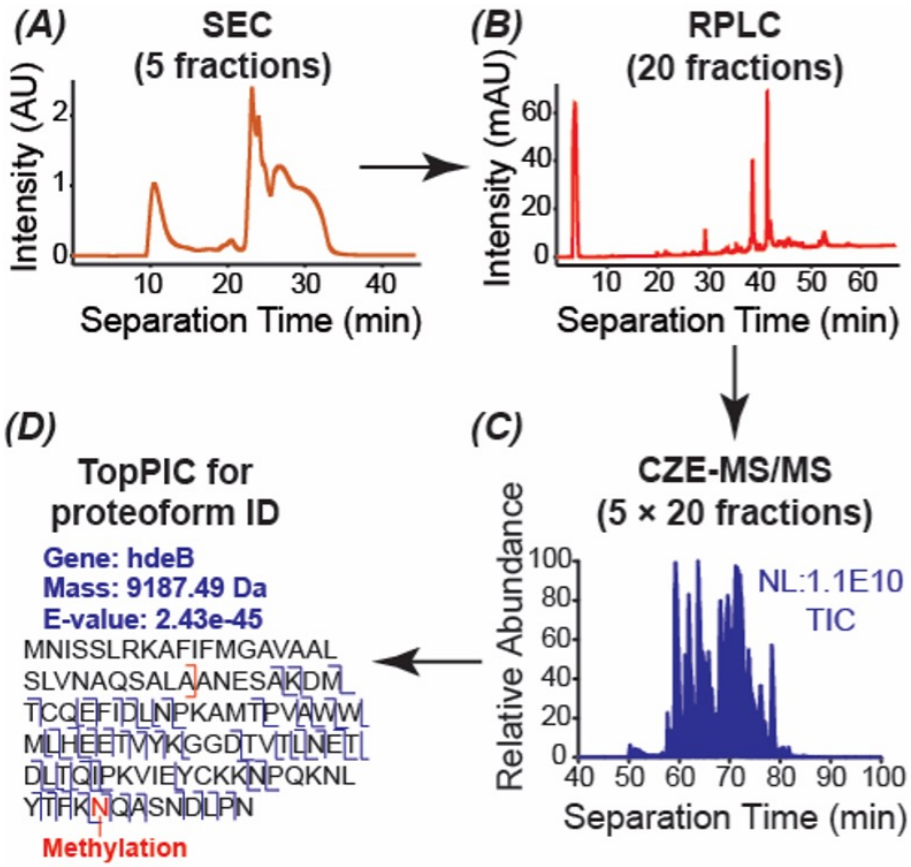


Figure 3.1: Multidimensional platform with high peak capacity for separation of intact proteins in complex proteomes. (A) SEC chromatogram of an *E. coli* lysate. (B) RPLC chromatogram of an SEC fraction of the *E. coli* lysate. (C) TIC electropherogram of an RPLC fraction of the *E. coli* lysate (D) Fragmentation pattern of one identified proteoform from gene hdeB. AU = absorbance units; mAU = milli-absorbance units; NL = normalized level. This figure is reprinted with permission from reference [1].

junction-based CZE based on their size-to-charge ratios, followed by ESI-MS/MS analysis, Figure 3.1C. The proteins in each RPLC fraction were dissolved in 5 μL of 50 mM NH4HCO3 (pH 8.0) for CZE-MS/MS. The BGE of CZE was 10% (v/v) AA (pH 2.2). The electrokinetically pumped sheath flow interface was employed to couple CZE to MS [24, 27]. About 10% of the sample (500 nL) was injected into the separation capillary for CZE-MS/MS. The SEC-RPLC-CZE platform produced orthogonal and high-capacity separation of intact proteins. The peak capacity of the platform was estimated to be around 4000 based on the full width at half-maximum (fwhm) of protein peaks. The

acquired MS/MS spectra of proteins were subjected to a database search using TopPIC software for identification and characterization of proteoforms, Figure 3.1D [28, 29].

We identified over 58000 PrSMs, 5705 proteoforms, and 850 proteins from the *E. coli* proteome using the SEC-RPLC-CZE-MS/MS platform with a 1% spectrum-level FDR. We observed reasonable protein signal from 43 RPLC fractions using CZE-MS/MS, and the proteoform/protein IDs were from those 43 CZE-MS/MS runs. Example electropherograms are shown in Figure 3.2. The data set represents an order of magnitude improvement in

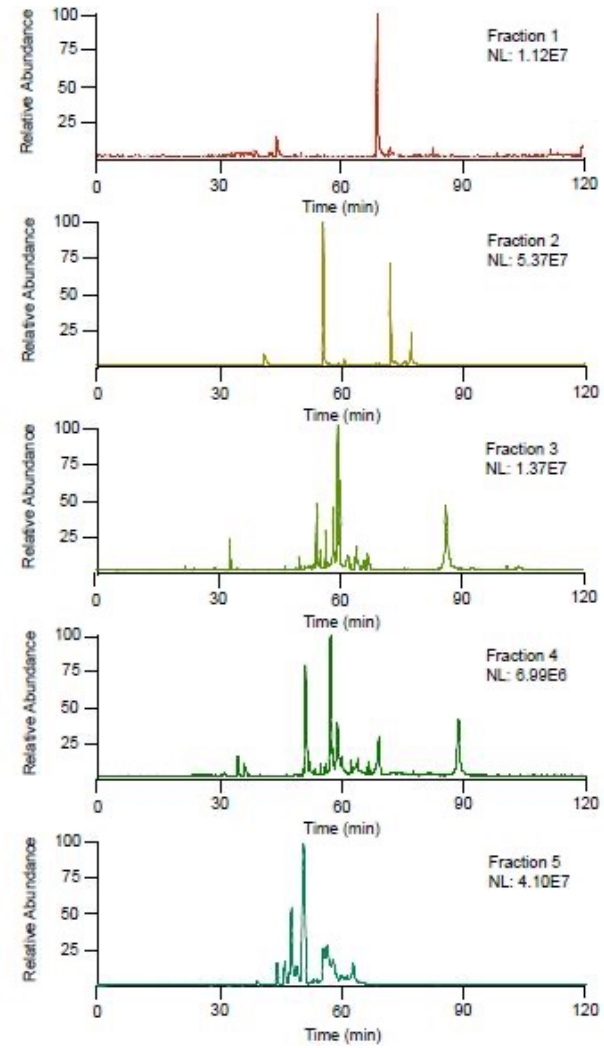


Figure 3.2: Base peak electropherograms of RPLC fractions 1-5 after CZE-MS/MS analysis. This figure is reprinted with permission from reference [1].

the number of proteoform IDs compared with previous CZE-MS/MS studies (5700 vs 300-600 proteoforms) [17, 19, 20]. The data set also represents the largest bacterial TDP data set reported to date. The details of the identified PrSMs and proteoforms are listed elsewhere [1].

We attribute the dramatic improvement in the number of proteoform IDs to two major reasons. First, the SEC-RPLC-CZE platform produced high peak capacity ($\sim 4,000$) for separation of intact proteins. The peak capacity is at least 4 times higher than that in previous TDP studies using CZE-MS/MS [17, 19, 20]. Second, the dynamic pH junction-based CZE-MS/MS system had high sample loading capacity. About 10% of the proteins in each RPLC fraction (500 nL vs 5 μL) was injected into the capillary for CZE-MS/MS, and the sample loading volume is 2-5 times higher than previous TDP studies using LC-CZE-MS/MS [17, 19]. Both the high peak capacity and high sample loading capacity benefit the identification of relatively low abundant proteins and proteoforms.

We then performed various analyses of the proteoforms and proteins that were identified from the *E. coli* proteome using the SEC-RPLC-CZE-MS/MS platform. Single-shot CZE-MS/MS produced nearly 500 proteoform IDs from two of the 43 RPLC fractions and yielded 200-400 proteoform IDs from most of the RPLC fractions, Figure 3.3A. The number of cumulative proteoform IDs increased steadily with the increase of the number of RPLC fraction or SEC fraction, indicating the efficient prefractionation performance of SEC and RPLC, Figure 3.3A. SEC fractions 3-5 made greater contribution to the proteoform IDs than SEC fraction 1, and we did not observe significant protein signal from SEC fraction 2, Figure 3.3A. The majority of the identified proteoforms had mass in a range of 10-20 kDa, and 52 proteoforms with mass bigger than 30 kDa were identified, indicating the potential of the platform for top-down characterization of large proteins, Figure 3.3B.

The number of proteoforms per gene ranged from 1 to 345, Figure 3.4. The detected

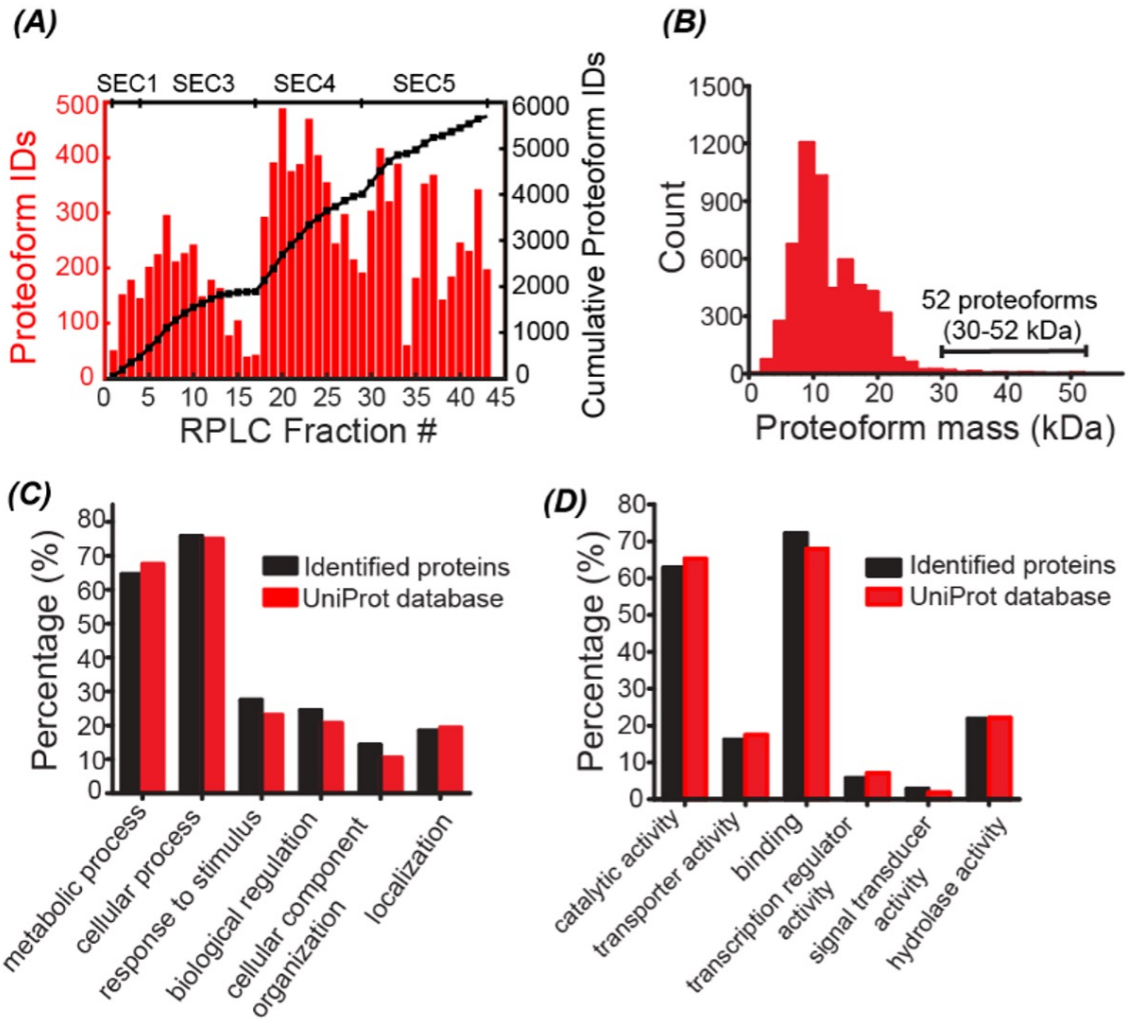


Figure 3.3: Summary of the identified proteins and proteoforms. (A) The number of proteoform IDs in each RPLC fraction (red bars); the cumulative proteoform IDs vs the number of RPLC fractions (black line with squares). (B) Mass distribution of identified proteoforms. (C) Distribution of biological processes of identified proteins in this work and proteins in the UniProt *E. coli* database. (D) Distribution of molecular functions of identified proteins in this work and the proteins in the UniProt *E. coli* database. The "Retrieve/ID mapping" tool in the UniProt Web site was used to obtain the gene ontology (GO) information on proteins. This figure is reprinted with permission from reference [1].

mass shifts from the identified proteoforms ranged from -600 to 600 Da, corresponding to various modifications, e.g., cysteine carbamidomethylation (57 Da), methylation (14 Da), acetylation (42 Da), and oxidation (16 Da), Figure 3.5. We also detected N-terminal methionine excision, signal peptide removal, and protein truncations. We observed good

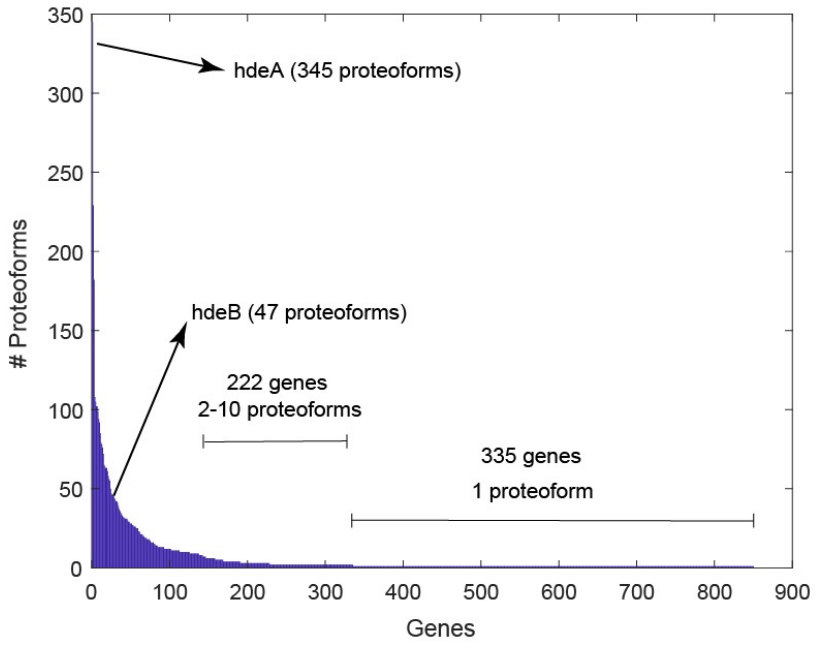


Figure 3.4: Distribution of the number of identified proteoforms from each *E. coli* gene. This figure is reprinted with permission from reference [1].

linear correlation between the number of PrSMs and the abundance (ppm) of 20 randomly selected proteins in a mass range of 6-20 kDa, Table 3.1 and Figure 3.6. Protein abundance was obtained from PaxDb (Protein Abundance Database, https://pax-db.org/). We used the *E. coli* protein abundance data derived from a BUP dataset deposited via PRIDE with the accession number PRD000485. The data suggested that the number of PrSMs of proteins (<20 kDa) could be used to roughly estimate their abundance in cells, which is similar to the spectral count idea used in BUP [32]. Similarly, we used the number of PrSMs to estimate the relative abundance of various proteoforms derived from the same gene and we took two genes, hdeA and hdeB, as the examples. We identified 345 proteoforms (6634 PrSMs) and 47 proteoforms (1084 PrSMs) for hdeA and hdeB, respectively, Figure 3.4. For hdeA, 62% of the identified proteoforms (214 out of the 345) related to various truncations at the termini of the protein molecules, and 131 proteoforms had no truncations. The data suggest that protein truncation is one major reason for the

Table 3.1: Summary of the gene, the number of PrSMs and abundance of the selected 20 proteins.

Gene	Number of PrSMs	Protein abundance (ppm)
hdeA	6634	16470
hdeB	1084	2403
rpsF	1603	2913
groS	1703	3297
wrbA	436	1933
acpP	1909	8302
greA	96	746
ygiW	536	1313
slyA	94	622
rpoZ	569	1515
yeeX	333	1844
trxA	338	1280
yjjA	48	62.7
copA	81	270
osmC	288	1152
sufE	13	18.7
ygiC	52	261
yibT	66	294
ybeL	95	235
hns	1680	7009

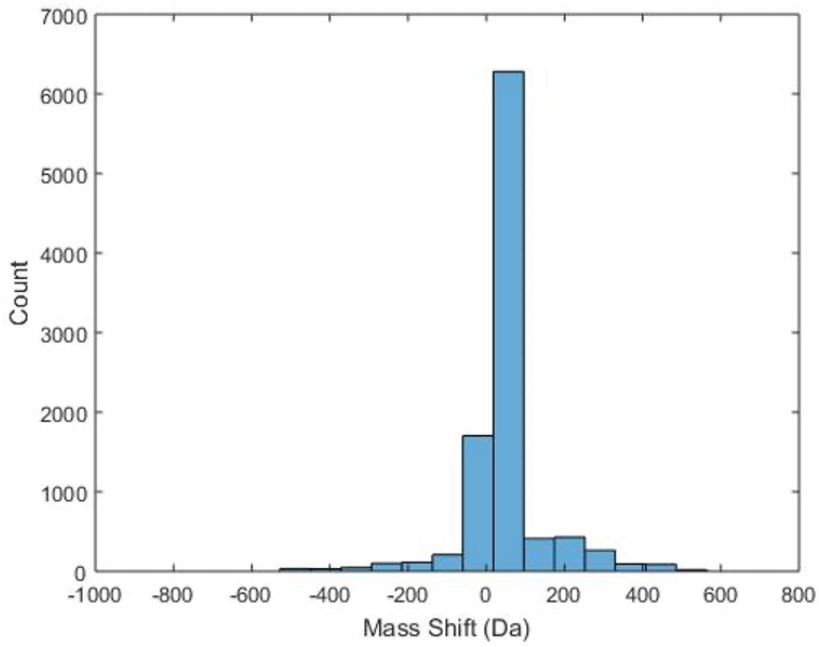


Figure 3.5: Distribution of the detected mass shifts from the identified proteoforms. This figure is reprinted with permission from reference [1].

large number of identified proteoforms of hdeA. The 131 proteoforms of hdeA that were not truncated corresponded to 87% of all the PrSMs of hdeA, and the 214 truncated proteoforms only accounted for 13% of the total PrSMs of hdeA. For hdeB, only 10% of the proteoforms (5 out of the 47) related to various truncations and those proteoforms only represented 1% of the total PrSMs. The data clearly indicate that the majority of the hdeA and hdeB protein molecules in the *E. coli* cells have no truncations. As shown in Table 3.2, the majority of the hdeA and hdeB protein molecules in *E. coli* cells had the mass shift as 0 Da based on their PrSM data. When the number of PrSMs listed in the table were calculated for different mass shifts, we did not consider whether the protein sequences were truncated or not. A small percentage of the hdeB protein molecules had methylation (mass shift as 14 Da), dimethylation (mass shift as 28 Da), acetylation (mass shift as 42 Da), or a combination of methylation and acetylation (mass shift as 56 Da), 3.1. Those PTMs of hdeB detected here agreed well with that in one *E. coli* PTM database

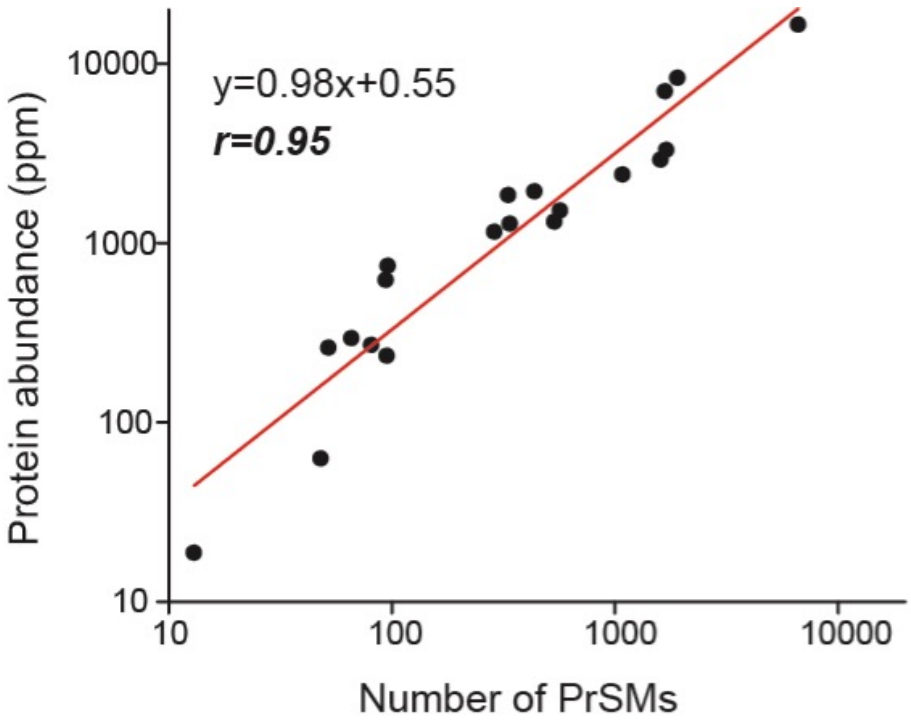


Figure 3.6: Correlation between the number of PrSMs and the abundance (ppm) of 20 randomly selected proteins with mass in a range of 6-20 kDa (log-log plot). This figure is reprinted with permission from reference [1].

established recently by the Smith group using BUP [33]. Similarly, we also identified some hdeA proteoforms with the same mass shifts as the hdeB proteoforms, e.g., 14, 28, and 42 Da. However, we could not find any PTM information about hdeA from UniProt database (http://www.uniprot.org/uniprot/P0AES9) and the *E. coli* PTM database in reference [33]. The results here highlight the capability of the CZE-MS/MS-based TDP for accurate characterization of proteins in cells.

We further compared the identified proteins (850) with the proteins in UniProt *E. coli* database ($\sim 4,000$ proteins) in terms of the gene ontology (GO) information, Figures 3.3C,D and 3.7. The detailed information about those proteins is shown in Table 3.1. Our SEC-RPLC-CZE-MS/MS platform had no obvious bias in protein ID with respect to the biological process and molecular function distributions. About 36% of the identified

Table 3.2: Summary of the number of PrSMs of various proteoforms derived from hdeA and hdeB with different mass shifts detected in the work.

