Published in final edited form as:

Anal Chem. 2017 November 21; 89(22): 12059–12067. doi:10.1021/acs.analchem.7b02532.

Single-shot top-down proteomics with capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry for identification of nearly 600 *Escherichia coli* proteoforms

Rachele A. Lubeckyj[1,†], **Elijah N. McCool**[1,†], **Xiaojing Shen**[1], **Qiang Kou**[2], **Xiaowen Liu**[2,3], and **Liangliang Sun**[1,*]

[1]Department of Chemistry, Michigan State University, 578 S Shaw Ln, East Lansing, MI 48824, USA

[2]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, 719 Indiana Avenue, Indianapolis, IN 46202, USA

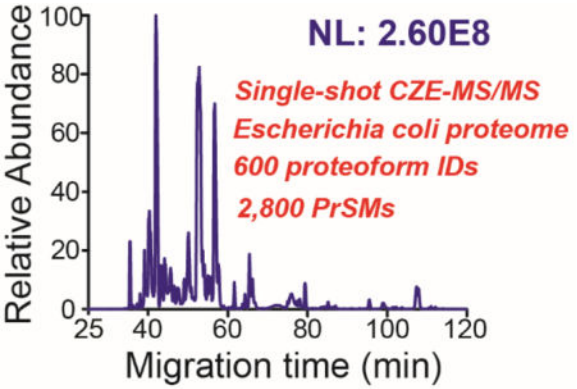
[3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 410 W. 10th Street, Indianapolis, IN 46202, USA

Abstract

Capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry (CZE-ESI-MS/MS) has been recognized as an invaluable platform for top-down proteomics. However, the scale of top-down proteomics using CZE-MS/MS is still limited due to the low loading capacity and narrow separation window of CZE. In this work, for the first time we systematically evaluated the dynamic pH junction method for focusing of intact proteins during CZE-MS. The optimized dynamic pH junction based CZE-MS/MS approached 1-μL loading capacity, 90-min separation window and high peak capacity (~280) for characterization of an *Escherichia coli* proteome. The results represent the largest loading capacity and the highest peak capacity of CZE for top-down characterization of complex proteomes. Single-shot CZE-MS/MS identified about 2,800 proteoform-spectrum matches, nearly 600 proteoforms, and 200 proteins from the *Escherichia coli* proteome with spectrum-level false discovery rate (FDR) less than 1%. The number of identified proteoforms in this work is over three times higher than that in previous single-shot CZE-MS/MS studies. Truncations, N-terminal methionine excision, signal peptide removal and some post-translational modifications including oxidation and acetylation were detected.

TOC image

*Corresponding author. lsun@chemistry.msu.edu; Phone: 517-353-0498.
†These two authors contributed equally to this work.



Introduction

Capillary zone electrophoresis (CZE)-electrospray ionization (ESI)-mass spectrometry (MS) has been well recognized for characterization of intact proteins due to its high separation efficiency.[1–4] CZE-ESI-MS/MS has been suggested as an alternative to widely used reversed-phase liquid chromatography (RPLC)-ESI-MS/MS for top-down proteomics. [5–15]

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. [5] However, the CZE-MS interface used in that work had limited lifetime and robustness, which impeded the wide application of the platform for top-down proteomics. An electrokinetically pumped sheath flow CE-MS interface with good sensitivity and robustness was developed by Dovichi group in 2010.[16] Sun *et al.* demonstrated fast, reproducible and sensitive characterization of intact proteins with the electrokinetically pumped sheath flow interface based CZE-MS/MS. [6] Later, Zhao *et al.* further applied the CZE-MS/MS system for top-down proteomics of *Mycobacterium marinum* secretome and yeast proteome.[7, 8] 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. [8] 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 (30–80 kDa), resulting in identification of 30 proteins in the mass range of 30–80 kDa. [9]

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.[17] Han *et al.* employed the sheathless interface based CZE-MS/MS for top-down proteomics of *a Pyrococcus furiosus* lysate, resulting in identification of 291 proteoforms with RPLC fractionation and CZE-MS/MS. [10] 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.[11] 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[12] and 138 proteoforms from recombinant human interferon-β1.[13] The sheathless interface based CZE-MS has also been applied for characterization of intact histones by the Lindner group.[14,15]

The current CZE-MS interfaces are robust and sensitive, enabling CZE-MS/MS to be used for top-down proteomics. However, two issues remain for CZE-MS/MS based top-down proteomics. First, the largest sample loading capacity of CZE-MS/MS systems reported in the literature for top-down proteomics is only about 200 nL.[8,10] 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 top-down proteomics is roughly 30 min.[8,10] 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 top-down proteomics 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. [18,19] However, coupling cIEF to MS is still not straightforward, which hinders its wide application for top-down proteomics.

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 bottom-up proteomics. We observed a 140-min separation window and a micro-liter scale sample loading capacity using the CZE-MS/MS system for analysis of complex proteome digests.[20] Dynamic pH junction is a simple method for sample stacking in CZE.[21,22] For instance, sample is dissolved in a basic buffer (*e.g.*, ammonium bicarbonate, pH 8) and the background electrolyte (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.[23–26] After those two pH boundaries meet, isotachophoresis (ITP) plays a role for stacking the analytes with NH_4^+ as the leading ion, followed by the typical CZE.[23]

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 top-down proteomics. 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 top-down proteomics. Zhao *et al.* performed top-down proteomics of a yeast lysate using dynamic pH junction based CZE-MS/MS.[8] 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.[8] 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 top-down proteomic analysis of *Escherichia coli*, thus leading to micro-liter scale loading capacity, 90-min separation window, high peak capacity (~280) and nearly 600 proteoform IDs with single-shot CZE-MS/MS.

Experimental section

Materials and Reagents

See Supporting Information I for details.

Sample Preparation

A mixture of standard proteins consisting of lysozyme (14.3 kDa, pI 11.0, 0.1 mg/mL), cytochrome c (Cyto.c, 12 kDa, pI 10.0, 0.1 mg/mL), myoglobin (16.9 kDa, pI 7.0, 0.1 mg/mL), β-casein (24 kDa, pI 4.5, 0.4 mg/mL), carbonic anhydrase (CA, 29 kDa, pI 5.1, 0.5 mg/mL) and bovine serum albumin (BSA, 66.5 kDa, pI 5.0, 1.0 mg/mL) was prepared in LC/MS grade water and used as a stock solution. The stock solution was diluted appropriately with different buffers for various experiments. The details for preparation of *Escherichia coli* (*E.coli*, strain K-12 substrain MG1655) samples are described in Supporting Information I.

CZE-ESI-MS/MS

An automated CZE-ESI-MS system was used in the experiments. The system contained an ECE-001 CE autosampler and a commercialized electro-kinetically pumped sheath flow CE −MS interface from CMP Scientific (Brooklyn, NY).[16,27] The CE system was coupled to a LTQ-XL or a Q-Exactive HF mass spectrometer (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 linear polyacrylamide (LPA) based on references [20] and [28]. One end of the capillary was etched with hydrofluoric acid based on reference [29] to reduce the outer diameter of the capillary. (*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 pressure 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 30–40 μm.

The detailed parameters of the LTQ-XL and Q-Exactive HF mass spectrometers are described in Supporting Information I.

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 the *E.coli* RAW files were analyzed with the TopFD[30] (TOP-Down Mass Spectrometry Feature Detection) and TopPIC (TOP-Down Mass Spectrometry Based Proteoform Identification and Characterization) pipeline.[31] TopFD is an improved version of MS-Deconv.[32] 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.[33] Then, spectral deconvolution was performed with TopFD to generate msalign files. Finally, TopPIC (version 1.1.3) was used for database searching with msalign files as input. *E. coli* (strain K12) UniProt database (UP000000625, 4307 entries, version June 7, 2017) was used for database search. The spectrum-level false discovery rate (FDR) was estimated using the target-decoy approach. [34] 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 proteoform-spectrum matches (PrSMs) were filtered with a 1% FDR at the spectrum level. 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).

Results and discussion

In order to approach large-scale top-down proteomics 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 microliter-scale sample loading capacity and 140-min separation window simultaneously for analysis of complex peptide mixtures. [20] 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 method of CZE. [6,35,36] We then optimized the dynamic pH junction based CZE-MS, and applied the optimized system for top-down proteomics of an *E. coli* proteome.

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) acetic acid (BGE) for control, 2.5% (v/v) acetic acid in water containing 35% (v/v) acetonitrile (lower conductivity than BGE) for FESS, and 10 mM NH_4HCO_3 (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) acetic acid (pH 2.4). Choices for the BGE and the sample buffer for dynamic pH junction method was based on references [8] and [20].

Figure 1 summarizes the results of the comparison experiments. Figure 1A–1C shows the change of protein intensity as a function of sample injection volume 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. The *m/z* used for extraction of protein peaks are presented in the legend of Figure 1. Figure 1D shows the EIEs 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 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. 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, but they co-migrated out of the separation capillary when the sample injection volume increased to 500 nL.