Mass shift (Da)	hdeA	hdeB
0	3191	553
14	249	68
28	282	126
42	138	68
56	0	22
Others	2774	247
In total	6634	1084

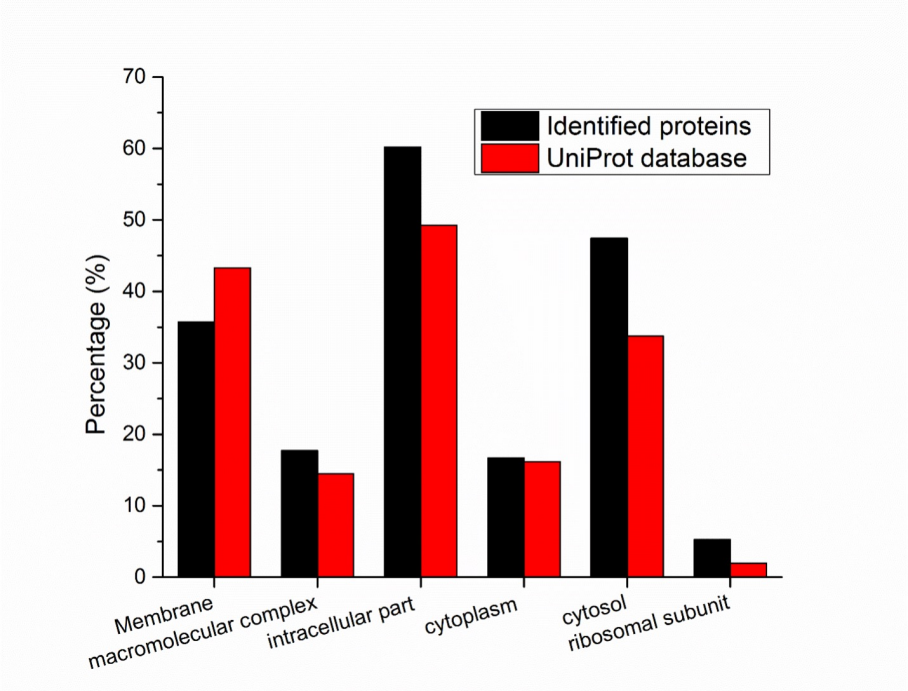


Figure 3.7: Distribution of the cellular component of the identified proteins and the proteins in the UniProt *E.coli* database using the "Retrieve/ID mapping" tool in the UniProt website. This figure is reprinted with permission from reference [1].

proteins were membrane proteins, and this percentage was only slightly lower than that in the UniProt database (43%). The data indicated that our platform was efficient for identification of membrane proteins. The percentage of proteins that located in the intracellular part, cytosol, or ribosomal subunit was higher in the identified protein pool than that in the UniProt database.

We also compared our work with recent deep TDP studies that employed RPLC as the final dimension for separation of intact proteins prior to MS and MS/MS analysis. In our work, 5705 proteoform and 850 protein IDs were observed from the 43 CZE-MS/MS runs, corresponding to roughly 4680 min of instrument time. Tran et al. combined sIEF, GELFrEE, and RPLC-MS/MS for TDP of a human cell line, resulting in over 3000 proteoform IDs from 1063 proteins with 3825 min of instrument time [5]. Anderson et al. identified 3238 proteoforms and 684 proteins from human colorectal cancer cells using GELFrEE prefractionation followed by RPLC-MS/MS [8]. Overall, the data acquisition took roughly 4960 min. Catherman et al. combined subcellular fractionation, sIEF, GELFrEE, and RPLC-MS/MS for deep TDP of the transformed human cell line H1299 proteome [6]. Over 5000 proteoforms and 1220 proteins were identified, representing the largest TDP data set of the human proteome reported to date. Hundreds of RPLC-MS/MS runs ($\sim$ 90 min per run) were performed in that study.

3.4 Conclusion

In summary, our SEC-RPLC-CZE-MS/MS platform is comparable with the state-of-the-art RPLC-MS/MS based systems for deep TDP in terms of the number of proteoform IDs and the total instrument time. It is noteworthy that the total CZE-MS/MS analysis time can be easily reduced via boosting the electric field across the separation capillary. In this work, 20 kV was applied across the capillary for separation, and increasing the voltage to 30 kV will theoretically improve the throughput by 1.5-fold. In addition, in this work, we did not fully use the instrument time for proteoform IDs, and there was significant dead time in each CZE-MS/MS run. For instance, all of the identified PrSMs concentrated in a 10 min window for the RPLC fraction 15, and the dead time of this CZE-MS/MS run was 110 min, Figure 1.12A.

As another example, the identified PrSMs spread over an 80 min window for the RPLC

fraction 19, and about 40 PrSMs/min was approached across a 35 min window, Figure 1.12B. The dead time of that CZE-MS/MS run was still 40 min. We believe the sequential sample injection method that has been tested for high-throughput BUP using CZE-MS/MS recently will allow us to reduce the dead time in each CZE-MS/MS run [34–37]. Those improvements will be very helpful to increase the throughput of our SEC-RPLC-CZE-MS/MS platform for deep TDP.

We speculate that the number of proteoform and protein IDs from the SEC-RPLC-CZE-MS/MS platform can be significantly boosted via several improvements. First, the SEC separation can be further improved via simply increasing the length of the SEC column and employing the serial SEC method developed recently by the Ge group [7]. Second, the RPLC separation can be improved via investigating different RP beads and employing longer columns [37, 38]. Third, the performance of CZE can be improved with longer separation capillaries (i.e., 1.5 m) and higher separation voltage (i.e., 60 kV). Fourth, the improvement in mass resolution and scan speed of mass spectrometers definitely will benefit large-scale TDP of complex proteomes. In, addition, the combination of different protein fragmentation techniques, e.g., HCD, ETD, and UVPD, will be invaluable for boosting the scale of TDP and improving the quality of proteoform characterization [39–44].

3.5 Acknowledgments

We thank Prof. Heedeok Hong's group at Michigan State University (Department of Chemistry) for kindly providing the *Escherichia coli* cells for our research. We are thankful for the support from the National Institute of General Medical Sciences, National Institutes of Health (NIH), through Grants R01GM118470 (X.L.) and R01GM125991 (L.S. and X.L.).

BIBLIOGRAPHY

BIBLIOGRAPHY

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

(2) Toby, T. K.; Fornelli, L.; Kelleher, N. L. *Annu. Rev. Anal. Chem.* **2016**, *9*, 499–519.

(3) Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W. *J. Am. Chem. Soc.* **1999**, *121*, 806–812.

(4) Ge, Y.; Lawhorn, B. G.; ElNaggar, M.; Strauss, E.; Park, J. H.; Begley, T. P.; McLafferty, F. W. *J. Am. Chem. Soc.* **2002**, *124*, 672–678.

(5) Tran, J. C. et al. *Nature* **2011**, *480*, 254–258.

(6) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. *Mol. Cell. Proteomics* **2013**, *12*, 3465–3473.

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

(8) Anderson, L. C. et al. *J. Proteome Res.* **2017**, *16*, 1087–1096.

(9) Jorgenson, J. W.; Lukacs, K. D. *Science* **1983**, *222*, 266–272.

(10) Valaskovic, G. A.; Kelleher, N. L.; McLafferty, F. W. *Science* **1996**, *273*, 1199–1202.

(11) Sun, L.; Knierman, M. D.; Zhu, G.; Dovichi, N. J. *Anal. Chem.* **2013**, *85*, 5989–5995.

(12) Haselberg, R.; de Jong, G. J.; Somsen, G. W. *Anal. Chem.* **2013**, *85*, 2289–2296.

(13) Bush, D. R.; Zang, L.; Belov, A. M.; Ivanov, A. R.; Karger, B. L. *J. Chromatogr. A* **2017**, *1498*, 1138–1146.

(14) Sarg, B.; Faserl, K.; Kremser, L.; Halfinger, B.; Sebastiano, R.; Lindner, H. H. *Mol. Cell Proteomics* **2013**, *12*, 2640–2656.

(15) Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; III, J. R. Y. *J. Proteome Res.* **2014**, *13*, 6078–6086.

(16) Li, Y.; Tran, J. C.; Ntai, I.; Kelleher, N. L. *Proteomics* **2014**, *14*, 1158–1164.

(17) Han, X.; Wang, Y.; Aslanian, A.; Bern, M.; Lavellée-Adam, M.; III, J. R. Y. *Anal. Chem.* **2014**, *86*, 11006–11012.

(18) Zhao, Y.; Sun, L.; Champion, M. M.; Knierman, M. D.; Dovichi, N. J. *Anal. Chem.* **2014**, *86*, 4873–4878.

(19) Zhao, Y.; Sun, L.; Zhu, G.; Dovichi, N. J. *J. Proteome Res.* **2016**, *15*, 3679–3685.

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

(21) Britz-Mckibbin, P.; Chen, D. D. Y. *Anal. Chem.* **2000**, *72*, 1242–1252.

(22) Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J. *Anal. Chem.* **2014**, *86*, 6331–6336.

(23) Chen, D.; Shen, X.; Sun, L. *Analyst* **2017**, *142*, 2118–2127.

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

(25) Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J. *J. Proteome Res.* **2015**, *14*, 2312–2321.

(26) Zhu, G.; Sun, L.; Dovichi, N. J. *Talanta* **2016**, *146*, 839–843.

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

(28) Kou, Q.; Xun, L.; Liu, X. *Bioinformatics* **2016**, *32*, 3495–3497.

(29) Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. *Mol. Cell. Proteom.* **2010**, *9*, 2772–2782.

(30) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534–2536.

(31) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207–214.

(32) Liu, H.; Sadygov, R. G.; III, J. R. Y. *Anal. Chem.* **2004**, *76*, 4193–4201.

(33) Dai, Y.; Shortreed, M. R.; Scalf, M.; Frey, B. L.; Cesnik, A. J.; Solntsev, S.; Schaffer, L. V.; Smith, L. M. *J. Proteome Res.* **2017**, *16*, 4156–4165.

(34) Faserl, K.; Sarg, B.; Sola, L.; Linder, H. H. *Proteomics* **2017**, *17*, doi.org/10.1002/pmic.201700310.

(35) Boley, D. A.; Zhang, Z.; Dovichi, N. J. *J. Chromatogr. A* **2017**, *1523*, 123–126.

(36) Garza, S.; Moini, M. *Anal. Chem.* **2006**, *78*, 7309–7316.

(37) Ansong, C. et al. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 10153–10158.

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

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

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

(41) Swaney, D. L.; McAlister, G. C.; Wirtala, M.; Schwartz, J. C.; Syka, J. E.; Coon, J. J. *Anal. Chem.* **2007**, *79*, 477–485.

(42) Xia, Y.; Han, H.; McLuckey, S. A. *Anal. Chem.* **2008**, *80*, 1111–1117.

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

(44) O'Brien, J. P.; Li, W.; Zhang, Y.; Brodbelt, J. S. *J. Am. Chem. Soc.* **2014**, *136*, 12920–12928.

Chapter 4

Large-scale top-down proteomics of human colorectal cancer cell lines using capillary zone electrophoresis-tandem mass spectrometry

4.1 Introduction

As the third most common cancer in the world and due to its high mortality rate, colorectal cancer (CRC) has drawn considerable attention from the medical community [1–3]. Metastasis is the main cause of CRC death and gaining insight into the main actors in cells behind CRC progression is highly desirable to develop new drug targets and for choices related to patient care and treatment regimen [4–6]. A comprehensive list of genes that are known to be associated with unfavorable prognosis in CRC are provided through The Human Protein Atlas [1]. However, cellular behavior during cancer progression is the result of subtle changes at the protein level and genetic analysis does not provide a clear picture of how proteins are regulated or protein level information, such as PTMs, that also play a key role in CRC metastasis [2, 4, 7, 8]. Hummon, et al. identified phosphorylated proteins upregulated during CRC metastasis by analyzing SW480 and SW620 (primary and metastatic) cell lines [5]. Ghosh, et al. utilized iTRAQ for a complete BUP profile of

SW480 and SW620 cell lines [9]. However, BUP is limited by the protein inference problem where it is not possible to identify or quantify specific proteoforms that may contain multiple PTMs or combinations of different modifications, including those with SAAVs and other genetic alterations [10]. Herein, we provide a clear picture of the diversity of modifications and proteoform family variation at play during CRC metastasis in the largest TDP study of CRC to date. RPLC-MS is traditionally used in proteomics experiments, however, our TDP platform utilizes CZE-MS/MS which is a recently proven method for in-depth analysis of proteoforms within complex proteomes [11–15]. We also combine LC prefractionation with CZE-MS/MS for increased separation efficiency of proteoforms and to ease the burden on the mass spectrometer. High separation efficiency coupled to the resolving power and speed of mass analysis routinely allows for analysis of thousands of proteoforms, many of which are low abundant but still play key roles in biological processes [11, 16–20].

4.2 Experimental

4.2.1 Materials and reagents

MS-grade water, ACN, MeOH, FA and HPLC-grade (AA) were purchased from Fisher Scientific (Pittsburgh, PA). NH4HCO3, urea, DTT, IAA and 3-(trimethoxysilyl)propyl methacrylate were from Sigma-Aldrich (St. Louis, MO). HF (48-51% solution in water) and acrylamide were purchased from Acros Organics (NJ, USA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ). Complete, mini protease inhibitor cocktail (EASYpacks) was from Roche (Indianapolis, IN).

4.2.2 Sample preparation

Cell lysis buffer consisted of 8M urea. 50 mM (pH 8.2), 1 mM β-glycerophosphate, 1 mM PMSF, 75 mM NaCl, 1 mM NaF, 1 mM sodium orthovanadate, 10 mM sodium pyrophosphate, and protease inhibitor cocktail. Protein concentrations were determined by BCA assay. SW480 and SW620 proteins were denatured at 37°C for 30 minutes, reduced at 37°C for 30 minutes after adding 1 M DTT, and then alkylated at room temperature in the dark for 20 minutes after adding 1M IAA. The reactions were quenched by adding 1M DTT for 5 minutes at room temperature.

For study 1, 200 μg of protein for SW480 and SW620 were reduced, alkylated, and acidified prior to being sent directly through RPLC fractionation. For study 2, 2 mg of protein for SW480 and SW620 were reduced and alkylated prior to being sent directly through multidimensional separation. For study 3, 420 μg of both SW480 and SW620 were reduced and alkylated prior to separation by RPLC. For study 4, the samples were desalted after reduction and alkylation using a C4 trap column (4×10 mm, 3 μm particles, 300 Å pore size). Specifically, 500 μg of protein from SW480 and SW620 was loaded onto the column and flushed with mobile phase A (2% (v/v) ACN, 0.1% FA) for 10 minutes at a flow rate of 1 mL/min. The proteins were eluted with mobile phase B (80% ACN, 0.1% FA) for 3 minutes at flow rate of 1 mL/min. The eluates were lyophilized with a speed vacuum and redissolved in 150 μL 0.1% FA.

4.2.3 Fractionation of the SW480 and SW620 proteome

All separations were performed on a 1260 Infinity II HPLC system from Agilent (Santa Clara, CA). Detection was performed using a UV-visible detector at a wavelength of 254 nm. Data was collected and analyzed using OpenLAB software. RPLC (C4, 2.1 x 250 mm, Sepax Technologies) and SEC (4.6 x 300 mm, 500 Å pores, Agilent) were performed offline (Agilent HPLC) for prefractionation. Fractions from SW620 and SW480 from study 1

(13×2), study 2 (52×2), study 3 (6×2), and study 4 (6×2) were analyzed by CZE-MS/MS, respectively.

In study 1, a 0.25 mL/min flow rate and gradient of 0-80% mobile phase (MP) B over 90 minutes (MPA: 2% ACN, 0.1% FA in water MPB: 80% ACN, 0.1% FA in water) were used. Fractions were collected from 15 to 22 minutes (fraction 1) and 22 to 70 minutes (12, 4-minute fractions). For study 2, both SEC and RPLC were used for fractionation prior to CZE-MS/MS. For SEC, the flow rate was 0.35 mL/min with a 0.05% TFA mobile phase. A 20 μL loading volume (40×) was used for more efficient SEC separation with fractions being combined from successive separations. Fractions were collected from 5-8 minutes (fraction 1) and 8-12.5 minutes (3, 1.5-minute fractions). One RPLC run was performed for each SEC fraction with a flow rate of 0.25 mL/min and gradient of 0-80% MPB (MPA: 2% ACN, 0.1% TFA in water MPB: 10% IPA, 0.1% TFA in ACN) over 90 minutes with a 10-minute flush with 100% MPA at the beginning of the separation. Fractions were collected from 20 to 25 minutes (fraction 1) and 25 to 65 minutes (20, 2-min fractions). In study 3, the same mobile phases were used as in study 1, and a 90-minute gradient was used with a 10-minute flush with 100% MPA at the beginning of the separation. Fractions were collected from 25 to 55 minutes (fraction 1), 50 to 70 minutes (4, 5-minute fractions), and 70 to 95 minutes (fraction 6). In study 4 SEC fractionation was performed with an Agilent Bio SEC-5 column (4.6 × 300 mm, 5 μm particles, 500 Å pore size). 220 μg each of SW 480 and SW 620 (1.5 mg/mL, 75 μL×2 injections) proteins were loaded into the SEC column and separated isocratically at the flow rate of 0.3 mL/min with 0.1% FA as mobile phase. The first fraction is collected from 5.6 to 8.6 minutes. The second to the fifth fraction was from 8.6 to 14.6 minutes with 1.5 minutes per fraction. The final fraction was collected from 14.6 to 19.0 min. In studies 1-3, samples were dried down and redissolved in 50 mM NH4HCO3 (pH 8.0, $\sim$ 2 mg/mL) for CZE-ESI-MS/MS. In study 4, fractions were dried down and redissolved in 20 μL of 10 mM ammonium bicarbonate for a concentration of $\sim$ 2 mg/mL.

4.2.4 CZE-MS/MS

CZE separation was performed using a CESI 8000 Plus CE system (Beckman Coulter). A commercialized electrokinetically pumped sheath-flow CE-MS nanospray interface (CMP Scientific Corp) was applied for online coupling the CE system and mass spectrometer [21, 22]. A glass emitter (orifice size: 20-30 μm) installed on the interface was filled with sheath buffer (0.2% FA, 10% MeOH) to generate electrospray at voltage of 2-2.3 kV.

In studies 1, 2, and 4 an LPA-coated fused silica capillary (1 m, 50 μm i.d., 360 μm o.d.) was used with BGE (5% AA). In study 3, a 70 cm capillary was used with all other parameters being the same as studies 1, 2, and 4. The inner wall of the capillary was coated with LPA based on the procedure described in references [23, 24]. One end of the capillary was etched with HF to reduce the outer diameter of the capillary to about 70-80 μm based on the procedure described in reference [25]. (Caution: use appropriate safety procedures while handling HF solutions)

In study 1 and 2, the capillary was loaded with 500 nL of sample using 5 psi for 95 seconds. In study 3, the 70 cm capillary was loaded with $\sim$ 350 nL of sample and in study 4, the capillary was loaded with 500 nL of sample. After sample loading, the capillaries were inserted into background electrolyte, containing 5% AA, and 30 kV voltage was applied at the sample injection end to carry out separations. In study 1, 30 kV was applied for 100 minutes, followed by a 10-minute flush and in study 2, 30 kV was applied for 70 minutes for separation. In study 3 and 4, 30 kV was applied for 70 and 100 minutes for separation, respectively.

MS1 and MS2 data were collected on a QEHF mass spectrometer (Thermo Fisher Scientific) under DDA mode. The temperature of ion transfer tube was set to 320°C and s-lens RF was 55. MS1 spectra were collected with following parameters: m/z range of 600-2000, mass resolution of 120,000 (at m/z 200), a microscan number of 3, AGC target value of 1E6, and maximum injection time of 100 ms. The top 5 high abundant precursor

ions (charge state higher than 3, 5, or charge state unassigned and intensity threshold 2E4) in the MS1 spectra were isolated with a window of 4 m/z and fragmented via HCD with NCE of 20%. The settings for MS2 spectra were m/z range of 300 to 2000, resolution of 120,000 (at m/z 200), a microscan number of 3, AGC target value of 1E5, and maximum injection time of 200 ms. The dynamic exclusion was set to a duration of 30 s and the isotopic peaks were excluded. Each fraction was analyzed three times using CZE-MS/MS.

4.2.5 Data analysis

The 422 RAW files, that contained significant protein signal from the CZE-MS/MS runs, were analyzed with the TopFD and TopPIC pipeline [26]. TopFD is an improved version of MS-Deconv [27]. It converts precursor and fragment isotope clusters into monoisotopic masses and finds possible proteoform features in CZE-MS data by combining precursor isotope clusters with similar monoisotopic masses and close migration times (the isotopic clusters may have different charge states). The 422 RAW files were converted into 422 mzXML files with msconvert [28]. Then, spectral deconvolution was performed with TopFD to generate msalign files for database search using TopPIC (version 1.1.0). The human UniProt database (UP000005640, 77027 entries, version October, 23, 2019) was used for database search with TopPIC (version 1.4.0). The database search parameters were as follows. Cysteine carbamidomethylation was set as a fixed modification, and the maximum number of unexpected modifications was 2. The precursor and fragment mass error tolerances were 15 ppm. The maximum mass shift of unknown modifications was 500 Da. The spectrum-level FDR was estimated using the target-decoy approach and was set to 1% and the proteoform-level FDR was also set to either 1% or 5%.9 If set to 5%, proteoforms were filtered to have proteoform-level FDRs of less than 1%.

4.3 Results and discussion

Briefly, this work consists of four separate studies, each consisting of a slightly different analytical approach. Offline fractionation was performed in all four studies, with study 1 and 3 consisting of RPLC fractionation with 13 and 6 fractions collected respectively. Study 2 consisted of SEC and RPLC fractionation with 52 total fractions collected and study 4 consisted of SEC fractionation with 6 fractions collected. Each sample was analyzed by CZE-MS/MS and studies 2-4 were ran in triplicate. The 422 raw files accumulated during this work represent 516 hours of instrument time, mostly from study 2 (378 hours).

A summary of identification results from this work is provided in Figure 4.1A with a combined total of 23319 and 2297 proteoform and protein identifications at 1% proteoform-level FDR, respectively across studies 1-4. This is an over 400% increase, in terms of proteoform identifications, from any TDP study reported to date, Figure 4.1B. Despite requiring the most instrument time, study 2 resulted in nearly as many identifications as study 4 on its own, highlighting the give-and-take between protein concentration, solubility, and efficiency of prefractionation. The high number of proteoforms after combination of study 3 and study 4 (11374) highlights the complementary nature of both RPLC and SEC. Study 1 resulted in 7545 proteoform identifications.

Using the combined list of proteoforms, we identified proteoforms from genes which are among those listed in The Human Protein Atlas to be directly related to unfavorable outcomes in CRC. This includes proteoforms from the gene AKAP8L which has a prognostic p-value of 3.49E-4 for colorectal cancer according to The Human Protein Atlas. The particular proteoforms shown in Figure 4.1C, show the presence of both a phosphorylated and unphosphorylated AKAP8L proteoform in the SW480 cell line. Existence of phosphorylation on S601 on AKAP8L is further confirmed through

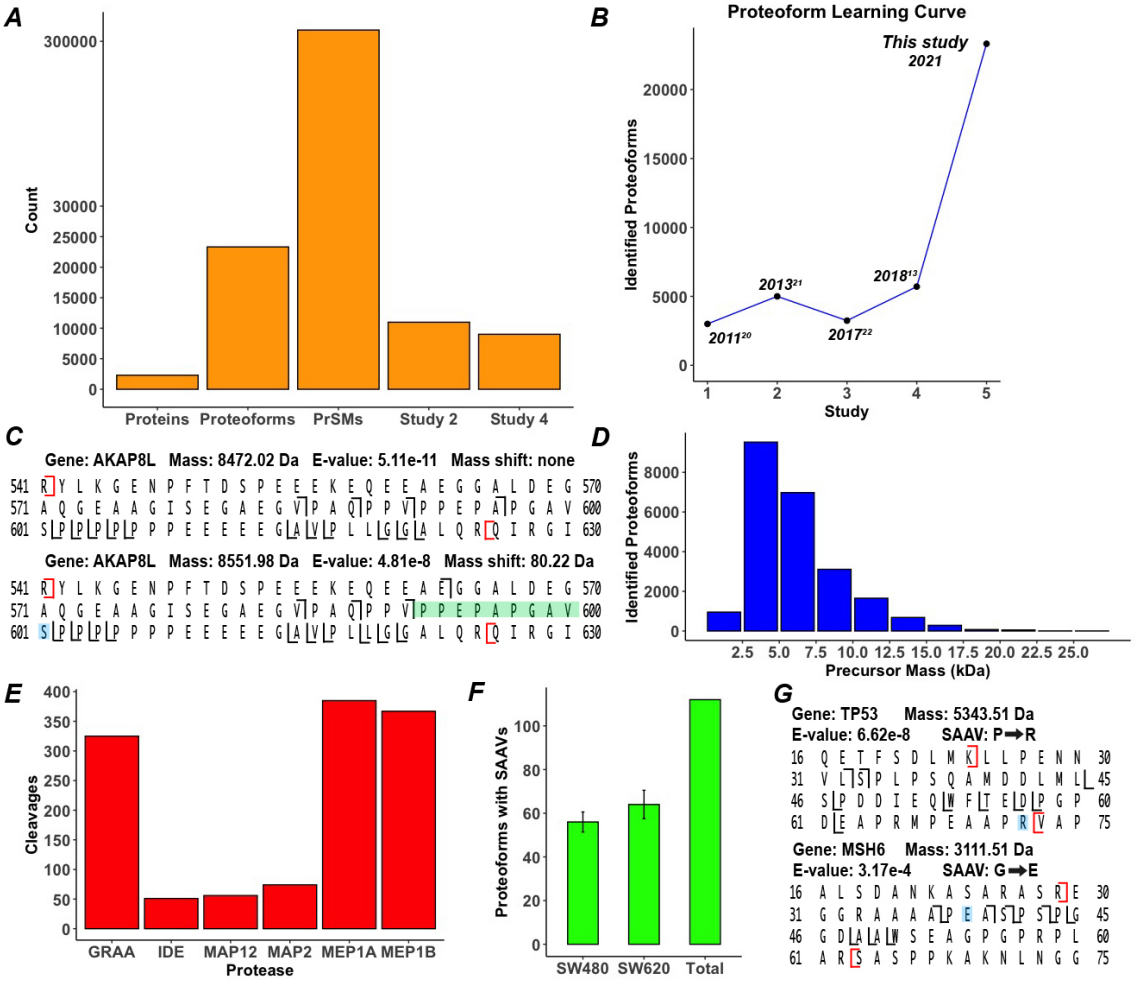


Figure 4.1: A, Graph showing an overview of identifications in this study. B, a plot of the proteoform learning curve, comparing the number of proteoform identifications in this study to other TDP studies. C, fragmentation patterns of two proteoforms from the AKAP8L gene. D, a histogram showing the molecular weight distribution for all proteoforms in this study. E, graph of various active proteases using TopFINDer. F, graph of the proteoforms with SAAVs for SW480, SW620, and the combined number of proteoforms. G, fragmentation patterns of TP53 and MSH6 proteoforms that contain SAAVs.

PhosphoSitePlus (version 6.5.9.3) [29]. As AKAP8L is already confirmed to be associated with an unfavorable prognosis in CRC, it is significant that we find the presence of these two proteoforms, as phosphorylation is also known to be an important factor in disease progression [1, 5]. Therefore, an additional layer of information is provided from analysis of these proteoforms by TDP that could ultimately lead to drug development or a different patient treatment regimen, significantly impacting CRC outcome.

As in most large-scale TDP experiments, this study was limited to mostly small proteoforms, with the complete mass distribution, shown in Figure 4.1D, filtered at 1% proteoform-level FDR. The relatively narrow mass range of top-down experiments is due to a combination of factors including limited separation efficiency for large proteoforms, particularly during fractionation using LC, the protein signal being spread over many charge states after ESI, and resolution and mass range limitations associated with the mass spectrometer. In addition to instrumentation limitations, there is a high number of proteoform fragments present within our dataset, which has long been a point of contention in TDP experiments. We believed that this could possibly be due to enzymatic degradation prior to sample preparation, and utilized the TopFINDer tool to confirm enzymatic activity [30]. A high number of cleaved proteoforms were confirmed through TopFINDer, Figure 4.1E, with high confidence and the cleavage enzyme responsible is also provided. This figure shows the proteases with the highest number of cleavages and a q-value less than 5%, demonstrating high confidence. The q-value represents the percentage of cleavages that are expected to be false positives, and a significant value is chosen by TopFINDer. Manual examination of the dataset also confirmed the presence of complementary proteoform sequences to the cleaved sequences, further validating these results. Complete protease information is shown in Table 4.1.

TDP is also aided by advances in other -omics fields, including analysis of the transcriptome, which can give information on expected mutations in proteoforms [31]. This study uses mRNA level data to match to our TDP dataset to confirm the presence of mutations at the proteoform-level [32]. Using a curated database, we were able to identify SAAVs present in SW480 and SW620, Figure 4.1F. Proteoforms with SAAVs from genes TP53 and MSH6 are shown in Figure 4.1G. According to many studies, TP53 mutations are recognized as a hallmark of colorectal cancer, and the regulation of TP53 is directly related to patient outcome in CRC [33, 34]. MSH6 mutations are used as a marker for clinical diagnosis of Lynch syndrome which is known to increase a patient's risk of

Table 4.1: Complete protease information from TopFINDer.

Protease	Accession	List count	Adjusted Fisher Exact Test (q-value)
MEP1A	Q16819	385	1.26E-17
MAP12	Q6UB28	56	6.65E-20
GRAM	P51124	239	8.40E-01
GRAA	P12544	325	2.84E-29
MEP1B	Q16820	367	4.91E-05
CASP6	P55212	20	1.00E+00
MAP2	P50579	74	1.41E-31
MPPB	O75439	15	2.65E-02
GRAB	P10144	273	1.00E+00
CASP3	P42574	56	1.00E+00
IDE	P14735	51	2.08E-07
CATS	P25774	173	1.00E+00
CATB	P07858	105	1.00E+00
CATL1	P07711	176	1.00E+00
MMP14	P50281	10	5.75E-01
HTRA2	O43464	29	1.00E+00
FURIN	P09958	3	1.00E+00
CASP1	P29466	15	1.00E+00
PPCE	P48147	13	1.14E-05
CMA1	P23946	5	7.75E-01
CASP7	P55210	6	1.00E+00
MMP7	P09237	5	1.00E+00
CASP8	Q14790	2	1.00E+00
KLK5	Q9Y337	1	1.00E+00
CAN1	P07384	1	1.00E+00
PARL	Q9H300	1	1.00E+00
UROK	P00749	1	1.00E+00
GRAK	P49863	1	1.00E+00
CBPA6	Q8N4T0	2	7.75E-01
MMP2	P08253	2	1.00E+00
MMP9	P14780	2	1.00E+00
MMP12	P39900	1	1.00E+00
MMP25	Q9NPA2	1	1.00E+00
MMP3	P08254	1	1.00E+00
MMP26	Q9NRE1	1	1.00E+00
MMP1	P03956	1	1.00E+00
MMP8	P22894	1	1.00E+00
ELNE	P08246	2	1.00E+00
TRY3	P35030	1	1.00E+00
KLK3	P07288	1	1.00E+00
LGMN	Q99538	2	1.00E+00
GRAH	P20718	2	7.75E-01
CATD	P07339	1	1.00E+00

developing CRC and other cancers [35]. Therefore, the ability of TDP to detect mutated TP53 and MSH6 proteoforms directly, possibly with modifications that further impact disease progression and outcome, is a significant upper hand in a clinical setting when it comes to treating patients and for drugs that target these genes.

A quantitative analysis was performed with the TopDiff (version 1.3.4) tool, based on proteoforms reported by TopPIC, available in the TopPIC suite, for the data obtained with SEC-CZE-MS/MS and RPLC-CZE-MS/MS (study 3 and 4). Quantitative data provides differentially expressed proteoforms that can be further analyzed for biological significance within the context of CRC. Differentially expressed proteoforms in study 4 can be visualized in Figure 4.2A, with 460 proteoforms that have significant differences between SW480 and SW620. Pink dots have higher abundance in SW480 and blue dots have higher abundance in SW620. GO analysis was performed for study 3 and 4 and select biological processes, with p-values less than 0.01, are shown in Figure 4.2B. Several of the differentially expressed proteoforms with significant difference between SW480 and SW620 also contained phosphorylations. The genes associated with these proteoforms, along with a log transformation of the ratio of their abundance in SW480 and SW620 and their p-values are shown in Figure 4.2C. Complete proteoform information from Figure 4.2C is shown in Table 4.2.

Interestingly, there are two proteoforms from the death-associated protein (DAP) gene that are either higher abundant in SW480 (positive ratio) or SW620 (negative ratio). The DAP gene is integral to programmed cell death and is therefore critical to cancer progression [36]. A closer look at the fragmentation pattern and localization of the phosphorylation modification in these proteoforms is shown in Figure 4.2D. The DAP proteoform that is higher abundant in SW620 has a better localized phosphorylation site on the proteoform sequence than the proteoform that is higher abundant in SW480, in addition to being a smaller proteoform fragment. However, the phosphorylation site for the

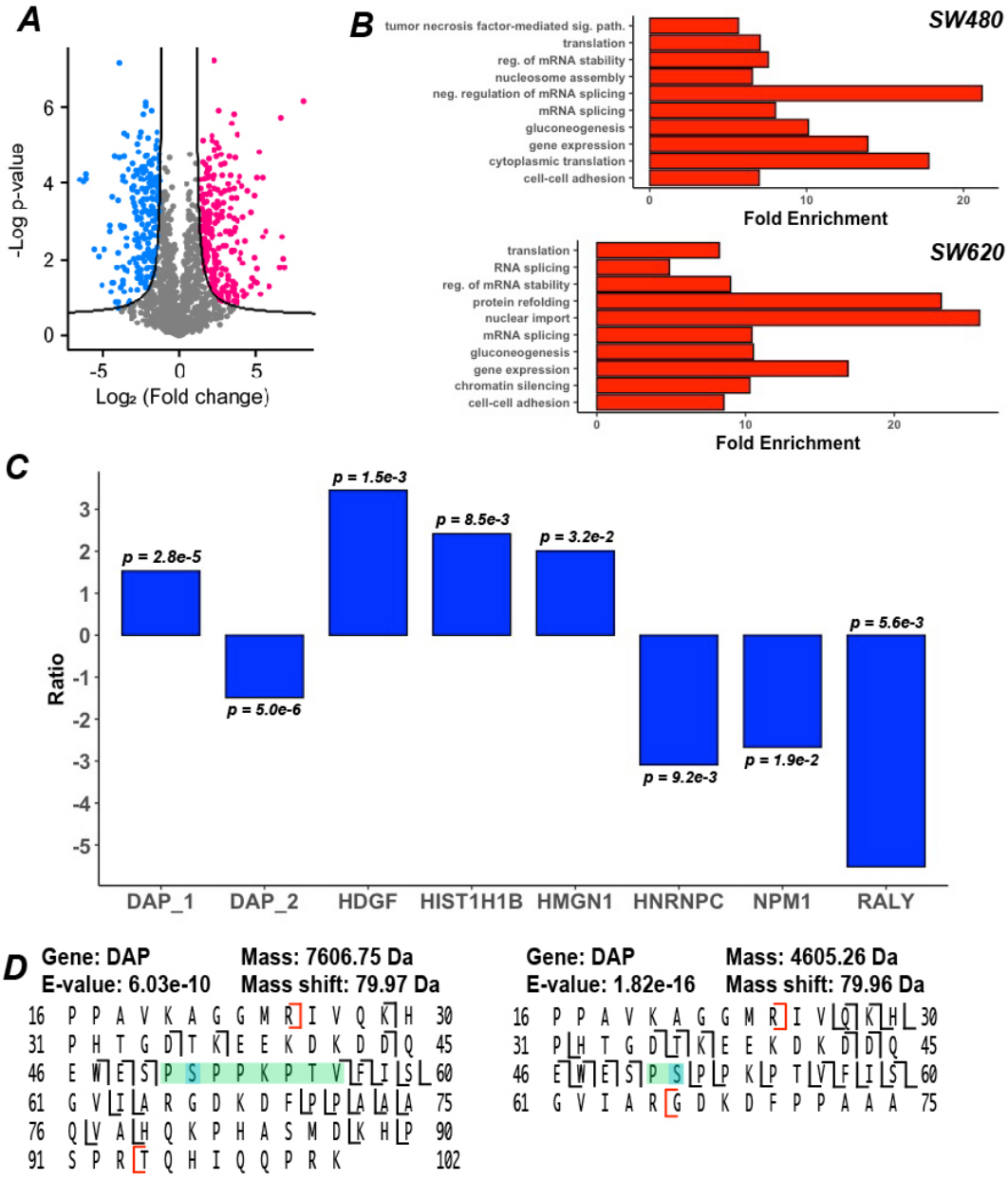


Figure 4.2: A, volcano plot showing differentially expressed proteoforms with S0 = 1 and FDR = 0.05. B, graph showing a select few biological processes and fold enrichment after GO analysis of proteoforms quantified in study 3 and study 4 for SW480 and SW620. C, graph of genes with quantified proteoforms from study 4 that also contain phosphorylations. D, fragmentation patterns for phosphorylated proteoforms from the DAP gene that are differentially expressed in SW480 and SW620.

DAP proteoform in SW620 also includes a T56, in addition to S51. Both of these sites are known to be phosphorylated according to PhosphoSitePlus, with S51 phosphorylation

Table 4.2: Proteoform information from Figure 4.2C. Proteoform ratio is given as the log_2(average abundance in SW480/average abundance in SW620). Proteoform sequences are provided with the part of the sequence that contains the phosphorylation in parentheses.

Gene	Ratio	Proteoform Sequence
DAP	1.53	R.IVQKHPHTGDTKEEKDKDDQEWES(PSPPKPTV)[79.9696]FIS GVIARGDKDFPPAAAQVAHQKPHASMDKHPSPR.T
DAP	-1.49	R.IVQKHPHTGDTKEEKDKDDQEWES(PS)[79.9692]PPKPTVFIS GVIAR.G
HDGF	3.45	R.AGDLLED(SPK)[79.9689]RPKEAENPEGEEKEAATLEVERPLP MEVEKNSTPSEPGSGRGPPQEEEEEEDEEEEATKEDAEAPGIR DHESL.
HIST1H1B	2.42	M.(S)[Acetyl]ETAPAETATPA(PVEKS)[79.9702]PAKKKATK.K
HMGN1	2.01	K.QAEVANQETKEDLPAEN(GETKTEESPAS)[159.9318]DEAGE KEAKSD.
HNRNPC	-3.09	R.SAAEMYGSVTEH(PS)[79.9690]PSPLLSSSFDLDYDFQRDYY DR.M
NPM1	-2.67	K.(C)[Carbamidomethylation]GSGPVHISGQHLVAVEEDAE (SE)[79.9682]DEEEEDVKLLSISGKR.S
RALY	-5.52	R.TRDDGDEEGLLTH(SEEELE)[79.9695]HSQDTDADDGALQ.

being the most common phosphorylation site. Ideally, more efficient fragmentation would be able to localize this modification better, but this is a perfect example of what TDP has to offer in terms of identifying proteoforms from the same gene that otherwise would not be differentiated by BUP or any other -omics method.

Other interesting proteoforms that are differentiated by their abundance in SW480 and SW620 includes proteoforms from the nucleophosmin (NPM1) gene and hepatoma-derived growth factor (HDGF). Both of these genes are known to play significant roles in cancer, and phosphorylated proteoforms from both genes are differentially expressed in CRC, Figure 4.2C [37, 38]. Another proteoform from HDGF, that was quantified without significant expression difference between SW480 and SW620, also contains a possible double phosphorylation event (159.9 Da mass shift). This event was localized to part of the proteoform sequence that contains S132 and S133 that are both commonly phosphorylated in HDGF according to PhosphoSitePlus.

4.4 Conclusion

Big data produced by TDP experiments poses a unique challenge to analytical chemists attempting to get a complete picture about the state of the biological system being analyzed. As a result, TDP has not carved out a niche in a clinical setting although more targeted TDP analyses are not far from clinical application. However, in the lab we can piece together results from a variety of input/outputs ranging from the analytical methods being used to specialized TDP software to gain some biologically relevant insight. This will involve moving away somewhat but not completely from characterizing the brute strength of TDP in terms of identifying large numbers of proteoforms and proteins and into completing other incomplete, but very useful, pictures of biological processes created from other -omics studies. In this study we offered a comprehensive top-down analysis of two human CRC cell lines, SW480 and SW620, resulting in the largest TDP dataset and irreplaceable insight into the biological state of primary and metastatic CRC. Drastic proteoform differences between SW480 and SW620 make it apparent that greater attention needs to be paid to the clinical ramifications caused by tumor proteoform heterogeneity between patients.

4.5 Acknowledgments

We thank Professor Amanda Hummon and her lab for generously providing the CRC cancer cell line samples. This project was funded by Michigan State University. We thank the support from the National Science Foundation (CAREER Award, Grant DBI1846913) and the National Institutes of Health (Grant R01GM125991).

BIBLIOGRAPHY

BIBLIOGRAPHY

(1) The Human Protein Atlas `https://www.proteinatlas.org/`.

(2) Schmitt, M.; Greten, F. R. *Nat. Rev. Immunol.* **2021**.

(3) Rehman, S. K. et al. *Cell* **2021**, *184*, 226–242.e21.

(4) Markowitz, S. D.; Bertagnolli, M. M. *N. Engl. J. Med.* **2009**, *361*, 2449–2460.

(5) Schunter, A. J.; Yue, X.; Hummon, A. B. *Anal. Bional. Chem.* **2017**, *409*, 1749–1763.

(6) Zhang, B. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 256–268.

(7) National Cancer Institute, Clinical Proteomic Tumor Analysis Consortium `https://proteomics.cancer.gov/programs/cptac`.

(8) Lee, J.; Mckinney, K. Q.; Pavlopoulos, A. J.; Park, J.; Hwang, S. *J. Proteom.* **2015**, 326–336.

(9) Ghosh, D.; Yu, H.; Tan, X. F.; Lim, T. K.; Zubaidah, R. M.; Tan, H. T.; Chung, M. C. M. *J. Proteome. Res.* **2011**, *10*, 4373–4387.

(10) Nesvizhskii, A. I.; Aebersold, R. *Mol. Cell. Proteom.* **2005**, *4*, 1419–1440.

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

(12) McCool, E. N.; Sun, L. *Se Pu* **2019**, *37*, 878–886.

(13) Zhao, Y.; Sun, L.; Zhu, G.; Dovichi, N. J. *J. Proteome Res.* **2016**, *15*, 3679–3685.

(14) Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L. *Trends Anal. Chem.* **2019**, *120*, 115644.

(15) McCool, E. N.; Lubeckyj, R.; Shen, X.; Kou, Q.; Liu, X.; Sun, L. *J. Vis. Exp.* **2018**, *120*, 115644.

(16) McCool, E. N.; Lodge, J. M.; Basharat, A. R.; Liu, X.; Coon, J. J.; Sun, L. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 2470–2479.

(17) McCool, E. N.; Chen, D.; Li, W.; Liu, Y.; Sun, L. *Anal. Methods* **2019**, *11*, 2855–2861.

(18) Tran, J. C. et al. *Nature* **2011**, *480*, 254–258.

(19) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. *Mol. Cell. Proteomics* **2013**, *12*, 3465–3473.

(20) Anderson, L. C. et al. *J. Proteome Res.* **2017**, *16*, 1087–1096.

(21) Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2554–2560.

(22) Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J. *J. Proteome Res.* **2015**, *14*, 2312–2321.

(23) Zhu, G.; Sun, L.; Dovichi, N. J. *Talanta* **2016**, *146*, 839–843.

(24) Chen, D.; Shen, X.; Sun, L. *Analyst* **2017**, *142*, 2118–2127.

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

(26) Kou, Q.; Xun, L.; Liu, X. *Bioinformatics* **2016**, *32*, 3495–3497.

(27) Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. *Mol. Cell. Proteom.* **2010**, *9*, 2772–2782.

(28) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534–2536.

(29) PhosphoSitePlus Cell Signaling Technology, `https://www.phosphosite.org/homeAction`.

(30) Fortelny, N.; Yang, S.; Pavlidis, P.; Lange, P. F.; Overall, C. M. *Nucleic Acids Res.* **2015**, *43*, D290–D297.

(31) Smith, L. M.; Kelleher, N. L. *Science* **2018**, *359*, 1106–1107.

(32) Chen, D.; Yang, Z.; Shen, X.; Sun, L. *Anal. Chem.* **2021**, *93*, 4417–4424.

(33) WIlliams, D. S. et al. *Mod. Pathol.* **2020**, *33*, 483–495.

(34) Nguyen, T.; Menendez, D.; Resnick, M. A.; Anderson, C. W. *Hum. Mutat.* **2014**, *35*, 738–755.

(35) Hendriks, Y. M. C. et al. *Gastroenterology* **2004**, *127*, 17–25.

(36) Levy-Strumpf, N.; Kimchi, A. *Oncogene* **2004**, *17*, 3331–3340.

(37) Grisendi, S.; Mecucci, C.; Falini, B.; Pandolfi, P. P. *Nat. Rev. Cancer* **2006**, *6*, 493–505.

(38) Liao, F.; Dong, W.; Fan, L. *Med. Oncol.* **2010**, *27*, 1219–1226.