As shown in Figure 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 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 separation in control experiments using the 500-nL sample injection. 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 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 signicant 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 of the five proteins from dynamic pH junction experiments were comparable to that from FESS experiments with the same sample injection volume that was 50, 100 or 200 nL. However, on average, the intensity of those proteins was improved by 80% using dynamic pH junction method compared with FESS method with 500-nL sample injection. As shown in Figure 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 [6,37] 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 [6], which were the monoisotopic masses of those three forms exported from MS-Deconv software.[32] 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, Figure S1 in Supporting Information I. The stock solution of the standard protein mixture was diluted with NH_4HCO_3 buffers by four different dilution factors that were 2, 6, 18 and 54, respectively. All of the dilute samples were dissolved in 10 mM NH_4HCO_3 (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 top-down proteomics.

Optimization of the dynamic pH junction based CZE-MS

We chose 10 mM NH_4HCO_3 (pH 8.0) as the sample buffer for dynamic pH junction based CZE-MS at the beginning based on references [20] and [38]. Imami *et al.* systematically investigated the effect of the concentration of NH_4HCO_3 in the sample buffer on the concentration performance of dynamic pH junction method using a peptide mixture.[25] They increased the concentration of NH_4HCO_3 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. [20] When we increased the concentration of NH_4HCO_3 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 NH_4HCO_3 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.[39–41] Therefore, we tested 50 mM NH_4HCO_3 (pH 8.0) as the sample buffer for dynamic pH junction based CZE-MS using the mixture of the six standard proteins, Figure 2 and Figure S2 in Supporting Information I. The BGE was 5% (v/v) acetic acid. The stock solution of the standard protein mixture was diluted with a NH_4HCO_3 buffer (pH 8.0) by a factor of 10, and the concentration of NH_4HCO_3 in the dilute sample was 50 mM. The dilute sample was used for all of the following experiments.

As shown in Figure S2, all of the six proteins (lysozyme, BSA, cyto.c, myoglobin, CA and β-casein) in the protein mixture were well separated with 500-nL sample injection and even 1-μL sample injection, which corresponded to 50% of the total capillary volume. Figure 2 shows the corresponding EIEs of the standard protein mixture using 500-nL sample injection (Figure 2A) and 1-μL sample injection (Figure 2B). The CZE-MS system using 50 mM NH_4HCO_3 as the sample buffer produced high separation efficiency for proteins with both 500-nL and 1-μL sample injection. As shown in Figure 2C, 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 NH_4HCO_3 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 NH_4HCO_3 and 50 mM NH_4HCO_3 as the sample buffers based on the EIEs in Figure 1D and Figure 2A. 50 mM NH_4HCO_3 sample buffer generated, on average, comparable intensity of proteins with 5-times lower protein concentration compared with 10 mM NH_4HCO_3 sample buffer. Therefore, we chose 50 mM NH_4HCO_3 (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) formic acid and 5–10% (v/v) acetic acid. The sample injection volume was 500 nL per CZE-MS run. We observed that the overall performance of 0.1% (v/v) formic acid BGE (pH 2.8) was better than that of 0.3% and 0.5% (v/v) formic acid (pH 2.3 and 2.1) in terms of protein intensity, Figure S3A in Supporting Information I. We also observed comparable protein intensity from 0.1% (v/v) formic acid BGE and acetic acid BGEs (5% and 10% (v/v)), Figure S3B and Figure S3C in

Supporting Information I. However, 5% and 10% (v/v) acetic acid BGEs (pH ~2.4 and ~2.2) produced significantly wider separation window than 0.1% (v/v) formic acid BGE for the standard protein mixture. In addition, the migration time of protein analytes in the capillary in 5% and 10% (v/v) acetic acid BGEs was significantly longer than that in 0.1% (v/v) formic acid BGE (*e.g.*, 10 minutes longer for lysozyme). There are two potential reasons for the phenomenon. First, 5–10% (v/v) acetic acid has much lower pH than 0.1% (v/v) formic acid (2.4–2.2 *vs.* 2.8), which further reduces the remaining electroosmotic flow in the LPA-coated separation capillary. We measured the electroosmotic mobility in the LPA-coated capillary based on the method reported in literature. [42,43] The electroosmotic mobility in 10% (v/v) acetic acid BGE was lower than that in 0.1% (v/v) formic acid BGE (6.8×10^{-6} $cm^2\cdot V^{-1}\cdot s^{-1}$ *vs.* 1.1×10^{-5} $cm^2\cdot V^{-1}\cdot s^{-1}$). The details about measurement of the electroosmotic mobility is described in Supporting Information I. Second, the more acidic BGEs (5–10% (v/v) acetic acid) typically lead 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) acetic acid and 50 mM NH_4HCO_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% (v/v) acetic acid as the BGE and 500-nL sample injection. 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 S1 in Supporting Information I). 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 experience, suggesting that the LPA coating on the inner wall of the separation capillary is stable.

Single-shot top-down proteomics with CZE-MS/MS

We further applied the optimized CZE-MS/MS system for top-down proteomics of *E.coli*. An *E.coli* protein sample (2 mg/mL) in 50 mM NH_4HCO_3 (pH 8.0) was used for the experiments. 10% (v/v) acetic acid was used as the BGE. A Q-Exactive HF 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 3A. 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 injection. 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 voltage and 500-nL sample loading, Figures 3B and 3C. 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 are shown in Supporting Information II. The corresponding raw files have been deposited to the ProteomeXchange Consortium via the PRIDE [44] partner repository with the dataset identifier PXD007273. 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).[8,10] Results clearly demonstrate the capability of CZE-MS/MS for large-scale top-down proteomics. 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 S4 in Supporting Information I. The molecular weight of identified proteoforms ranged from around 2,000 Da to about 24,000 Da. About 33% of the identified proteoforms had molecular weight higher than 10 kDa.

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. [8,10] Third, the dynamic pH junction based CZE produced high peak capacity for separation of the *E.coli* proteome. Based on the electropherograms in Figure 3B, the peak capacity of the system was about 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. [8,10]

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 4A. 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 4B 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 4C and 4D show sequences and observed fragmentation patterns 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 4C), while there was N-terminal methionine excision that was detected for 30S ribosomal protein S17 (Figure 4D).

Conclusions

We presented a CZE-MS/MS system with microliter scale sample loading capacity, 90-min separation window and high peak capacity (~280) for large-scale top-down proteomics, 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 21 T FT-ICR mass spectrometer.[45] This CZE-MS/MS system established the foundation for large-scale top-down proteomics using CZE-MS/MS.

RPLC-MS/MS is typically used for large-scale top-down proteomics, 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.[11] In addition, CZE can separate protein(s)/protein complexes in native condition.[46,47] 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.[47] The results demonstrate the potential of CZE-MS/MS for top-down proteomics of complex proteomes in native conditions.

However, CZE-MS/MS based large-scale top-down proteomics is still at the early stage. The 600 proteoform IDs in this work represents the largest top-down proteomics dataset using CZE-MS/MS. The number of proteoform IDs from CZE-MS/MS is still far away from the state of the art of LC-MS/MS based top-down proteomics, which has reached thousands of proteoform IDs from mammalian cell lines. [45, 48–52] Around 1,000 proteoform IDs from complex proteome samples has been approached using one-dimension high-resolution RPLC-MS/MS. [53,54] To improve the scale of CZE-MS/MS based top-down proteomics, we need to further improve the CZE-MS/MS system in terms of the separation window and sample loading capacity. One solution is to use longer separation capillary (*e.g.*, 1.5 meters) and higher separation voltage (*e.g.*, 60 kV or higher). In addition, coupling LC fractionation (size exclusion chromatography and RPLC) to CZE-MS/MS will enable high capacity separation of complex proteomes, and will significantly improve the scale of top-down proteomics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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. Qiang Kou and Xiaowen Liu were supported by the National Institute of General Medical Sciences, National Institutes of Health (NIH) through Grant R01GM118470.