Chapter 5

Fragmentation of intact proteins using activated ion electron transfer dissociation and ultraviolet photodissociation

5.1 Introduction

TDP aims to characterize proteoforms in their intact state and often in complex protein mixtures, with many advantages and disadvantages which have been discussed thoroughly to this point [1–4]. [1] The ability to elucidate these forms of biological variation is vital for understanding the roles played by proteoforms in disease and development [7, 8]. The state-of-the-art RPLC-MS/MS-based systems have achieved identification, and even quantification, of thousands of proteoforms from complex samples [9–17]. Much effort has been made to improve the separation of proteoforms with RPLC. Monolithic columns and packed columns with beads having various sizes, different lengths of carbon chains, varied porosity, and longer columns have been investigated for proteoform separation [10, 18–20].

Also, the recent improvements in CE-MS interface have facilitated the CZE-MS/MS for TDP [21–23]. The McLafferty group reported identifications (IDs) of intact proteins using

[1] This chapter was adapted with permission from references [5, 6].

CZE-MS with attomole amounts of materials in 1996 [24]. The Yates group demonstrated that CZE-MS achieved similar signal-to-noise ratios to RPLC-MS for analysis of a protein complex sample with 100-fold less sample consumption [25]. The Dovichi group has reported 600 proteoform IDs using RPLC-CZE-MS/MS from yeast cells [26]. And, for large protein characterization, the Kelleher group identified 30 proteins with masses in a range of 30-80 kDa from *Pseudomonas aeruginosa* PA01 cell lysate using CZE-MS/MS [27], demonstrating the potential of CZE-MS/MS for characterization of large proteins.

Challenges remain for large-scale TDP using CZE-MS/MS, including the narrow separation window (typically 30 min) and low sample loading capacity (low nL) of CZE. Recently, our group achieved a 90-min separation window and a 1-μL sample loading volume using CZE-MS for analysis of an *E. coli* cell lysate, leading to IDs of 600 proteoforms in a single CZE-MS/MS run [14, 28]. We employed a separation capillary with high-quality LPA coating on its inner wall to eliminate EOF in the capillary, widening the separation window [29]. We used a protein stacking method, dynamic pH junction, for highly efficient and online concentration of proteins in the capillary, boosting the sample loading volume [30, 31]. We coupled SEC-RPLC fractionation to the dynamic pH junction-based CZE-MS/MS for deep TDP of *E. coli* cells [15]. Nearly 6000 proteoforms and 850 proteins were identified using the multidimensional system. The dynamic pH junction-based CZE-MS/MS has established the foundation of TDP using CZE-MS/MS.

Extensive fragmentation of proteoforms in the gas phase is another challenge in TDP. Collision-based dissociation methods (e.g., HCD) are widely used for fragmentation of proteoforms [9–12, 18, 19, 26, 27, 32, 33]. However, HCD often fails to provide extensive fragmentation of proteoforms, and has preferential cleavage sites [34–37], limiting its utility for thorough characterization of proteoforms. Alternative fragmentation methods are vital for TDP.

Direct dissociation methods, without or concurrent with internal energy redistribution

prior to dissociation, include ETD, AI-ETD, and UVPD [17, 36, 38]. AI-ETD disrupts non-covalent interactions, that can hold fragments together following fragmentation by ETD, by using infrared radiation to heat the ion [17]. AI-ETD has been used for more efficient fragmentation of proteoforms in high-throughput TDP studies [17]. UVPD utilizes high-energy photons to heat the proteins and directly dissociate along the backbone, generating a wide variety of fragments and superior sequence coverage [38]. Close to complete sequence coverage of intact proteins using UVPD (193 nm) has been demonstrated, allowing for in-depth characterization of proteins [38]. However, 213 nm UVPD is the only commercially available method currently.

Recently, an AI-ETD method that combines infrared photoactivation concurrent with ETD has been developed and systematically evaluated for fragmentation of intact proteins [17, 39–42]. RPLC-AI-ETD has been evaluated for high-throughput top-down characterization of intact proteins (less than 20 kDa) in human CRC cells with a production of 935 proteoforms and 295 proteins [17]. More importantly, AI-ETD showed better performance than HCD and standard ETD regarding sequence coverage of identified proteoforms and proteoform characterization scores. CZE has also been coupled with AI-ETD for top-down characterization of a standard protein mixture and a bacterial secretome sample [42]. About 40 proteoforms were identified using the CZE-AI-ETD from the secretome sample, and other results have demonstrated a good complementarity of HCD and AI-ETD for intact protein fragmentation. UVPD (193 nm) has been compared to HCD for high-throughput TDP in a recent study and was reported to yield better average proteoform sequence coverage compared to HCD [16]. Coupling size-based protein fractionation to RPLC-UVPD led to 153 protein and 489 proteoform IDs from a HeLa cell lysate [16]. RPLC-UVPD (193 nm) has also been applied for top-down characterization of histones and BUP [43, 44].

In this work, for the first time, we coupled the dynamic pH junction-based CZE to

AI-ETD and UVPD (213 nm), Figure 5.1, on an Orbitrap Fusion Lumos mass spectrometer for large-scale TDP. An *E. coli* cell lysate was employed to evaluate the

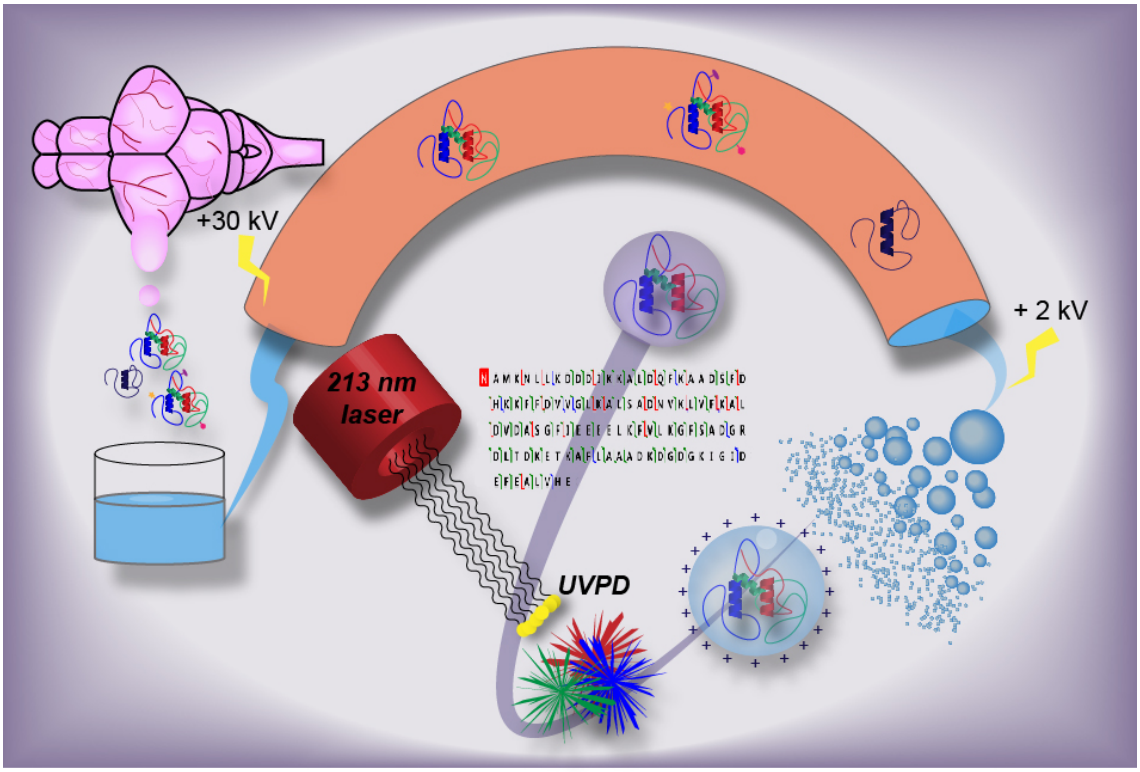


Figure 5.1: A figure of the workflow in the CZE-UVPD study. This figure is reproduced with permission from reference [6].

performance of the system. First, we investigated how the laser power used for the AI-ETD influenced the proteoform IDs. Then, we compared CZE-AI-ETD and CZE-ETD, as well as CZE-AI-ETD and CZE-HCD, for TDP of the *E. coli* cells. After that, we optimized the electric field and the DDA method for the CZE-AI-ETD system. After evaluating the reproducibility of the CZE-AI-ETD system, we coupled SEC fractionation to CZE-AI-ETD for large-scale TDP of the *E. coli* cells. For UVPD, a commercialized UVPD source with a 213-nm laser was used and optimized in the experiment. SEC-based fractionation was then coupled to CZE-UVPD for large-scale top-down proteomics of a zebrafish brain sample.

5.2 Experimental

5.2.1 Materials and reagents

MS-grade water, ACN, MeOH, FA and HPLC-grade AA were purchased from Fisher Scientific (Pittsburgh, PA). NH4HCO3, urea, ammonium persulfate and 3-(trimethoxysilyl)propyl methacrylate were from Sigma-Aldrich (St. Louis, MO). HF (48-51% solution in water) and acrylamide were purchased from Acros Organics (NJ, USA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ). Complete, mini protease inhibitor cocktail and phosphatase inhibitor cocktail were bought from Roche (Indianapolis, IN).

5.2.2 Sample preparation

E. coli (K-12 MG1655) was cultured in Lysogeny broth (LB) medium (37°C) while shaking (225 rpm) until the OD600 reached 0.7. *E. coli* cells were harvested through centrifugation (4,000 rpm, 10 min.) and washed three times with PBS. The *E. coli* cells were lysed in a lysis buffer containing 8 M urea, phosphatase inhibitor and protease inhibitor cocktail with the assistance of sonication on ice for 15 min with a Branson Sonifier 250 from VWR Scientific (Batavia, IL) after homogenization with a Homogenizer 150 from Fisher Scientific (Pittsburgh, PA). Following centrifugation (18,000 x g, 20 min), the supernatant containing extracted proteins was collected. A BCA assay was performed to determine the protein concentration using a small aliquot of the extracted proteins while leftover proteins were stored at -80°C for later use. The *E. coli* sample ($\sim$ 780 μg of proteins) was desalted using a C4 trap column (Bio-C4, 3 μm, 300 Å, 4.0 mm i.d., 10 mm long) from Sepax Technologies, Inc. (Newark, DE) on a 1260 Infinity II HPLC system from Agilent (Santa Clara, CA). The eluate containing the *E. coli* proteins was collected and lyophilized. The *E. coli* protein sample was redissolved in 50 mM NH4HCO3 (pH 8.0), and an aliquot

($\sim$ 117 μg) with $\sim$ 2 mg/mL protein concentration was used for the single-shot CZE-MS/MS experiments. The leftover *E. coli* proteins ($\sim$ 663 μg) was fractionated with SEC, followed by CZE-MS/MS analyses.

Zebrafish brain samples were kindly provided by Professor Jose Cibelli's group at the Department of Animal Science of Michigan State University. The whole protocol related to the zebrafish were performed in compliance with relevant laws or guidelines, and the protocol followed guidelines defined by the Institutional Animal Care and Use Committee of Michigan State University. Using zebrafish for scientific research has been approved by the Institutional Animal Care and Use Committee of Michigan State University. Male zebrafish brains were lysed in 8 M urea and 100 mM NH4HCO3 (pH 8.0) containing complete protease inhibitor cocktail and PhosSTOP (EASYpacks) from Roche (Indianapolis, IN). Homogenization with a Homogenizer 150 from Fisher Scientific (Pittsburgh, PA) and sonication with a Branson Sonifier 250 from VWR Scientific (Batavia, IL) were performed on ice for protein extraction. Samples were then centrifuged at 15,000 x g for 10 minutes to effectively separate lipids and cell debris from the proteins. The supernatant was collected for further preparation. The protein concentration was determined using a BCA assay. 0.7 mg of zebrafish brain proteins were reduced with DTT (1 M, 2 μL/mg of proteins) at 37°C for 30 minutes, alkylated with IAA (1 M, 5 μL/mg of proteins) at room temperature for 20 minutes in dark, and quenched with DTT (1M, 2 μL) before SEC-CZE-MS/MS analyses.

5.2.3 SEC prefractionation

The SEC fractionation was performed on a 1260 Infinity II HPLC system from Agilent (Santa Clara, CA). Detection was performed using a UV-visible detector at a wavelength of 254 nm. Data was collected and analyzed using OpenLAB software. An SEC column (4.6 x 300 mm, 3 μm particles, 300 Å pores) from Agilent was used for separation of proteins. The mobile phase was 0.1% (v/v) FA, and the flow rate was 0.15 mL/min. The column

temperature was kept at 40°C. $\sim$ 663 μg of *E. coli* proteins were loaded onto the column for separation. We collected one fraction during 11 to 12 minutes, one fraction per half a minute during 12 to 18 minutes, and one fraction during 18 to 21 minutes. All 0.7 mg of zebrafish brain sample was loaded onto the SEC column for fraction collection. One fraction was collected from 8 to 11 minutes and then 1-minute/fraction from 11 to 20 minutes was performed. Each fraction was dried down with a vacuum concentrator and redissolved in 50 mM NH4HCO3 (pH 8.0) with a final concentration of $\sim$ 2 mg/mL for CZE-MS/MS analysis.

5.2.4 CZE-ESI-MS and MS/MS

An ECE-001 CE autosampler and a commercialized electrokinetically pumped sheath flow CE-MS interface from CMP Scientific (Brooklyn, NY) were used in all experiments [22, 23]. The automated CE system was coupled to an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, Waltham, MA) via the electrokinetically pumped sheath flow interface. A 1-m-long fused silica capillary (50 μm i.d., 360 μm o.d.) with LPA coating on the inner wall was used for CZE separation. The LPA coating was made based on the procedure in references [28, 29]. One end of the capillary was etched with HF to reduce the outer diameter of the capillary to $\sim$ 70-80 μm based on the procedure described in reference [28].

The sample was injected into the capillary via applying 5 psi for 95 s corresponding to a 500 nL volume based on the Poiseuille's law. The separation voltage applied at the injection end was either 10 kV for 240 min, 20 kV for 120 min, or 30 kV for 90 min in the CZE-AIETD experiments and 20 kV for 120 min for CZE-UVPD experiments. Between CZE-MS/MS runs, the separation capillary was flushed with BGE using 10 psi for 10 min. For optimizing the laser power for AI-ETD and comparing the AI-ETD and ETD, 30 kV for 90 min was applied. For comparing AI-ETD and HCD, optimizing the DDA method, and evaluating the reproducibility of CZE-AI-ETD, 20 kV for 120 min was applied. For

analyzing the SEC fractions of the *E. coli* proteome, 20 kV for 120 min was used for the first nine fractions and 30 kV for 120 min for the last five fractions.

The ESI voltage was 2-2.3 kV. The ESI spray emitter was pulled from a glass capillary (1.0-mm o.d., 0.75-mm i.d., 10-cm long) with a Sutter P-1000 flaming/brown micropipette puller. The size of the emitter orifice was 20-40 μm. The BGE for CZE was 5% (v/v) AA (pH 2.4) and the sheath buffer for ESI consisted of 0.2% (v/v) FA and 10% (v/v) MeOH.

An Orbitrap Fusion Lumos mass spectrometer was used for all experiments. For all experiments, DDA was utilized with intact protein mode turned on, advanced peak determination set to true, and default charge state set to 10. The ion transfer tube temperature was set to 275°C. Charge exclusion and exclude isotopes settings were turned on for proteins with charge state between 7 and 24 able to be fragmented. Include undetermined charges states was set to false and include charge states 25 and higher was set to true. Dynamic exclusion was used with a setting of 30 s. The same MS settings were used for all experiments. Use wide quad isolation was set to true, the orbitrap resolution was 120,000, AGC target was 500,000, the number of microscans was 4, and the RF lens (%) was 60.

The scan range, m/z 600-2000, and maximum injection time, 50 ms, was the same for all CZE-AI-ETD experiments. For optimizing the laser power for AI-ETD (12, 18, 24, and 30 W) and comparing AI-ETD and ETD, a top 2 DDA method was used. The option for performing a dependent scan on a single charge state per precursor was set to false. For MS/MS, the isolation window was set to 3, orbitrap resolution was 60,000, maximum injection time was 118 ms, AGC target was 500,000, and the number of microscans was 4. For AI-ETD and ETD, the ETD reaction time was set to 20 ms, ETD reagent target was 700,000, and maximum ETD reagent injection time was 200 ms. For optimizing the DDA methods (top N) for AI-ETD (18 W laser power), top 2, top 4, and top 5 DDA methods were investigated. The option to perform a dependent scan on a single charge state per

precursor only was set to true. For optimizing the separation voltage, the AI-ETD method (18 W laser power) including a top 2 DDA method was used. For analyzing the SEC fractions of the *E. coli* proteome, AI-ETD with 18 W laser power and a top 4 DDA method were employed. The option to perform a dependent scan on a single charge state per precursor only was set to true. For comparing the AI-ETD (18 W) and HCD, top 5 DDA methods were used for both AI-ETD and HCD. The details of MS/MS with AI-ETD were the same as that described above. For MS/MS with HCD, a normalized collision energy 20% was used for fragmentation. Other parameters were the same as that for AI-ETD.

For CZE-UVPD experiments, a top 5 DDA method was used for acquiring MS/MS spectra of proteoforms. The setting for performing dependent scan on single charge state per precursors only was set to False. An isolation window of 0.5 m/z was used for all experiments. The orbitrap scan range was m/z 150-2000 and maximum injection time was 246 ms. Low and high mass tolerance was set to 1.5 m/z.

5.2.5 Data analysis

For the raw files from single-shot analyses of the whole *E. coli* cell lysate in the CZE-AIETD experiment and for zebrafish brain samples in the CZE-UVPD experiment, we employed the Proteome Discoverer 2.2 (Thermo Fisher Scientific) with the ProSight PD Top Down High/High node for database search [45]. Briefly, MS/MS spectra of proteoforms were deconvoluted with Xtract (signal-to-noise ratio threshold of three) and searched against the whole *E. coli* or zebrafish database downloaded from the http://proteinaceous. net/database-warehouse-legacy/. A three-tier search was used. Tier one consisted of an absolute mass search with 2.0 Da precursor mass tolerance and 10 ppm fragment ion mass tolerance. Tier two contained a biomarker search with 10 ppm precursor mass tolerance and 10 ppm fragment ion mass tolerance. Tier three had an absolute mass search with 1000 Da precursor mass tolerance and 10 ppm fragment ion mass tolerance. [45] Only b- and y-types of fragment ions were considered for HCD fragmentation; only c-

and z-types of fragment ions were considered for ETD and AI-ETD fragmentation. Fragments considered for UVPD included a and x, b and y, and c and z ions. The target-decoy approach was used to evaluate the FDRs of PrSMs and proteoform IDs [46, 47]. A 1% spectrum-level FDR was used to filter the PrSMs and a 5% proteoform-level FDR was used to filter the proteoform IDs.

Although the ProSight PD node integrated in Proteome Discoverer 2.2 can perform database search of individual raw files, it cannot combine raw files from different SEC fractions from the CZE-AIETD of zebrafish brain samples for database search. Therefore, the total numbers of protein and proteoform IDs from the SEC-CZE-MS/MS experiment were from our manual combinations of proteoform and protein IDs from each SEC fraction with the removal of redundant proteoforms and proteins.

For the raw files from the fractionated *E. coli* sample using SEC in the CZE-AIETD experiment, we employed the TopFD and TopPIC pipeline for database search [48]. The 14 raw files corresponding to the 14 SEC fractions were analyzed. First, the 14 raw files were converted into 14 mzML files with the Msconvert tool [49]. Then, the spectral deconvolution was performed with TopFD to generate msalign files for database search using TopPIC (version 1.2.2). The *E. coli* (strain K12) UniProt database (UP000000625, 4313 entries, version June 28, 2018) was used for database search. The database search parameters were as follows. The maximum number of unexpected modifications was 2. The precursor and fragment mass error tolerances were 15 ppm. The maximum mass shift of unknown modifications was 500 Da. The FDRs were estimated using the target-decoy approach [46, 47] To reduce the redundancy of proteoform IDs, we reviewed the proteoforms that were identified by multiple MS/MS spectra as one same proteoform ID if these MS/MS spectra corresponded to the same proteoform feature reported by the TopFD or those proteoforms were from one same protein and had smaller than 1.2-Da mass differences.

Two steps of analyses were performed. In the first step, we used TopPIC to search each raw file against the *E. coli* proteome database separately. In the second step, we combined all the PrSMs identified from the 14 data files and filtered the PrSM IDs with a 1% spectrum-level FDR. The proteoform IDs were then filtered with a 5% proteoform-level FDR. A list of identified proteoforms can be found elsewhere [5].

5.3 Results and discussion

5.3.1 Comparing CZE-AI-ETD, CZE-ETD, and CZE-HCD

We first optimized the laser power for AI-ETD before comparing it with ETD and HCD. The laser power can significantly affect the performance of AI-ETD based on a very recent RPLC-AI-ETD report [17]. Here, we evaluated the performance of CZE-AI-ETD with four different laser powers: 12 W, 18 W, 24 W, and 30 W, Figure 5.2. The CZE-MS system

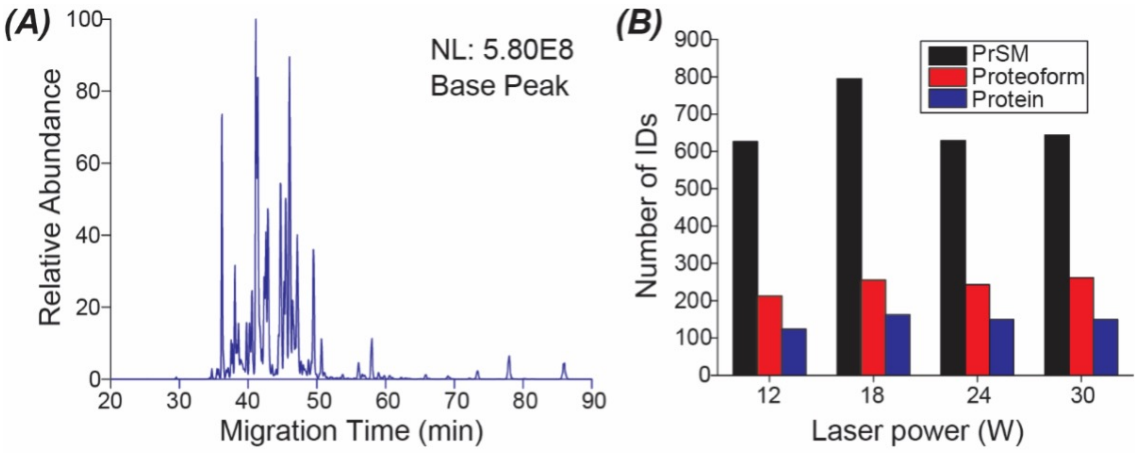


Figure 5.2: (A) Base peak electropherogram of the *E. coli* proteome after CZE-MS analysis. (B) Numbers of identified proteoform spectrum matches (PrSMs), proteoforms, and proteins using CZE-AI-ETD with four different laser powers. This figure is reproduced with permission from reference [5].

obtained a 1-h separation window and reasonably good signal (NL: 5.8E8) with only 1 μg of *E. coli* proteins. CZE-AI-ETD with the 18-W laser power produced over 20% and 8%

more PrSMs and proteins than other three laser powers. The 18-W laser power generated over 5% more proteoform IDs than 12-W and 24-W laser powers and yielded similar proteoform IDs to the 30-W laser power.