References

1. Domínguez-Vega E, Haselberg R, Somsen GW. Methods Mol Biol. 2016; 1466:25–41. [PubMed: 27473479]
2. Haselberg R, de Jong GJ, Somsen GW. Electrophoresis. 2013; 34:99–112. [PubMed: 23161520]
3. Jorgenson JW, Lukacs KD. Science. 1983; 222:266–272. [PubMed: 6623076]
4. Harstad RK, Johnson AC, Weisenberger MM, Bowser MT. Anal Chem. 2016; 88:299–319. [PubMed: 26640960]
5. Valaskovic GA, Kelleher NL, McLafferty FW. Science. 1996; 273:1199–1202. [PubMed: 8703047]
6. Sun L, Knierman MD, Zhu G, Dovichi NJ. Anal Chem. 2013; 85:5989–5995. [PubMed: 23692435]
7. Zhao Y, Sun L, Champion MM, Knierman MD, Dovichi NJ. Anal Chem. 2014; 86:4873–4878. [PubMed: 24725189]
8. Zhao Y, Sun L, Zhu G, Dovichi NJ. J Proteome Res. 2016; 15:3679–3685. [PubMed: 27490796]
9. Li Y, Compton PD, Tran JC, Ntai I, Kelleher NL. Proteomics. 2014; 14:1158–1164. [PubMed: 24596178]
10. Han X, Wang Y, Aslanian A, Bern M, Lavallée-Adam M, Yates JR III. Anal Chem. 2014; 86:11006–11012. [PubMed: 25346219]
11. Han X, Wang Y, Aslanian A, Fonslow B, Graczyk B, Davis TN, Yates JR III. J Proteome Res. 2014; 13:6078–6086. [PubMed: 25382489]
12. Haselberg R, de Jong GJ, Somsen GW. Anal Chem. 2013; 85:2289–2296. [PubMed: 23323765]
13. Bush DR, Zang L, Belov AM, Ivanov AR, Karger BL. Anal Chem. 2016; 88:1138–1146. [PubMed: 26641950]
14. Faserl K, Sarg B, Maurer V, Lindner HH. J Chromatogr A. 2017; 1498:215–223. [PubMed: 28179079]
15. Sarg B, Faserl K, Kremser L, Halfinger B, Sebastiano R, Lindner HH. Mol Cell Proteomics. 2013; 12:2640–2656. [PubMed: 23720761]
16. Wojcik R, Dada OO, Sadilek M, Dovichi NJ. Rapid Commun Mass Spectrom. 2010; 24:2554–2560. [PubMed: 20740530]
17. Moini M. Anal Chem. 2007; 79:4241–4246. [PubMed: 17447730]
18. Yang L, Lee CS, Hofstadler SA, Pasa-Tolic L, Smith RD. Anal Chem. 1998; 70:3235–3241. [PubMed: 11013724]
19. Jensen PK, Pasa-Tolić L, Anderson GA, Horner JA, Lipton MS, Bruce JE, Smith RD. Anal Chem. 1999; 71:2076–2084. [PubMed: 10366890]
20. Chen D, Shen X, Sun L. Analyst. 2017; 142:2118–2127. [PubMed: 28513658]
21. Aebersold R, Morrison HD. J Chromatogr. 1990; 516:79–88. [PubMed: 2286630]
22. Britz-McKibbin P, Chen DDY. Anal Chem. 2000; 72:1242–1252. [PubMed: 10740866]
23. Wang L, MacDonald D, Huang X, Chen DD. Electrophoresis. 2016; 37:1143–1150. [PubMed: 26949078]
24. Cao CX, Fan LY, Zhang W. Analyst. 2008; 133:1139–1157. [PubMed: 18709186]
25. Imami K, Monton MR, Ishihama Y, Terabe S. J Chromatogr A. 2007; 1148:250–255. [PubMed: 17382949]