We then compared the performance of CZE-AI-ETD (18 W) and CZE-ETD for top-down characterization of the *E. coli* proteome. CZE-AI-ETD identified about 12% more proteoforms and proteins compared to CZE-ETD (Figure 5.3a). More importantly,

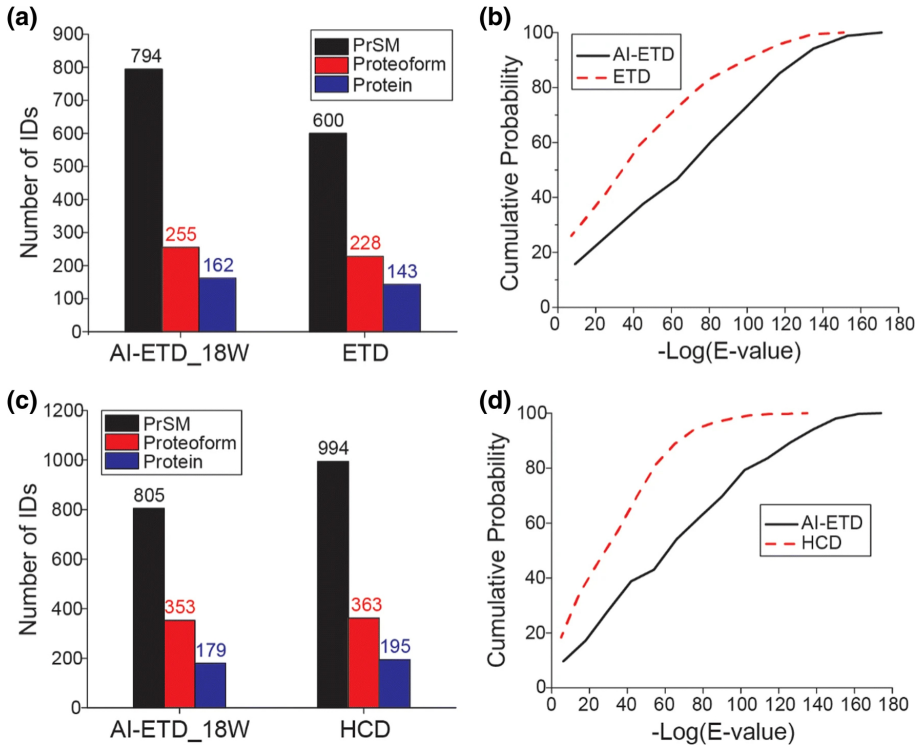


Figure 5.3: Summary of the comparisons between CZE-AI-ETD and CZE-ETD as well as CZE-AI-ETD and CZE-HCD. (a) Number of IDs from AI-ETD (18 W) and ETD. (b) Distribution of –log (E value) of identified proteoforms using AI-ETD (18 W) and ETD. (c) Number of IDs from AI-ETD (18 W) and HCD. (d) Distribution of –log (E value) of identified proteoforms using AI-ETD (18 W) and HCD. This figure is reproduced with permission from reference [5].

CZE-AI-ETD tended to obtain better expectation values (E values) of proteoform IDs than CZE-ETD (Figure 5.3b). E value represents a nonlinear transformation of the number of matching fragment ions in a spectrum. The data suggest that AI-ETD can produce better fragmentation of proteoforms compared to ETD. As shown in Figure 5.4, AI-ETD (18 W)

yielded much better residue cleavage and a much higher number of matching fragment ions than ETD (52% vs. 8%; 73 vs. 9 fragment ions) for thioredoxin 1. The disulfide bond was

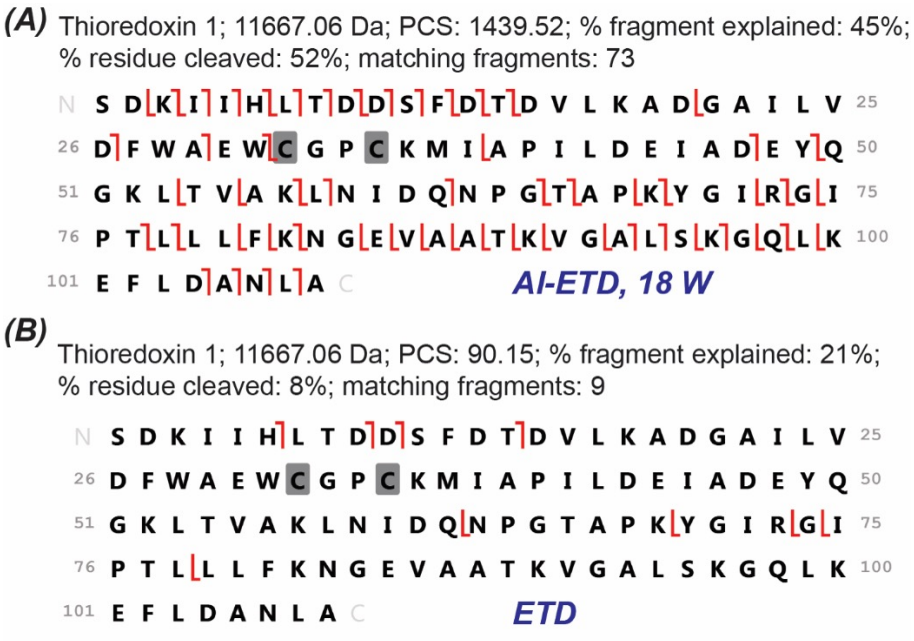


Figure 5.4: Sequences and fragmentation patterns of thioredoxin 1 observed with AI-ETD (A) and ETD (B). This figure was reproduced with permission from reference [5].

localized accurately based on the fragment ions from AI-ETD, with the two cysteine residues marked in gray forming a disulfide bond in Figure 5.4. The 18 W laser power was used in all following AI-ETD experiments.

We further compared the CZE-AI-ETD (18 W) with CZE-HCD. Single-shot CZE-HCD identified 994 PrSMs, 363 proteoforms, and 195 proteins from the *E. coli* sample. CZE-HCD produced a moderate increase in PrSMs and slightly better proteoform and protein IDs than CZE-AI-ETD (Figure 5.3c). In the experiments, CZE-HCD generated 50% more MS/MS spectra than CZE-AI-ETD per 120-min analysis but resulted in only minor improvement in the number of proteoform and protein IDs. Interestingly, CZE-AI-ETD inclined to gain better E values of proteoform IDs than HCD (Figure 5.3d). We need to note that different CZE separation conditions (30 kV for 90 min vs. 20 kV for 120 min) and MS/ MS conditions (top 2 vs. top 5 DDA methods) were used for the

experiments for Figure 5.3a and Figure 5.3c, leading to significant differences in the number of IDs from the CZE-AI-ETD.

5.3.2 Optimizing the CZE-AI-ETD method for TDP

We optimized the CZE separation voltage and the maximum number of MS/MS spectra followed by one MS spectrum in the DDA method (top N). A high separation voltage shortens the analysis time but produces a limited number of MS/MS spectra for proteoform IDs. A low separation voltage slows down the separation, allowing the acquisition of a large number of MS/ MS spectra for proteoform IDs. However, the low separation voltage results in wider protein peaks and lower protein signal, which certainly affects the quality of MS/MS spectra. The top N method in DDA influences the number of proteoform IDs because of the production of different numbers of MS/MS spectra.

When the separation voltage of CZE was changed from 30 to 10 kV, the analysis required much longer time, and the protein signal decreased significantly (Figure 5.5a). CZE with 20 kV separation voltage produced better separation efficiency than 30 kV and 10 kV (Figure 5.5b). The mass tolerance for peak extraction was 20 ppm and Gaussian smoothing (5 points) was applied in Figure 5.5b and N is labeled in the figure. The separation efficiency was up to half a million for one proteoform (m/z 775.05, charge + 9). CZE-AI-ETD with 20 kV voltage generated more proteoform IDs than that with 10 kV and 30 kV voltages (292 vs. 278 or 255) (Figure 5.5c). Interestingly, CZE-AI-ETD with 30 kV voltage identified 5% and 12% more proteins than that with 20 kV and 10 kV voltage, respectively. CZE-AI-ETD with 10 kV separation voltage gained the highest number of PrSMs, most likely due to the wider proteoform peaks generated using the lower voltage. The 20 kV separation voltage was employed in the following experiments.

We then optimized the DDA method by comparing the proteoform and protein IDs from top 2, top 4, and top 5 methods. The top 4 method identified 384 proteoforms and

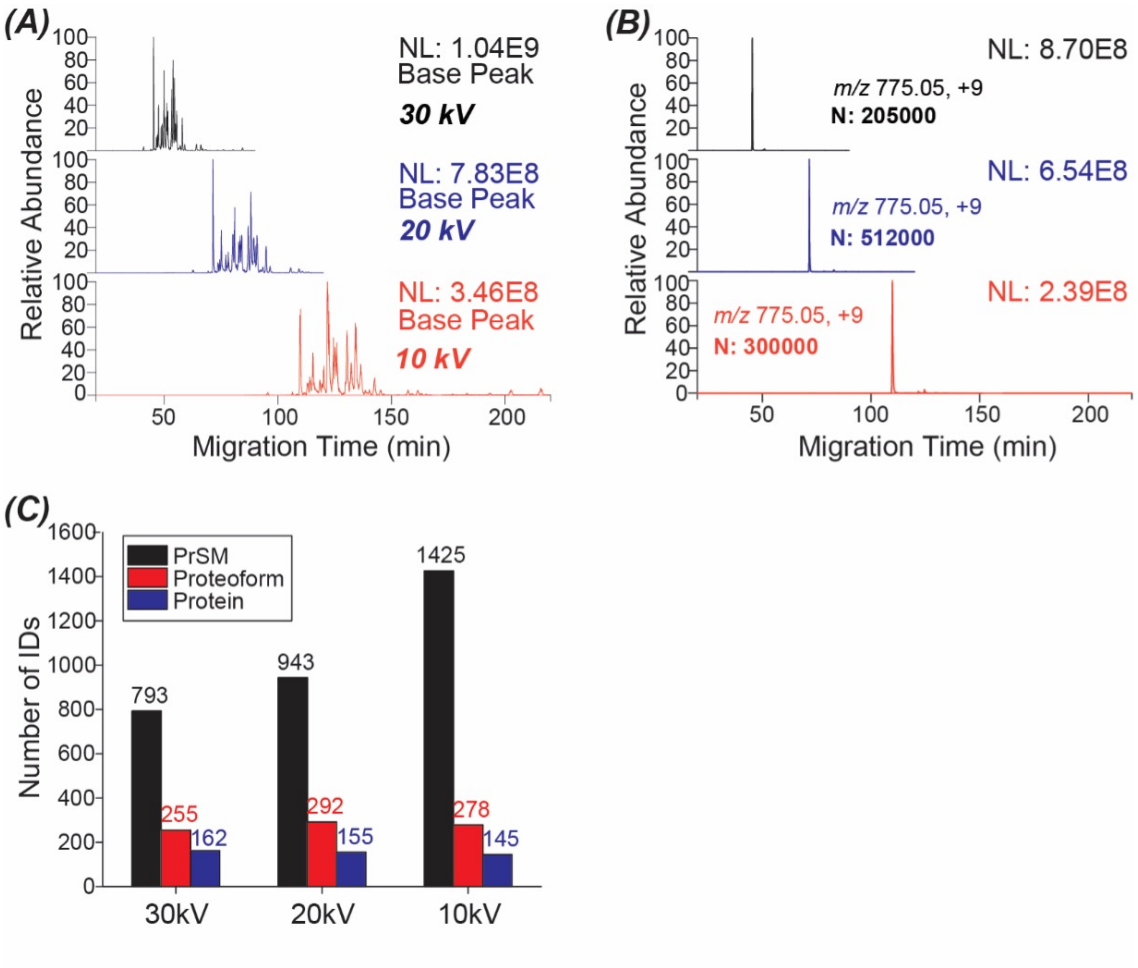


Figure 5.5: Summary of the data on optimizing the separation voltage of CZE. (A) Base peak electropherograms of the *E. coli* sample after CZE-MS analyses using 30 kV, 20 kV and 10 kV voltages. (B) EIEs of m/z 775.05 (charge +9) from one 30, 20, and 20 kV runs. (C) The PrSMs, proteoforms and proteins identified by CZE-AI-ETD with different separation voltages. This figure was reproduced with permission from reference [5].

191 proteins in a single CZE-AI-ETD run, and the number of proteoform IDs was 4% and 9% higher than that from the top 2 and top 5 methods. The top 4 method identified 2% and 7% more proteins than the top 2 and top 5 methods. The top 4 method was used in the following experiments.

We also evaluated the reproducibility of the optimized CZE-AI-ETD method for top-down characterization of the *E. coli* proteoform (Figure 5.6). The CZE-AI-ETD system produced reproducible separation profiles and base peak intensity across triplicate

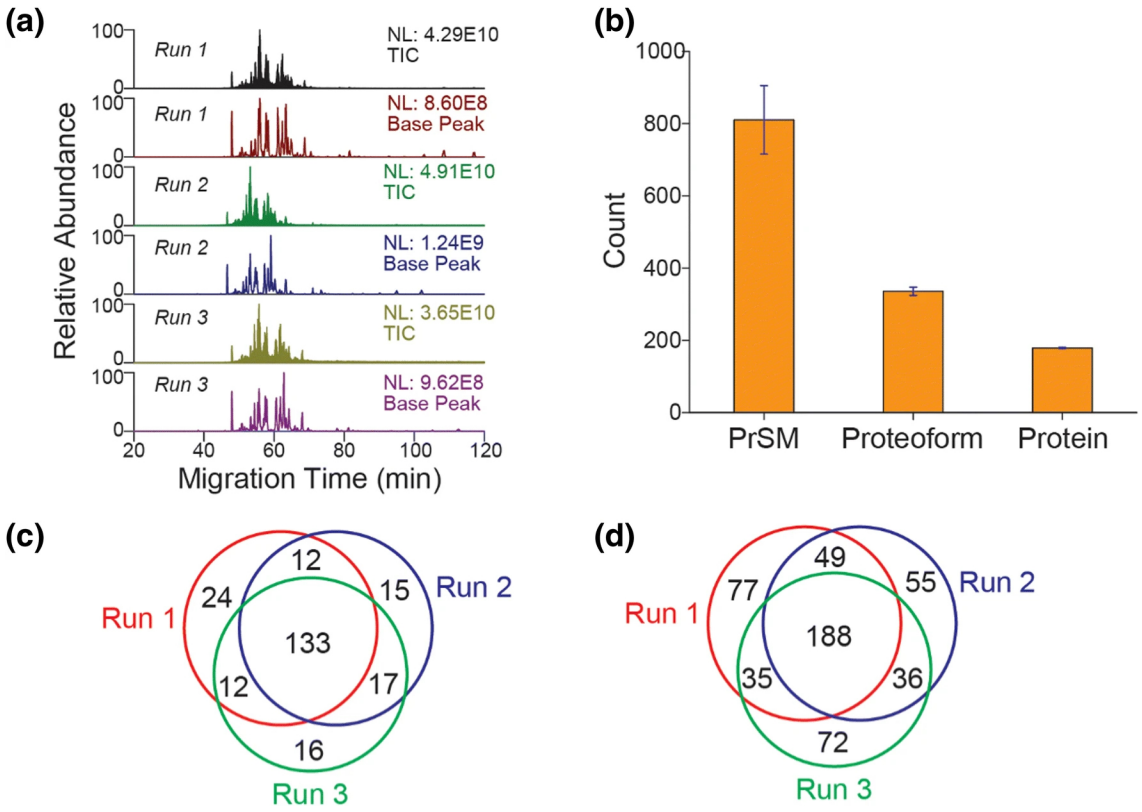


Figure 5.6: Reproducibility of the optimized CZE-AI-ETD system for TDP. (a) Base peak and TIC electropherograms of the *E. coli* sample analyzed by the optimized CZE-AI-ETD in triplicate. (b) Numbers of PrSMs, proteoforms, and proteins identified by the optimized CZE-AI-ETD. (c) The protein-level overlaps among the CZE-AI-ETD runs. (d) The proteoform-level overlaps among the CZE-AI-ETD runs. This figure was reproduced with permission from reference [5].

analyses (Figure 5.6a). The RSDs of PrSM IDs, proteoform IDs, and protein IDs were 12%, 3%, and 1%, respectively (Figure 5.6b). Error bars show the standard deviations of the IDs from the triplicate CZE-AI-ETD analyses. We further examined the protein-level and proteoform-level overlaps among the three CZE-AI-ETD runs (Figure 5.6c and Figure 5.6d). The overlaps were about 58% (protein-level) and 37% (proteoform-level) among the three runs.

5.3.3 SEC-CZE-AI-ETD for large-scale top-down characterization of the *E. coli* proteome

We fractionated the *E. coli* proteome into 14 fractions using SEC based on the size of proteoforms. Each SEC fraction was analyzed by the optimized CZE-AI-ETD in 120 min. Analyses of these 14 SEC fractions took 28 h. As shown in Figure 5.7a, the SEC fraction 12 was analyzed by the CZE-AI-ETD system, and a 50-min separation window was obtained in the run. The base peak electropherograms of the 14 SEC fractions are shown in Figure 5.8. The corresponding raw files have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD012247 [50].

Nearly 12,000 PrSMs, 3,028 proteoforms, and 387 proteins were identified from the *E. coli* proteome using the SEC-CZE-AI-ETD system with 1% spectrum-level and 5% proteoform-level FDRs. The list of identified proteoforms is shown elsewhere [5]. The data represents the largest TDP dataset using the AI-ETD method so far. The PrSM, proteoform, and protein IDs were not uniformly distributed across the 14 SEC fractions (Figure 5.7b). The number of proteoform IDs per SEC fraction ranged from as few as 25 proteoforms (fraction 2) to 957 proteoforms (fraction 14). On average, 216 proteoforms were identified per SEC fraction. Later SEC fractions tended to produce more PrSM, proteoform, and protein IDs. We need to note that single-shot CZE-AI-ETD of the SEC fraction 14 identified 957 proteoforms and 253 proteins in 120 min, and the number of proteoform and protein IDs from the fraction accounted for about 32% and 65% of the total proteoform and protein IDs.

The 3,028 proteoforms corresponded to 387 *E. coli* genes, an average of about 8 proteoforms per gene. The genes were classified into three categories based on the number of identified proteoforms: 1 proteoform per gene for 191 genes, 2-10 proteoforms per gene for 127 genes, and 10-144 proteoforms per gene for 69 genes (Figure 5.7c). We identified 144, 130, and 111 proteoforms for genes rbsB, rplL, and mglB, respectively. The mass of

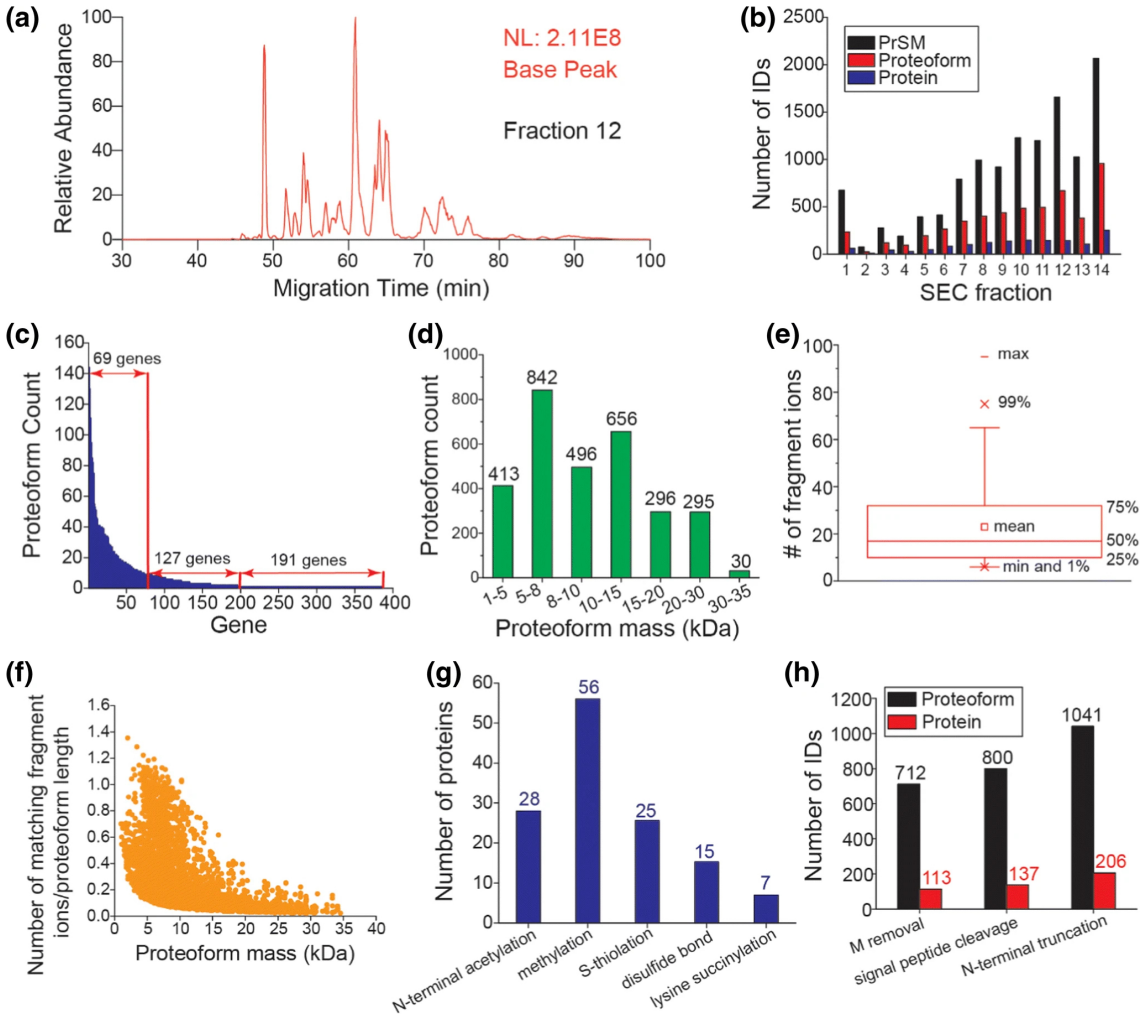


Figure 5.7: SEC-CZE-AI-ETD for large-scale TDP of the *E. coli* cells. (a) Base peak electropherogram of SEC fraction 12 of the *E. coli* proteome. (b) Distributions of the PrSM, proteoform, and protein IDs across the 14 SEC fractions. (c) Distribution of proteoform IDs per gene. (d) Distribution of the mass of identified proteoforms. (e) Box chart of the number of matching fragment ions of identified proteoforms. (f) Correlation between the proteoform mass and the normalized number of matching fragment ions. (g) Summary of the detected PTMs. (h) Summary of the detected N-terminal methionine (M) removal, potential signal peptide cleavage, and N-terminal truncations. This figure was reproduced with permission from reference [5].

identified proteoforms ranged from 1-35 kDa, and most of the proteoforms (89%) were smaller than 20 kDa (Figure 5.7d). 325 proteoforms from 51 proteins and 30 proteoforms from 6 proteins were larger than 20 kDa and 30 kDa, respectively. The proteoforms larger than 30 kDa were identified with at least 7 fragment ions, and the average number of

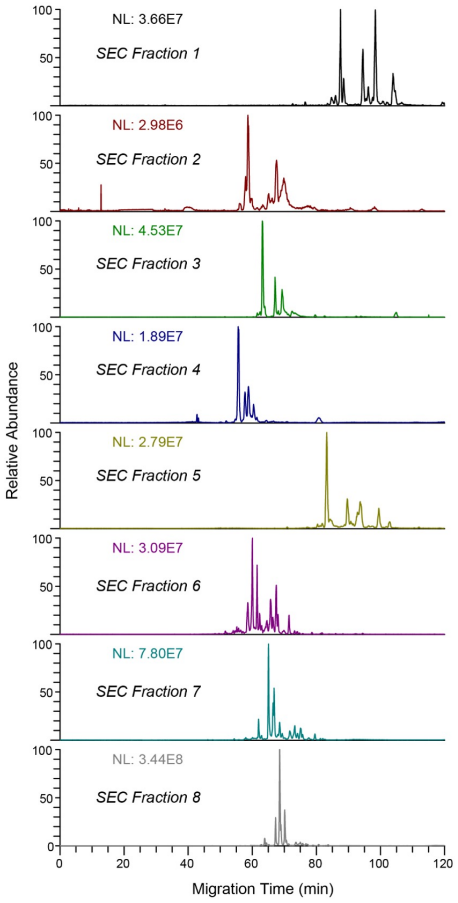


Figure 5.8: Base peak electropherograms of the SEC fractions 1-8 analyzed by the CZE-AI-ETD system. This figure was reproduced with permission from reference [5].

matching fragment ions was 18.

The number of matching fragment ions of identified proteoforms ranged from 6 to nearly 100 (Figure 5.7e). The mean was 23 and the median was 17. Roughly, 25% of the proteoforms were identified with fewer than 10 fragment ions. The proteoform mass influenced the number of matching fragment ions (Figure 5.7f). The number of matching fragment ions was normalized to the proteoform length that is the number of amino acid residues in a proteoform sequence. The number of fragment ions of each proteoform was normalized to the length of each corresponding proteoform, and the normalized number of fragment ions was used to evaluate the performance of AI-ETD for generation of sequence-informative fragment ions. When the proteoform mass increased, the performance

of AI-ETD tended to decrease (Figure 5.7f). However, the normalized number of fragment ions varied obviously for proteoforms with similar masses, suggesting that the performance of AI-ETD for proteoform fragmentation was also influenced by other proteoform features.

5.3.4 PTMs with SEC-CZE-AI-ETD

We detected several kinds of PTMs from the *E. coli* proteome, including protein N-terminal acetylation, methylation, S-thiolation, disulfide bonds, and lysine succinylation (Figure 5.6g). Only a few proteins in the *E. coli* sample had these PTMs. We detected 28 proteins with N-terminal acetylation, 56 proteins with methylation, 25 proteins with S-thiolation, 15 proteins with disulfide bonds (S-S), and 7 proteins with lysine succinylation. We identified 712 proteoforms from 113 proteins with N-terminal methionine removal, 800 proteoforms from 137 proteins with potential signal peptide cleavage, and 1041 proteoforms from 206 proteins with N-terminal truncations (Figure 5.7h).

The N-terminal acetylation was determined by the TopPIC software with a 42-Da mass shift at the N-terminus of one proteoform. The methylation was determined with a 14±1 Da or 28±1 Da mass shift. The S-thiolation was determined with a 305±2 Da mass shift for glutathionylation and a 119±2 Da mass shift for cysteinylation. We also manually checked that there was one cysteine residue in the sequence corresponding to the mass shift. For the S-S, if they are reported in the literature, we confirmed the detection through a -2±1 Da mass shift and two cysteine residues for one S-S and through a -4±1 Da mass shift and four cysteine residues for two S-S. If the S-S were not reported before, we required more accurate masses of the mass shifts (-2 Da for one S-S and -4 Da for two S-S). The lysine succinylation was determined with a 100±2 Da mass shift for one succinylation site, a 200±2 Da mass shift for two succinylation sites, and a 300±2 Da mass shift for three succinylation sites. If the first 7-50 amino acids of a proteoform were cleaved from its N-terminus, we considered the proteoform had a potential signal peptide cleavage based on information from the "Center for Biological Sequence Analysis"

(http://www.cbs.dtu.dk/services/SignalP-1.1/sp_lengths.html). If more than 50 amino acids were cleaved from the N-terminus of one proteoform, we reviewed the proteoform as truncated.

S-thiolation is a kind of PTM in which free thiol groups on proteins react with low mass thiols (e.g., glutathione and cysteine) to form disulfides. S-glutathionylation and S-cysteinylation are two kinds of S-thiolation. Protein S-thiolation can occur in response to oxidative stress and protect cysteine from irreversible oxidation, and it can happen under physical conditions to influence protein function [51–53]. Recently, Ansong et al. reported that Gram-negative bacteria cultured in LB medium preferred to use S-glutathionylation as a way for thiol protection [12]. We cultured the *E. coli* cells in LB medium for the experiment. We detected 25 proteins with S-glutathionylation PTM and only four proteins with S-cysteinylation PTM. Interestingly, the four cysteinylated proteins had both cysteinylated and glutathionylated proteoforms. Information of these proteins is listed elsewhere [5]. We compared the relative abundance of cysteinylated and glutathionylated proteoforms of two proteins (Figure 5.9a and 5.9b). In Figure 5.9a, 1022.66 m/z (charge +8) and 944.49 m/z (charge +8) were extracted with a 20-ppm mass tolerance for the S-glutathionylation and S-cysteinylation proteoforms. In Figure 5.9b, 920.02 m/z (charge +15) and 907.55 m/z (charge +15) were extracted with a 20-ppm mass tolerance for the S-glutathionylation and S-cysteinylation proteoforms.The glutathionylated proteoform showed much higher intensity than the cysteinylated proteoform, suggesting that the *E. coli* cells cultured in LB medium preferentially used S-glutathionylation as a mechanism for thiol protection. Figure 5.10 shows the sequences and fragmentation patterns of the cysteinylated and glutathionylated proteoforms of the two proteins. We identified 15 proteins with S-S, including 9 proteins with one, and 5 proteins with two. Interestingly, we also identified one protein, RNA polymerase-binding transcription factor DksA, which had one proteoform with one S-S and another proteoform with two S-S. Five out of the 15 proteins have been reported as S-S containing proteins in the literature. Information of the

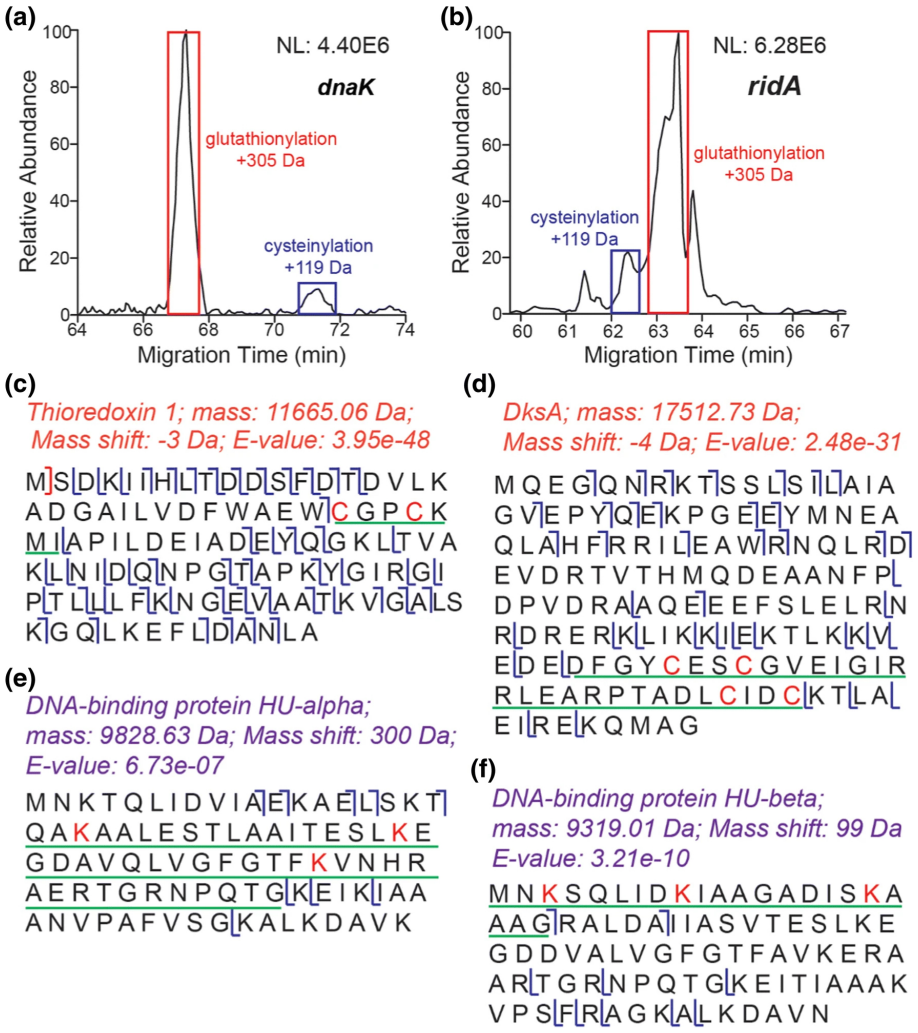


Figure 5.9: Examples of the S-thiolation, disulfide bond, and lysine succinylation PTMs. (a) EIE of the dnaK proteoforms with S-glutathionylation and S-cysteinylation. (b) EIE of the ridA proteoforms with S-glutathionylation and S-cysteinylation. (c) The sequence and fragmentation pattern of thioredoxin 1. (d) The sequence and fragmentation pattern of DksA. (e) The sequence and fragmentation pattern of DNA-binding protein HU-alpha. (f) The sequence and fragmentation pattern of DNA-binding protein HU-beta. This figure was reproduced with permission from reference [5].

15 proteins is listed elsewhere [5]. Figure 5.9c and 5.9d show two examples of these proteins. In Figure 5.9c, the sequence underlined with a green line had a -3 Da mass shift corresponding to a disulfide bond between the two cysteine residues. In Figure 5.9d, the sequence underlined with a green line had a -4 Da mass shift corresponding to two disulfide bonds between the four cysteine residues. These two proteins, thioredoxin 1 and DksA,

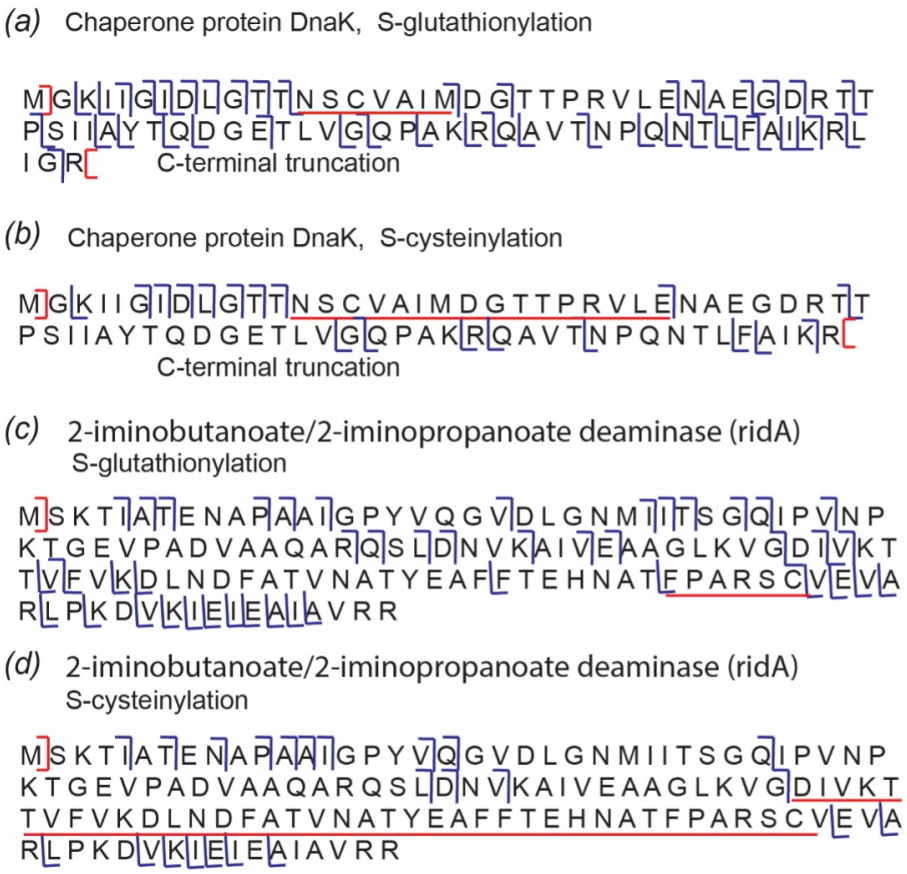


Figure 5.10: The sequences and fragmentation patterns of DnaK with S-glutathionylation (a), DnaK with S-cysteinylation (b), ridA with S-glutathionylation (c), and ridA with S-cysteinylation. (d)The S-thiolation modifications are in the regions highlighted with red underlines. This figure was reproduced with permission from reference [5].

were well fragmented with AI-ETD. The disulfide bonds were well localized based on the matching fragment ions. Thioredoxin 1 has one S-S between the two cysteine residues highlighted in red in Figure 5.9c [54]. Figure 5.9d and Figure 5.11 show the sequences and fragmentation patterns of DksA proteoforms with two and one S-S between the cysteine residues highlighted in red. The sequence underlined with a green line had a -2 Da mass shift corresponding to one disulfide bond between the two cysteine residues in Figure 5.11. DksA does not have S-S based on the literature, and instead, it binds one zinc ion through the four cysteine residues highlighted in Figure 5.9d [55]. The detected S-S on protein DksA might be endogenous or might form after cell lysis because the *E. coli* cells were

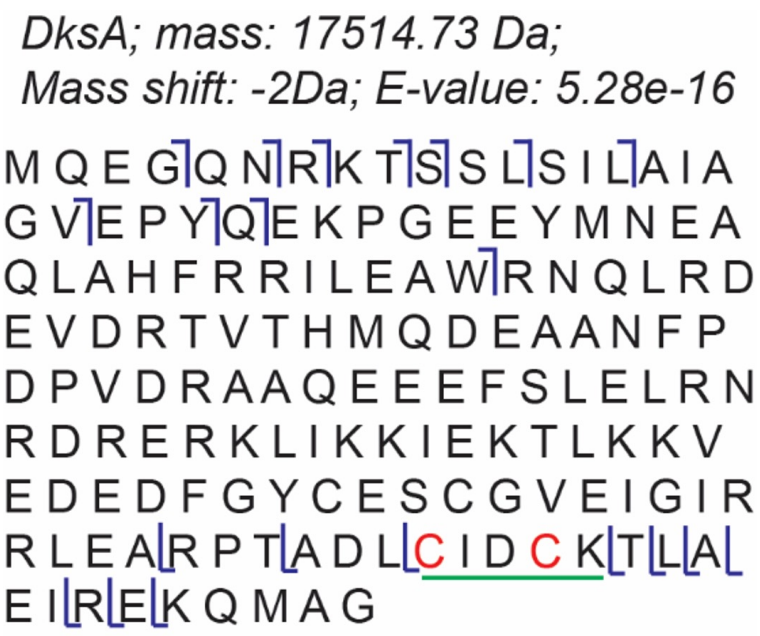


Figure 5.11: The sequence and fragmentation pattern of DksA. This figure was reproduced with permission from reference [5].

lysed under a denaturing condition and the DksA-zinc complex was most likely destroyed during the process.

We also identified seven proteins with the lysine succinylation PTM, and these seven proteins were reported as succinylated proteins in the literature [56]. Three proteins had one modification site (100-Da mass shift), two proteins had three modification sites (300-Da mass shift), and one protein had two modification sites (200-Da mass shift). Interestingly, the lysine residues on ribose import binding protein RbsB were not succinylated consistently across different proteoforms. Two RbsB proteoforms had two succinylation sites, but the sites were different between the proteoforms. We also identified one RbsB proteoform with only one succinylation site. The information on proteins with lysine succinylation is shown elsewhere [5]. As shown in Figure 5.9e, the three modification sites on DNA-binding protein HU-alpha were localized based on the fragment ions generated by AI-ETD. The sequence underlined with a green line had a 300-Da mass shift corresponding to succinylations on the three lysine residues. Figure 5.9f shows the sequence

and fragmentation pattern of another succinylated protein, DNA-binding protein HU-beta, indicating one succinylation site on one of the three lysine residues highlighted in red. The sequence underlined with a green line had a 99-Da mass shift corresponding to succinylation on one of the three lysine residues.

5.4 CZE-UVPD

As shown in Figure 5.12, proteins were extracted from zebrafish brains and fractionated by SEC into ten fractions based on their size. The SEC fractions were analyzed by CZE-MS/MS using UVPD (213 nm) for proteoform fragmentation. The ProSight PD

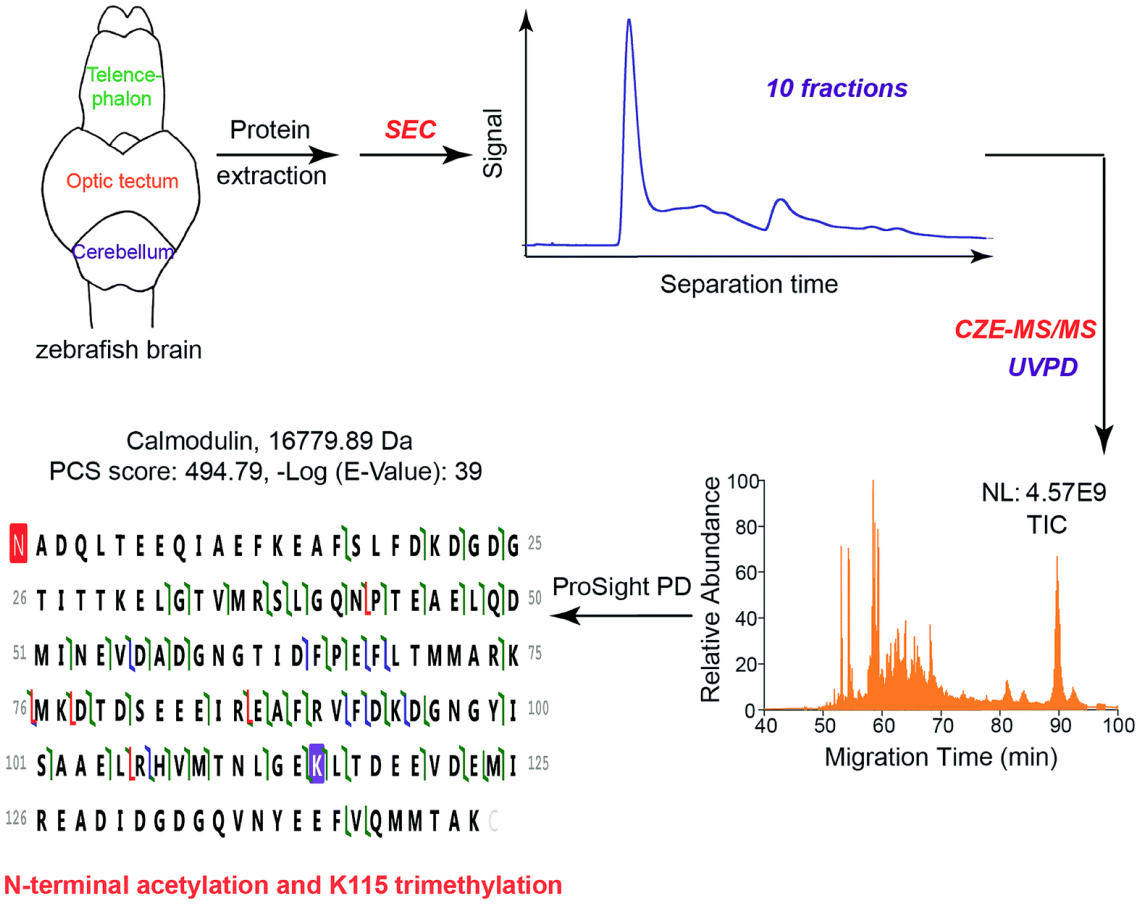


Figure 5.12: Diagram of the experimental design. NL: normalized level. TIC: total ion current. This figure was reproduced with permission from reference [6].

software was used for database search for proteoform IDs. A 1% PrSM-level FDR and a 5% proteoform-level FDR were employed to filter the database search results. The workflow was able to identify proteoforms with high confidence. For example, one proteoform of calmodulin was identified with a good E-Value and a high proteoform characterization score (PCS), Figure 5.12. N-terminal acetylation and K115 trimethylation were also detected on this calmodulin proteoform.

In total, about 600 proteoforms and 369 proteins were identified from the zebrafish brain samples using the SEC-CZE-MS/MS in roughly 20 hours. The identified proteoforms from each SEC fraction are listed elsewhere [6]. The dataset represents one of the largest TDP datasets using UVPD. In the meantime, this work represents the first application of CZE-UVPD for TDP. The number of proteoform and protein IDs are not uniformly distributed across the ten SEC fractions, Figure 5.13A. Single-shot CZE-UVPD analysis of the SEC fraction 8 identified 227 proteoforms from 139 proteins. The number of proteoform and protein IDs in each SEC fraction are in ranges of 7-227 and 5-139. On average, $\sim$ 95 proteoforms and $\sim$ 62 proteins were identified per SEC fractions.