26. Ptolemy AS, Britz-McKibbin P. Analyst. 2008; 133:1643–1648. [PubMed: 19082065]
27. Sun L, Zhu G, Zhang Z, Mou S, Dovichi NJ. J Proteome Res. 2015; 14:2312–2321. [PubMed: 25786131]
28. Zhu G, Sun L, Dovichi NJ. Talanta. 2016; 146:839–843. [PubMed: 26695337]
29. Sun L, Zhu G, Zhao Y, Yan X, Mou S, Dovichi NJ. Angew Chem Int Ed. 2013; 52:13661–13664.
30. http://proteomics.informatics.iupui.edu/software/topfd/. Date of access: June 8, 2017
31. Kou Q, Xun L, Liu X. Bioinformatics. 2016; 32:3495–3497. [PubMed: 27423895]
32. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, Whitelegge JP, Bafna V, Pevzner PA. Mol Cell Proteomics. 2010; 9:2772–2782. [PubMed: 20855543]
33. Kessner D, Chambers M, Burke R, Agus D, Mallick P. Bioinformatics. 2008; 24:2534–2536. [PubMed: 18606607]
34. Elias JE, Gygi SP. Nature Methods. 2007; 4:207–214. [PubMed: 17327847]
35. Sun L, Hebert AS, Yan X, Zhao Y, Westphall MS, Rush MJ, Zhu G, Champion MM, Coon JJ, Dovichi NJ. Angew Chem Int Ed. 2014; 53:13931–13933.
36. Simpson SL Jr, Quirino JP, Terabe S. J Chromatogr A. 2008; 1184:504–541. [PubMed: 18035364]
37. Wu S, Lourette NM, Tolić N, Zhao R, Robinson EW, Tolmachev AV, Smith RD, Pasa-Tolić L. J Proteome Res. 2009; 8:1347–1357. [PubMed: 19206473]
38. Zhu G, Sun L, Heidbrink-Thompson J, Kuntumalla S, Lin HY, Larkin CJ, McGivney JB IV, Dovichi NJ. Electrophoresis. 2016; 37:616–622. [PubMed: 26530276]
39. Busnel JM, Schoenmaker B, Ramautar R, Carrasco-Pancorbo A, Ratnayake C, Feitelson JS, Chapman JD, Deelder AM, Mayboroda OA. Anal Chem. 2010; 82:9476–9483. [PubMed: 21028888]
40. Faserl K, Kremser L, Müller M, Teis D, Lindner HH. Anal Chem. 2015; 87:4633–4640. [PubMed: 25839223]
41. Wang Y, Fonslow BR, Wong CC, Nakorchevsky A, Yates JR 3rd. Anal Chem. 2012; 84:8505–8513. [PubMed: 23004022]
42. Williams BA, Vigh G. Anal Chem. 1996; 68:1174–1180. [PubMed: 21619150]
43. Zhang Z, Peuchen EH, Dovichi NJ. Anal Chem. 2017; 89:6774–6780. [PubMed: 28540730]
44. Vizcaíno JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu QW, Wang R, Hermjakob H. Nucleic Acids Res. 2016; 44:D447–D456. [PubMed: 26527722]
45. Anderson LC, DeHart CJ, Kaiser NK, Fellers RT, Smith DF, Greer JB, LeDuc RD, Blakney GT, Thomas PM, Kelleher NL, Hendrickson CL. J Proteome Res. 2017; 16:1087–1096. [PubMed: 27936753]
46. Nguyen A, Moini M. Anal Chem. 2008; 80:7169–7173. [PubMed: 18710259]
47. Belov AM, Viner R, Santos MR, Horn DM, Bern M, Karger BL, Ivanov AR. J Am Soc Mass Spectrom. 2017; doi: 10.1007/s13361-017-1781-1
48. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M, Wu C, Sweet SM, Early BP, Siuti N, LeDuc RD, Compton PD, Thomas PM, Kelleher NL. Nature. 2011; 480:254–258. [PubMed: 22037311]
49. Fornelli L, Durbin KR, Fellers RT, Early BP, Greer JB, LeDuc RD, Compton PD, Kelleher NL. J Proteome Res. 2017; 16:609–618. [PubMed: 28152595]
50. Durbin KR, Fornelli L, Fellers RT, Doubleday PF, Narita M, Kelleher NL. J Proteome Res. 2016; 15:976–982. [PubMed: 26795204]
51. Cai W, Tucholski T, Chen B, Alpert AJ, McIlwain S, Kohmoto T, Jin S, Ge Y. Anal Chem. 2017; 89:5467–5475. [PubMed: 28406609]
52. Valeja SG, Xiu L, Gregorich ZR, Guner H, Jin S, Ge Y. Anal Chem. 2015; 87:5363–5371. [PubMed: 25867201]
53. Ansong C, Wu S, Meng D, Liu X, Brewer HM, Deatherage Kaiser BL, Nakayasu ES, Cort JR, Pevzner P, Smith RD, Heffron F, Adkins JN, Pasa-Tolic L. Proc Natl Acad Sci USA. 2013; 110:10153–10158. [PubMed: 23720318]

54. Shen Y, Tolić N, Piehowski PD, Shukla AK, Kim S, Zhao R, Qu Y, Robinson E, Smith RD, Paša-Tolić L. J Chromatogr A. 2017; 1498:99–110. [PubMed: 28077236]