SEC separates proteoforms based on their size. We plotted the mass distribution of the identified proteoforms in each SEC fraction, Figure 5.13B. On average, the identified proteoforms in early SEC fractions have larger masses than that in late SEC fractions, which agrees well with the separation principle of SEC. We noted that the adjacent SEC fractions have obvious overlaps regarding proteoform mass, suggesting the relatively low resolution of the SEC column used in the experiment for proteoform separation based on their masses. In our future study, we will increase the length of the SEC column for better separation resolution or try the serial SEC method developed by the Ge group recently [11]. Our SEC-CZE-UVPD system identified proteoforms in a mass range of $\sim$ 3-21 kDa. TDP usually has difficulty in the identification of large proteins, partially due to the limited mass resolution of mass spectrometers and inefficient gas-phase fragmentation of large proteins.

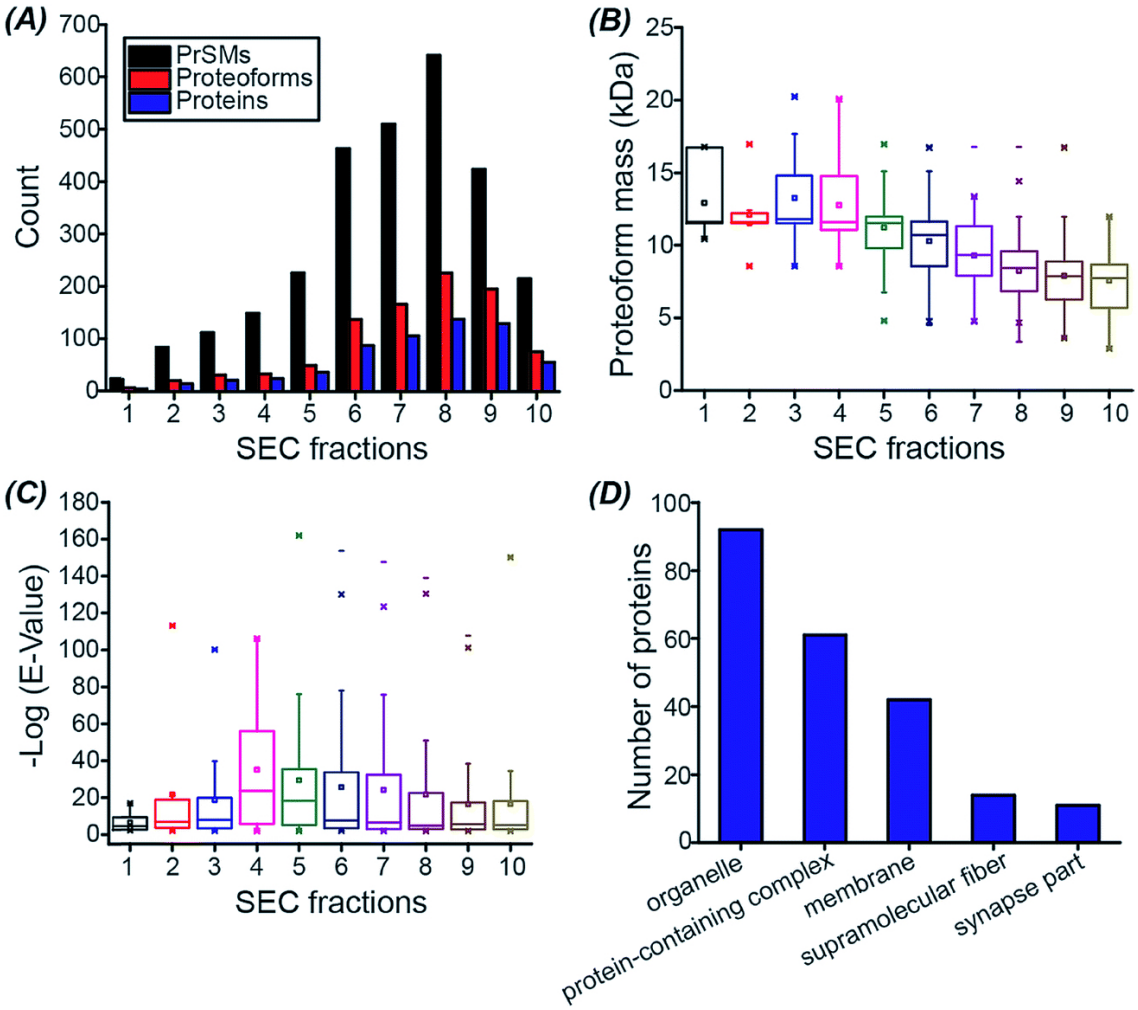


Figure 5.13: Summary of the SEC-CZE-UVPD data. (A) Numbers of PrSMs, proteoforms, and proteins vs. SEC fractions. (B) Box-plots of the masses of proteoforms identified from each SEC fraction. (C) Box-plots of the -log (E-Value) of identified proteoforms from each SEC fraction. (D) Cellular component distribution of identified proteins obtained from the UniProt website using the Retrieve/ID mapping tool. This figure was reproduced with permission from reference [6].

In our work, a target-decoy approach was used to evaluate the FDR of proteoform IDs and a 5% proteoform-level FDR was used to filter the proteoform IDs. For each identified proteoform listed elsewhere, a P-Score (Probability Score) and an E-Value (Expectation Value) were reported for the ID [6]. Lower P-Scores and E-Values indicates better confidence in proteoform IDs; Higher -log (P-Score) and -log (E-Value) are better regarding the confidence of proteoform ID. For example, as shown in Figure 5.12, one proteoform of

calmodulin was identified with -log (E-Value) as indicating a proteoform ID with very high confidence. The proteoform was well fragmented across the proteoform sequence, producing a good number of fragment ions. E-Value is a non-linear transformation of the number of matched fragment ions in a MS/MS spectrum. The -log (E-Value) ranges from 2 to over 160 for the identified proteoforms and the median value of -log (E-Value) is in a range of 4.6-24 for the ten SEC fractions, Figure 5.13C. Some of the proteoform IDs have -log (E-Values) well below 10, indicating small numbers of fragment ions produced during UVPD. Better UVPD fragmentation will be helpful for more confident identification and characterization of these proteoforms. We also analyzed the cellular component (CC) information of the identified proteins, Figure 5.13D. The Top 5 CCs are organelle, protein-containing complex, membrane, supramolecular fiber, and synapse part. Proteoforms of over 40 membrane proteins were identified in this work.

UVPD (213 nm) has produced reasonably good gas-phase fragmentation for some proteoforms. As shown in Figures 5.14A and 5.12B, 75% and 73% backbone cleavages for Parvalbumin-7 (11932 Da) and Si:dkey-46i9.1 (7184 Da) were observed using UVPD. a/x ions, b/y ions, and c/z ions were marked in green, blue, and red. For a relatively large protein, ATP synthase subunit d (18158 Da), only 25% backbone cleavage was obtained, Figure 5.14C. The data suggest that the extensive fragmentation of large proteins is still challenging with the UVPD (213 nm). Interestingly, 87% backbone cleavage was reported by the Brodbelt's group for carbonic anhydrase II (29 kDa) using UVPD (193 nm) under an optimal condition [38]. The data suggest that the fragmentation performance of UVPD for large proteins can certainly be improved with further systematic optimization. The dominant fragment ion types from UVPD (213 nm) for Parvalbumin-7 and Si:dkey-46i9.1 are a, x, y and z ions, Figure 5.14D. For the ATP synthase subunit d, a and y ions are the dominant ones. The data reasonably agrees with that reported in the literature [38].

Our SEC-CZE-UVPD system detected various PTMs from the zebrafish brain sample,

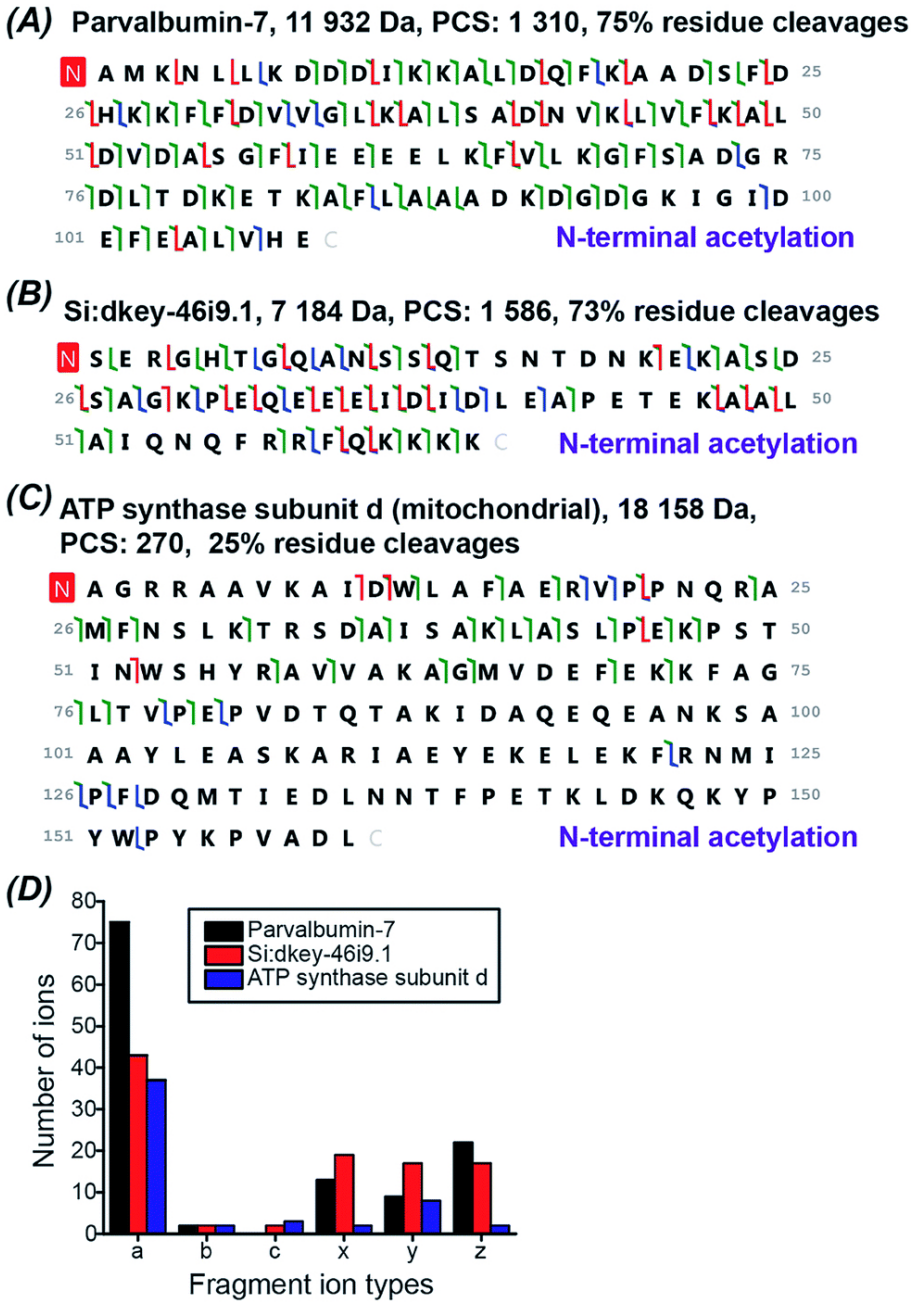


Figure 5.14: Proteoform fragmentation data. (A)-(C): sequences and fragmentation patterns of Parvalbumin-7, Si:dkey-46i9.1, and ATP synthase subunit d (mitochondrial). (D) Distribution of the fragment ion types for the three proteoforms shown in (A)-(C). This figure was reproduced with permission from reference [6].

including N-terminal acetylation, trimethylation and myristoylation of N-terminal glycine.

Acetylation was the most abundant PTM (most commonly N-terminal acetylation) with a

total of 156 acetylated proteoforms. An example of an acetylated protein well-characterized by CZE-UVPD in the zebrafish brain sample is calmodulin from the calm1a gene. The calm1a gene has one ortholog in human. The calmodulin-calcium complex is known to control kinases, phosphatases, and other proteins. In this work, we detected two different proteoforms from the calm1a gene. One proteoform (Proteoform 1, theoretical mass: 16739 Da) has only N-terminal acetylation and the other proteoform (Proteoform 2, theoretical mass: 16781 Da) has both N-terminal acetylation and K115 trimethylation, Figures 5.15A and 5.15B. a/x ions, b/y ions, and c/z ions were marked in green, blue, and red. Both

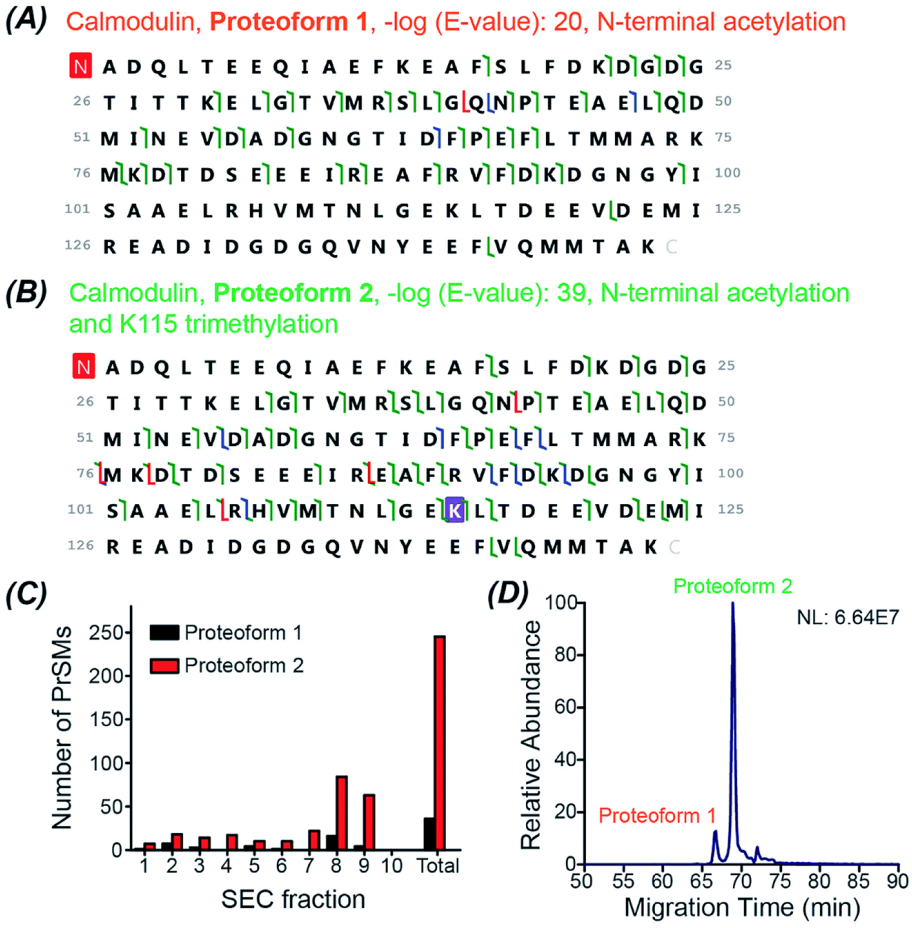


Figure 5.15: Data about proteoforms of the calmodulin. (A)-(B): sequences and fragmentation patterns of the Proteoform 1 and Proteoform 2 of calmodulin. (C): Distribution of the PrSMs of the Proteoform 1 and Proteoform 2 across different SEC fractions. (D): EIE of the Proteoform 1 and Proteoform 2 from the data of the SEC fraction 8. This figure was reproduced with permission from reference [6].

proteoforms were identified with good confidence with E-Values better than 10^{-20} (Proteoform 1) and 10^{-39} (Proteoform 2). The identification confidence of Proteoform 2 is much higher than that of Proteoform 1, indicated by the much lower E-Value. It has been reported that calmodulin in the human brain has both N-terminal acetylation and K115 trimethylation [57]. However, there is no experimental evidence in the literature on the N-terminal acetylation and K115 trimethylation of calmodulin in zebrafish brain based on the information in the UniProt Protein knowledgebase (https://www.uniprot.org/uniprot/Q6PI52). Here we identified the two proteoforms of calmodulin from the zebrafish brain for the first time. We also compared the relative abundance of the two proteoforms of calmodulin in each SEC fraction based on their PrSMs [15]. Proteoform 2 has much higher abundance than Proteoform 1 in all the SEC fractions, Figure 5.15C. The PrSM data agrees well with the proteoform intensity data, Figure 5.15D. For peak extraction, m/z 1132.54 (+15) was used for the Proteoform 1 and m/z 1120.33 (+15) was used for the Proteoform 2. The mass tolerance was 20 ppm for peak extraction and Gaussian smoothing (5 points) was applied. We noted that the Proteoform 2 migrated slower than the Proteoform 1 during CZE separation, most likely because K115 trimethylation reduced the overall charge of the protein, Figure 5.15D.

5.5 Conclusion

We demonstrated the first application of CZE-AI-ETD for large-scale TDP. CZE-AI-ETD outperformed CZE-ETD and CZE-HCD considering the number of proteoform and protein IDs as well as the number of sequence-informative fragment ions generated. Coupling SEC fractionation to CZE-AI-ETD enabled IDs of 3028 proteoforms and 387 proteins from the *E. coli* proteome, which represents the largest TDP dataset using the AI-ETD method so far. The SEC-CZE-AI-ETD system detected various PTMs, including protein N-terminal acetylation, methylation, S-thiolation, disulfide bonds, and lysine succinylation.

CZE-UVPD was applied for top-down proteomics for the first time. About 600 proteoforms and 369 proteins were identified from a zebrafish brain sample using the SEC-CZE-UVPD. The pilot data demonstrate the great potential of CZE-UVPD for large-scale TDP. We expect that further systematic optimization of the UVPD fragmentation and further improvements in the SEC-CZE separations will boost the number of proteoform and protein IDs drastically.

We noted that UVPD (213 nm) did not reach extensive fragmentation for many identified proteoforms, which made the complete characterization of these proteoforms challenging. We expect that combinations of various fragmentation methods, e.g., HCD, ETD/AI-ETD, and UVPD, will be useful for improving the quality of proteoform characterization.

5.6 Acknowledgments

We thank Prof. Heedeok Hong's group at Michigan State University (Department of Chemistry) for kindly providing the *E. coli* cells for this project. We thank Prof. Jose Cibelli's group at Michigan State University (Department of Animal Science) for kindly providing the zebrafish brains for this project. We thank the support from the National Institute of General Medical Sciences, National Institutes of Health (NIH), through Grant Nos. R01GM118470 (X. Liu), R01GM125991 (L. Sun and X. Liu), P41GM108538 (J. Coon), R35GM118110 (J. Lodge and J. Coon).

BIBLIOGRAPHY

BIBLIOGRAPHY

(1) Toby, T. K.; Fornelli, L.; Kelleher, N. L. *Annu. Rev. Anal. Chem.* **2016**, *9*, 499–519.

(2) Smith, L. M.; Kelleher, N. L. *Science* **2018**, *359*, 1106–1107.

(3) Evans, V. C.; Barker, G.; Heesom, K. J.; Fan, J.; Bessant, C.; Matthews, D. A. *Nat. Methods* **2012**, *9*, 1207–1211.

(4) Wang, X.; Slebos, R. J. C.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B. *J. Proteome Res.* **2012**, *11*, 1009–1017.

(5) McCool, E. N.; Lodge, J. M.; Basharat, A. R.; Liu, X.; Coon, J. J.; Sun, L. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 2470–2479.

(6) McCool, E. N.; Chen, D.; Li, W.; Liu, Y.; Sun, L. *Anal. Methods* **2019**, *11*, 2855–2861.

(7) Cravatt, B. F.; Simon, G. M.; III, J. R. Y. *Nature* **2007**, *9*, 991–1000.

(8) Lucitt, M. B.; Price, T. S.; Pizzaro, A.; Wu, W.; Yocum, A. K.; Seiler, C.; Pack, M. A.; Fitzgerald, G. A.; Grosser, T. **2008**, *7*, 981–994.

(9) Tran, J. C. et al. *Nature* **2011**, *480*, 254–258.

(10) Durbin, K. R.; Fornelli, L.; Fellers, R. T.; Doubleday, P. F.; Narita, M.; Kelleher, N. L. *J. Proteome Res.* **2016**, *15*, 976–982.

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

(12) Ansong, C. et al. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 10153–10158.

(13) Liang, Y.; Jin, Y.; Wu, Z.; Tucholski, T.; brown, K. A.; Zhang, L.; Zhang, Y.; Ge, Y. *Anal. Chem.* **2019**, *91*, 1743–1747.

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

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

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

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

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

(19) Roth, M. J.; Plymire, D. A.; Chang, A. N.; Kim, J.; Maresh, E. M.; Larson, S. E.; Patrie, S. M. *Anal. Chem.* **2011**, *83*, 9586–9592.

(20) Zhou, Y.; Zhang, X.; Fornelli, L.; Compton, P. D.; Kelleher, N.; Wirth, M. J. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **2017**, *1044-1045*, 47–53.

(21) Moini, M. *Anal. Chem.* **2007**, *79*, 4241–4246.

(22) Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2554–2560.

(23) Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J. *J. Proteome Res.* **2015**, *14*, 2312–2321.

(24) Valaskovic, G. A.; Kelleher, N. L.; McLafferty, F. W. *Science* **1996**, *273*, 1199–1202.

(25) Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; III, J. R. Y. *J. Proteome Res.* **2014**, *13*, 6078–6086.

(26) Zhao, Y.; Sun, L.; Zhu, G.; Dovichi, N. J. *J. Proteome Res.* **2016**, *15*, 3679–3685.

(27) Li, Y.; Tran, J. C.; Ntai, I.; Kelleher, N. L. *Proteomics* **2014**, *14*, 1158–1164.

(28) McCool, E. N.; Lubeckyj, R.; Shen, X.; Kou, Q.; Liu, X.; Sun, L. *J. Vis. Exp.* **2018**, *120*, 115644.

(29) Zhu, G.; Sun, L.; Dovichi, N. J. *Talanta* **2016**, *146*, 839–843.

(30) Aebersold, R.; Morrison, H. D. *J. Chromatogr.* **1990**, *516*, 79–88.

(31) Britz-Mckibbin, P.; Chen, D. D. Y. *Anal. Chem.* **2000**, *72*, 1242–1252.

(32) Han, X.; Wang, Y.; Aslanian, A.; Bern, M.; Lavellée-Adam, M.; III, J. R. Y. *Anal. Chem.* **2014**, *86*, 11006–11012.

(33) Sun, L.; Knierman, M. D.; Zhu, G.; Dovichi, N. J. *Anal. Chem.* **2013**, *85*, 5989–5995.