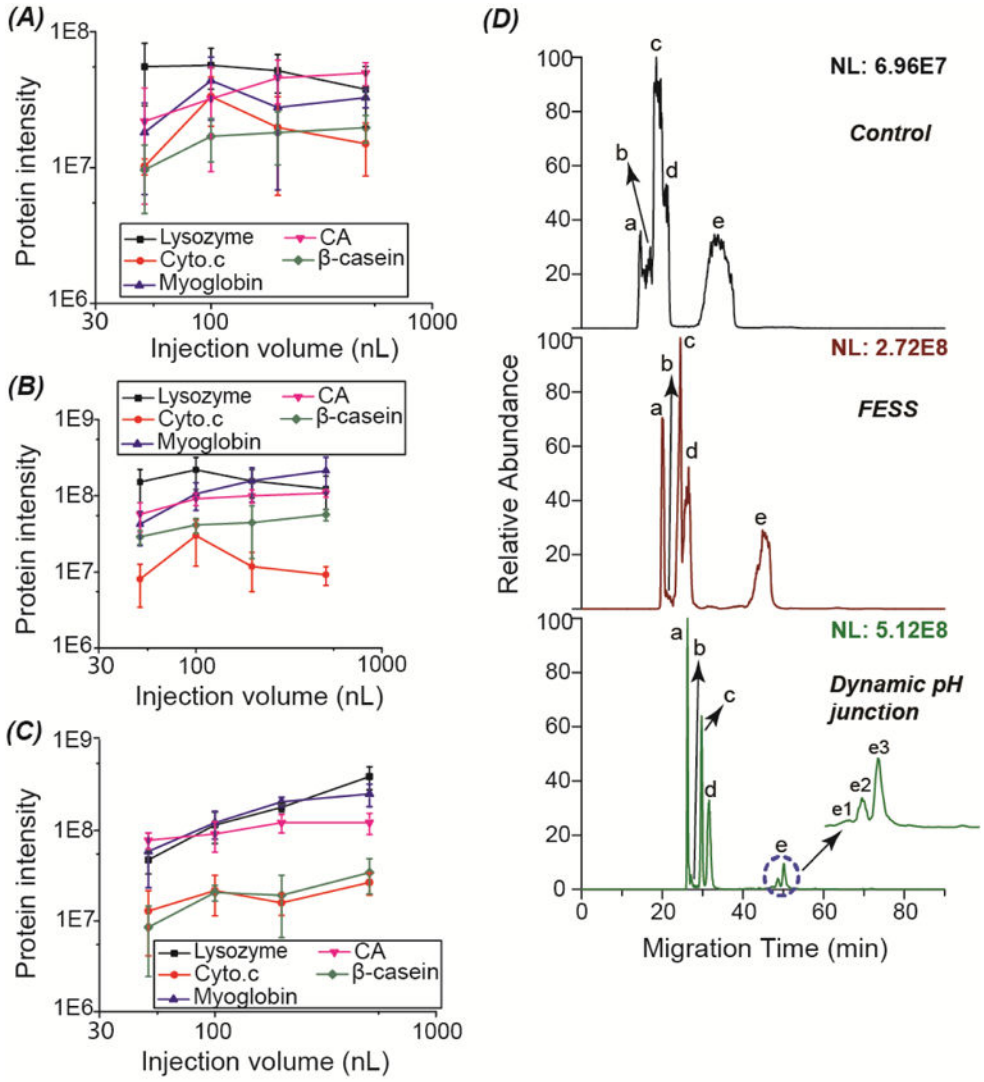
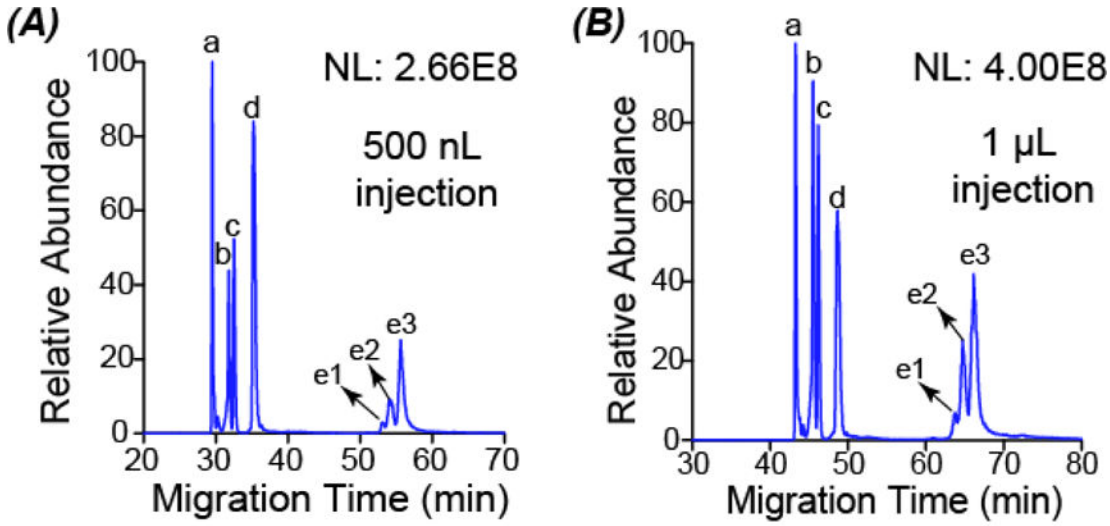


Figure 1.

Protein intensity change across different sample injection volumes (50, 100, 200 and 500 nL) for control (A), FESS (B) and dynamic pH junction (C). All the protein intensities were obtained from extracted ion electropherograms (EIEs). The error bars represent the standard deviations of protein intensity from triplicate CZE-MS analyses. (D) EIEs of the mixture of standard proteins from CZE-MS under the three different conditions. The sample injection volume was 500 nL for each condition. The proteins labelled in the electropherograms are lysozyme (a), cyto.c (b), myoglobin (c), CA (d) and β-casein (e). 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. The mass tolerance was 100 ppm. A Q-Exactive HF mass spectrometer was used for all of the experiments. For CZE, 30 kV was applied at the injection end for separation.