(34) Haverland, N. A.; Skinner, O. S.; Fellers, R. T.; Tariq, A. A.; Early, B. P.; Fornelli, R. D.; Compton, P. D.; Kelleher, N. L. *J. Am. Mass Spectrom.* **2017**, *28*, 1203–1215.

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

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

(37) Wysocki, V. H.; Tsapralis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* **2000**, *35*, 1399–1406.

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

(39) Rush, M. J. P.; Riley, N. M.; Westphall, M. S.; Coon, J. J. *Anal. Chem.* **2018**, *90*, 8946–8953.

(40) Riley, N. M.; Westphall, M. S.; Coon, J. J. *J. Proteome Res.* **2017**, *16*, 2653–2659.

(41) Riley, N. M.; Westphall, M. S.; Coon, J. J. *J. Am. Soc. Mass Spectrom.* **2018**, *29*, 140–149.

(42) Zhao, Y. et al. *Anal. Chem.* **2015**, *87*, 5422–5429.

(43) Gargano, A. F. G.; Shaw, J. B.; Zhou, M.; Wilkins, C. S.; Fillmore, T. L.; Moore, R. J.; Somsen, G. W.; Paša-Tolić, L. *J. Proteome Res.* **2018**, *17*, 3791–3800.

(44) Greer, S. M.; Bern, M.; Becker, C.; Brodbelt, J. S. *J. Proteome Res.* **2018**, *17*, 1340–1347.

(45) Zandborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. *Nucleic Acids Res.* **2007**, *35*, W701–W706.

(46) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383–5392.

(47) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207–214.

(48) Kou, Q.; Xun, L.; Liu, X. *Bioinformatics* **2016**, *32*, 3495–3497.

(49) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534–2536.

(50) Vizcaíno, J. A. et al. *Nucleic Acids Res.* **2016**, *44*, D447–D456.

(51) Dalle-Donne, I.; Rossi, R.; Colombo, G.; Giustarini, D.; Milzani, A. *Trends Biochem. Sci.* **2009**, *34*, 85–96.

(52) Hochgräfe, F.; Mostertz, J.; Pöther, D. C.; Becher, D.; Helmann, J. D.; Hecker, M. *J. Biol. Chem.* **2007**, *282*, 25981–25985.

(53) Chu, F.; Ward, N. E.; O'Brian, C. A. *Carcinogenesis* **2003**, *24*, 317–325.

(54) Schultz, L. W.; Chivers, P. T.; Raines, R. T. *Acta Crystallogr. D Biol. Crystallogr.* **1999**, *55*, 1533–1538.

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

(56) Weinert, B. T.; Schölz, C.; Wagner, S. A.; Iesmantavicius, V.; Su, D.; Daniel, J. A.; Choudhary, C. *Cell Rep.* **2013**, *4*, 842–851.

(57) Sasagawa, T.; Ericsson, L. H.; Walsh, K. A.; Schreiber, W. E.; Fischer, E. H.; Titani, K. *Biochemistry* **1982**, *21*, 2565–2569.

Chapter 6

Conclusion

6.1 Future directions

TDP of biological systems would not be possible without teams of researchers, across the country and abroad, continuously developing aspects of TDP, Figure 6.1. The TDP community has expanded rapidly in the past couple of decades, and has made a concerted effort to propel TDP into of a variety of fields and applications [1–5]. The natural limitations associated with efficiency of front-end intact protein separations, speed of mass analysis, and efficiency of fragmentation have been and will continue to hamper the advancement of TDP into more useful applications outside of the laboratory. In conjunction with other -omics approaches, TDP could become a routinely used method. However, without advancements like what was seen in the genome project, it is unlikely to carve out a standalone niche in a clinical or industrial setting. This is also a good thing, as outside influences, such as foreign proteins, peptides, transcripts, small molecules, and environmental factors also impact what will eventually be seen at the proteoform-level of a species. It is more important to develop sensitive and selective diagnostics that include information from a wide variety of angles to gain insight into patient phenotype [6].

As deep TDP is becoming more efficient and thorough, TDP is beginning to be recognized as a useful approach to personalized medicine,including therapeutics, diagnostics, and drug development, embryonic development, tissue imaging, and pathogen identification and characterization, to name a few [1, 6, 7]. In terms of cancer, it is

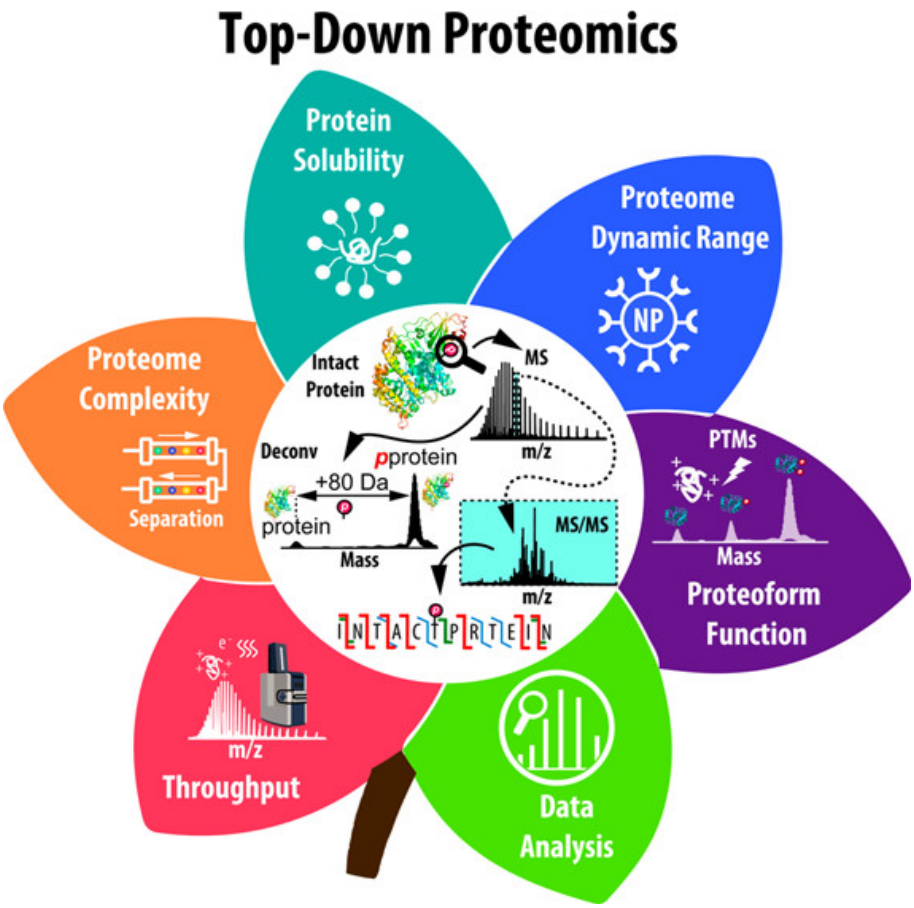


Figure 6.1: An overview of the challenges in TDP. This figure is reproduced with permission from reference [1].

universally recognized that no two patients' cancers are exactly the same, and that there is molecular heterogeneity between different tumor subtypes within a tumor type [8, 9]. Molecular heterogeneity has even been realized within the same tumors using single-cell proteomics [10]. Heterogeneity and the importance of personalized medical approaches continues to become more evident with the use of genomics and proteomics (proteogenomics) to analyze various cell types and tumors, Figure 6.2 [8, 11, 12]. A more precise medical approach tailored to an individual patient's cancer based on their phenotype is needed and may be accomplished using technologies that can be extended to a wide variety of tissue and serum samples [1, 13–18].

Proteomics and TDP will also be invaluable in the identification of biomarkers, and

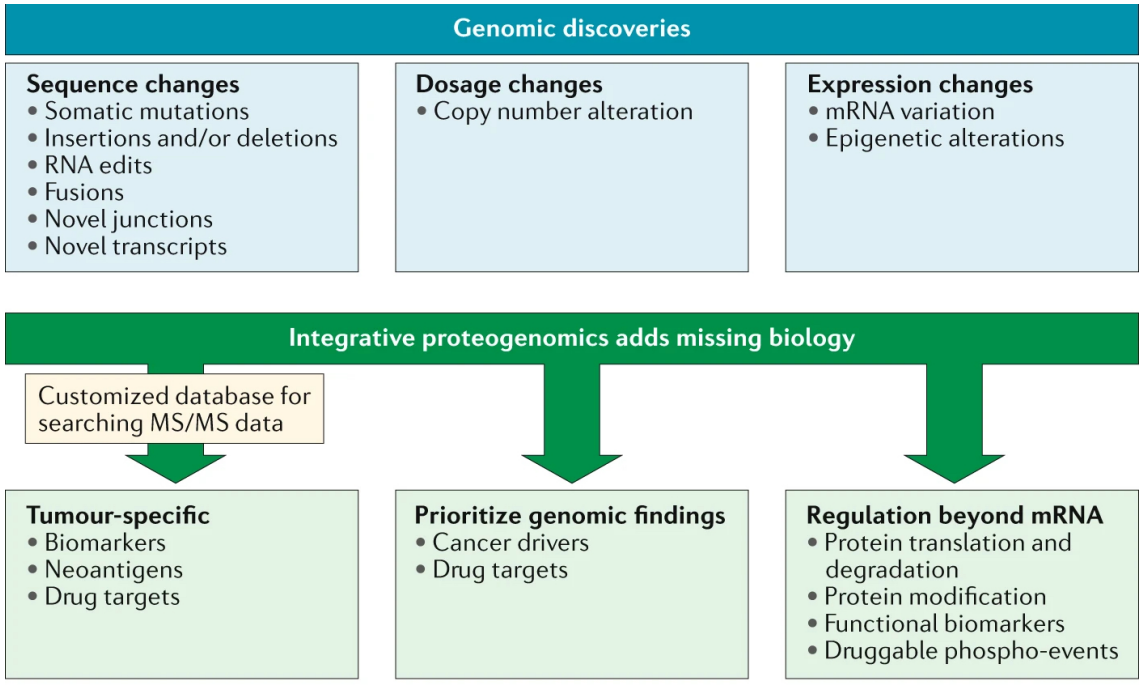


Figure 6.2: An overview of what proteomics and genomics, used in conjunction, can offer in terms of biological discovery and it's clinical potential. This figure was reproduced with permission from reference [11].

identification and characterization of drugs and therapeutics. Newer biotherapeutics are usually monoclonal antibodies, and characterizing and following up with them after they have been on the market involves a variety of analytical challenges that many people are trying to solve, including using CE-MS in some targeted studies [19]. A lot of recent work has identified, using TDP and BUP, potential biomarkers for phenotypic characterization using reference cell lines and tissue samples [14, 20].

Understanding embryonic development is another area that TDP could provide invaluable insight into. *Danio rerio* (zebrafish) has been viewed as an important model organism for vertebrate development studies. 71% of human genes have at least one ortholog in zebrafish and 82% of the known genes responsible for human disease are present in zebrafish [21]. Research on understanding zebrafish early embryogenesis could shed invaluable light on human early embryogenesis. Figure 6.3 shows a few of the important

stages of embryonic development in zebrafish. Although many transcriptome studies have

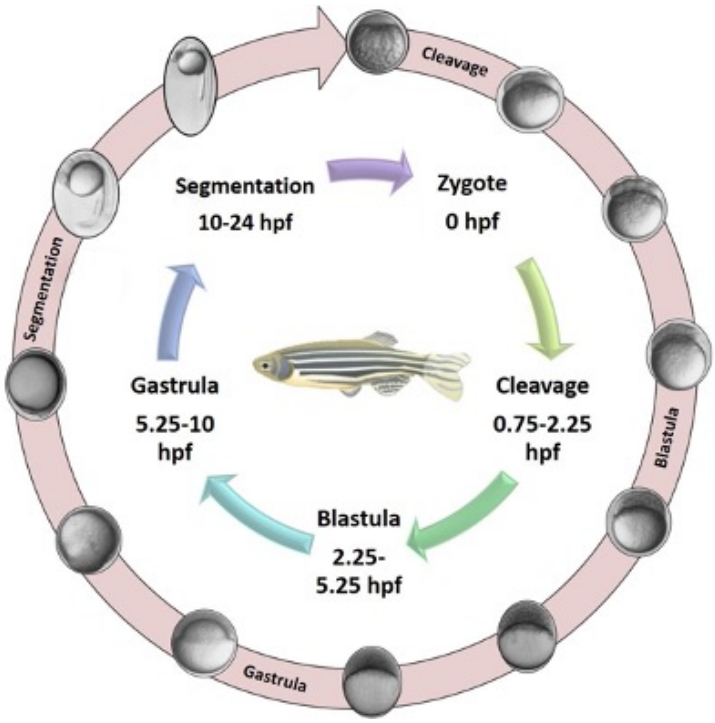


Figure 6.3: A figure demonstrated the stages of embryonic development in zebrafish. Parts of this figure are adapted with permission from reference [22].

been done on zebrafish early embryogenesis, transcriptome-level information cannot fully reflect proteome-level information during early embryogenesis due to several reasons [23–26]. First, zygotic transcription is silent before mid-blastula transition (MBT). Second, post-transcriptional regulation modulates gene expression. Third, protein PTMs affect protein function. Time-resolved, quantitative proteomics datasets for zebrafish early-stage embryos will provide new insights into early embryogenesis. Unfortunately, those datasets are not available, and it is a goal of our research group to create them.

Another relatively active field for TDP is pathogen identification and characterization Figure 6.4 [27–30]. In addition to analysis of the pathogens, understanding their affect on the body is important for development of potential therapies. A recent work looked at the multi-organ proteomic landscape of COVID-19 autopsies [31]. I anticipate there to be interest in using TDP to better characterize individual patient responses to pathogens,

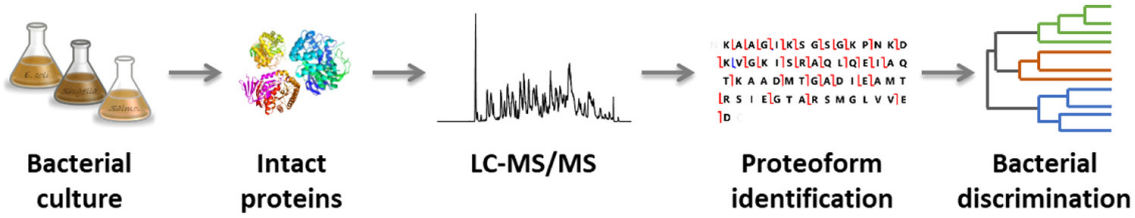


Figure 6.4: Workflow for bacterial discrimination. This figure is reproduced with permission from reference [30].

much like it is starting to be used for potential personalized medical approaches for diseases. In a very creative use of CZE, the Dovichi group collected fractions from an environmental microbiome using CZE to separate intact bacterial cells, followed by culturing and identification of bacteria using Sanger sequencing. Similar studies could be repeated with TDP used as the bacterial identification method, possibly providing even more accurate differentiation between species [32].

Finally, in addition to the parts of the TDP workflow mentioned earlier in this dissertation, I expect there to be more of a push towards identifying larger proteoforms in TDP studies. This not only includes specialized mass analyzers with high fields and extended mass ranges, but specially designed separations and fragmentation methods [33–35]. TDP can also work in conjunction with other proteomics methods for better proteome characterization. I expect this to happen more frequently with BUP and native proteomics for better databases for searching with TDP datasets and for analysis of protein complexes [36, 37]. Aebersold, et al. developed a method for native separation of protein complexes followed by BUP for protein complex analysis [37]. Similar work could be done with TDP for even better understanding of these complexes and to get the best of both worlds for each of these techniques for better proteome coverage and characterization. Overall, I expect there to be continued and more heavy investment in CZE-MS-MS and TDP in general, targeted and untargeted, by industry and academia, and more expeditious and creative improvements to the parts of the TDP workflow mentioned in this dissertation.

6.2 Summary

Efficient separations coupled to extensive proteoform fragmentation results in well characterized proteoforms. Our group has been able to substantially increase the number of proteoform identifications, we have been able to identify and localize interesting modifications, and we have been able to perform quantitative TDP for identification of differentially expressed proteoforms in primary and metastatic tumor samples using LC-CZE-MS/MS [38–44]. These are all extremely important leaps forward for the TDP workflow and sets the table for future TDP studies where I believe that TDP will be a significant part of the future of personalized medicine.

BIBLIOGRAPHY

BIBLIOGRAPHY

(1) Melby, J. A.; Roberts, D. S.; Larson, E. J.; Brown, K. A.; Bayne, E. F.; Jin, S.; Ge, Y. *J. Am. Mass Spectrom.* **2021**, *Just Accepted Manuscript*.

(2) Consortium for Top-Down Proteomics `https://www.topdownproteomics.org/`.

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

(4) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. *Anal. Chem.* **2018**, *90*, 110–127.

(5) Toby, T. K.; Fornelli, L.; Kelleher, N. L. *Annu. Rev. Anal. Chem.* **2016**, *9*, 499–519.

(6) Savaryn, J. P.; Catherman, A. D.; Thomas, P. M.; Abecassis, M. M.; Kelleher, N. L. *Genome Med.* **2013**, *5*, eCollection2013, DOI: `10.1186/gm457`.

(7) Gregorich, Z. R.; Ge, Y. *Proteomics* **2014**, *14*, 1195–1210.

(8) Rodriguez, H.; Zenklusen, J. C.; Staudt, L. M.; Doroshow, J. H.; Lowry, D. R. *Cell* **2021**, *184*, 1661–1670.

(9) Meding, S. et al. *J. Proteome Res.* **2012**, *11*, 1996–2003.

(10) Tsai, C. F. et al. *Mol. Cell. Proteomics* **2020**, *19*, 828–838.

(11) Zhang, B. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 256–268.

(12) Schroll, M. M.; Ludwig, K. R.; LaBonia, G. J.; Herring, E. L.; Hummon, A. B. *J. Am. Mass Spectrom.* **2018**, *29*, 2012–2022.

(13) Cai, W.; Tucholski, T. M.; Gregorich, Z. R.; Ge, Y. *Expert Rev. Proteomics* **2016**, *13*, 717–730.

(14) Seckler, H. D. S.; Park, H.; Lloyd-Jones, C. M.; Melani, R. D.; Camarillo, J. M.; Wilkins, J. T.; Compton, P. D.; Kelleher, N. L. *J. Am. Mass Spectrom.* **2021**, *32*, 1659–1670.

(15) Rinschen, M. M.; Saez-Rodriguez, J. *Nat. Rev. Nephrol.* **2021**, *17*, 205–219.

(16) Toby, T. K.; Abecassis, M.; Kim, K.; Thomas, P. M.; Fellers, R. T.; LeDuc, R. D.; Kelleher, N. L.; Demetris, J.; Levitsky, J. *Am. J. Transplant* **2017**, *17*, 2458–2467.

(17) Tucholski, T. et al. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 24691–24700.

(18) Tiambeng, T. N.; Roberts, D. S.; Brown, K. A.; Zhu, Y.; Chen, B.; Wu, Z.; Mitchell, S. D.; Guardado-Alvarez, T. M.; Jin, S.; Ge, Y. *Nat. Commun.* **2020**, *11*, 1–12.

(19) Han, M.; Wang, Y.; Cook, K.; Bala, N.; Soto, M.; Rock, D. A.; Pearson, J. T.; Rock, B. M. *Anal. Chem.* **2021**, *93*, 5562–5569.

(20) Salz, R.; Bouwmeester, R.; Gabriels, R.; Degroeve, S.; Martens, L.; Volders, P.; Hoen, P. A. C. '. *J. Proteome Res.* **2021**, *20*, 3353–3364.

(21) Howe, K. et al. *Nature* **2013**, *496*, 498–503.

(22) Kimmel, C. B.; Ballard, W. W.; Kimmel, S. R.; Ullman, B.; Schilling, T. F. *Dev. Dyn.* **1995**, *203*, 253–310.

(23) Satija, R.; Farrell, J. A.; Gennert, D.; Schier, A. F.; Regev, A. *Nature Biotechnol.* **2015**, *33*, 495–502.

(24) Mathavan, S. et al. *PLoS Genet.* **2005**, *1*, 260–276.

(25) Harvey, S. A.; Sealy, I.; Kettleborough, R.; Fenyes, F.; White, R.; Stemple, D.; Smith, J. C. *Development* **2013**, *140*, 2703–2710.

(26) Kelkar, D. S. et al. *Mol. Cell. Proteomics* **2014**, *131*, 3184–3198.

(27) Ho, Y.; Reddy, P. M. *Clin. Chem.* **2010**, *56*, 525–536.

(28) Ho, Y.; Reddy, P. M. *Mass Spectrom. Rev.* **2011**, *30*, 1203–1224.

(29) Ansong, C. et al. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 10153–10158.

(30) Dupré, M.; Duchateau, M.; Malosse, C.; Borges-lima, D.; Calvaresi, V.; Podglajen, I.; Clermont, D.; Rey, M.; Chamot-Rooke, J. *J. Proteome. Res.* **2021**, *20*, 202–211.

(31) Nie, X. et al. *Cell* **2021**, *184*, 775–791.

(32) Huge, B. J.; Champion, M. M.; Dovichi, N. J. *Anal. Chem.* **2019**, *91*, 4649–4655.

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

(34) Shaw, J. B.; Gorshkov, M. V.; Wu, Q.; Paša-Tolić, L. *Anal. Chem.* **2018**, *90*, 5557–5562.

(35) Lu, L.; Scalf, M.; Shortreed, M. R.; Smith, L. M. *Anal. Chem.* **2021**, *32*, 1319–1325.

(36) Schaffer, L. V.; Millikin, R. J.; Shortreed, M. R.; Scalf, M.; Smith, L. M. *J. Proteome Res.* **2020**, *19*, 3510–3517.

(37) Heusel, M.; Bludau, I.; Rosenberger, G.; Hafen, R.; Frank, M.; Banaei-Esfahani, A.; van Drogen, A.; Collins, B. C.; Gstaiger, M.; Aebersold, R. *Mol. Syst. Biol.* **2019**, *15*, e8438.

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

(39) McCool, E. N.; Lodge, J. M.; Basharat, A. R.; Liu, X.; Coon, J. J.; Sun, L. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 2470–2479.

(40) McCool, E. N.; Chen, D.; Li, W.; Liu, Y.; Sun, L. *Anal. Methods* **2019**, *11*, 2855–2861.

(41) McCool, E. N.; Sun, L. *Se Pu* **2019**, *37*, 878–886.

(42) Chen, D.; McCool, E. N.; Yang, Z.; Shen, X.; Lubeckyj, R. A.; Xu, T.; Wang, Q.; Sun, L. *Mass Spectrom. Rev.* **2021**, *Just Accepted Manuscript*.

(43) Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L. *Trends Anal. Chem.* **2019**, *120*, 115644.

(44) Yang, Z.; Shen, X.; Chen, D.; Sun, L. *Anal. Chem.* **2020**, *19*, 3315–3325.