(C)

Peak	Protein	N	
		500 nL	*1 µL*
a	Lysozyme	206,000	292,000
b	Cyto.c	101,000	187,000
c	Myoglobin	95,000	213,000
d	CA	31,000	41,000
e1	β-casein, peak e1	36,000	45,000
e2	β-casein, peak e2	21,000	46,000
e3	β-casein, peak e3	41,000	30,000

Figure 2.

Extracted ion electropherograms (EIEs) of the standard protein mixture dissolved in 50 mM NH_4HCO_3 (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). The m/z used for protein peak extraction were the same as those in Figure 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. A Q-Exactive HF mass spectrometer was used for all of the experiments. For CZE separation, 30 kV was applied at the injection end.

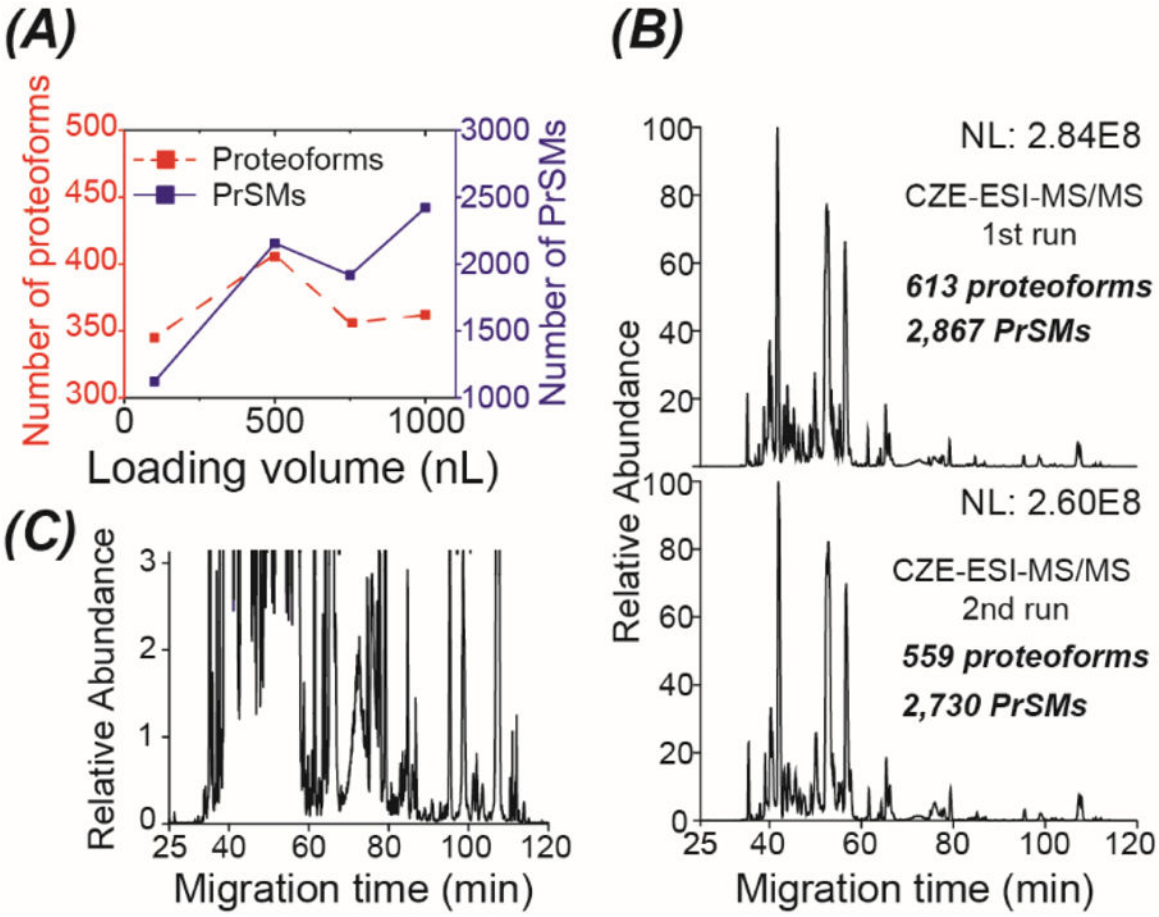


Figure 3.

Data about top-down proteomics 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). For the CZE-MS experiments, 30 kV was applied at the injection end for separation. (B) Electropherograms of the *E.coli* protein sample analyzed by CZE-MS/MS in duplicate runs. For the CZE-MS experiments, 20 kV was applied at the injection end for separation. (C) The zoom-in electropherogram of the *E.coli* protein sample from the 1st run CZE-MS/MS in (B). A Q-Exactive HF mass spectrometer was used for all of the experiments.

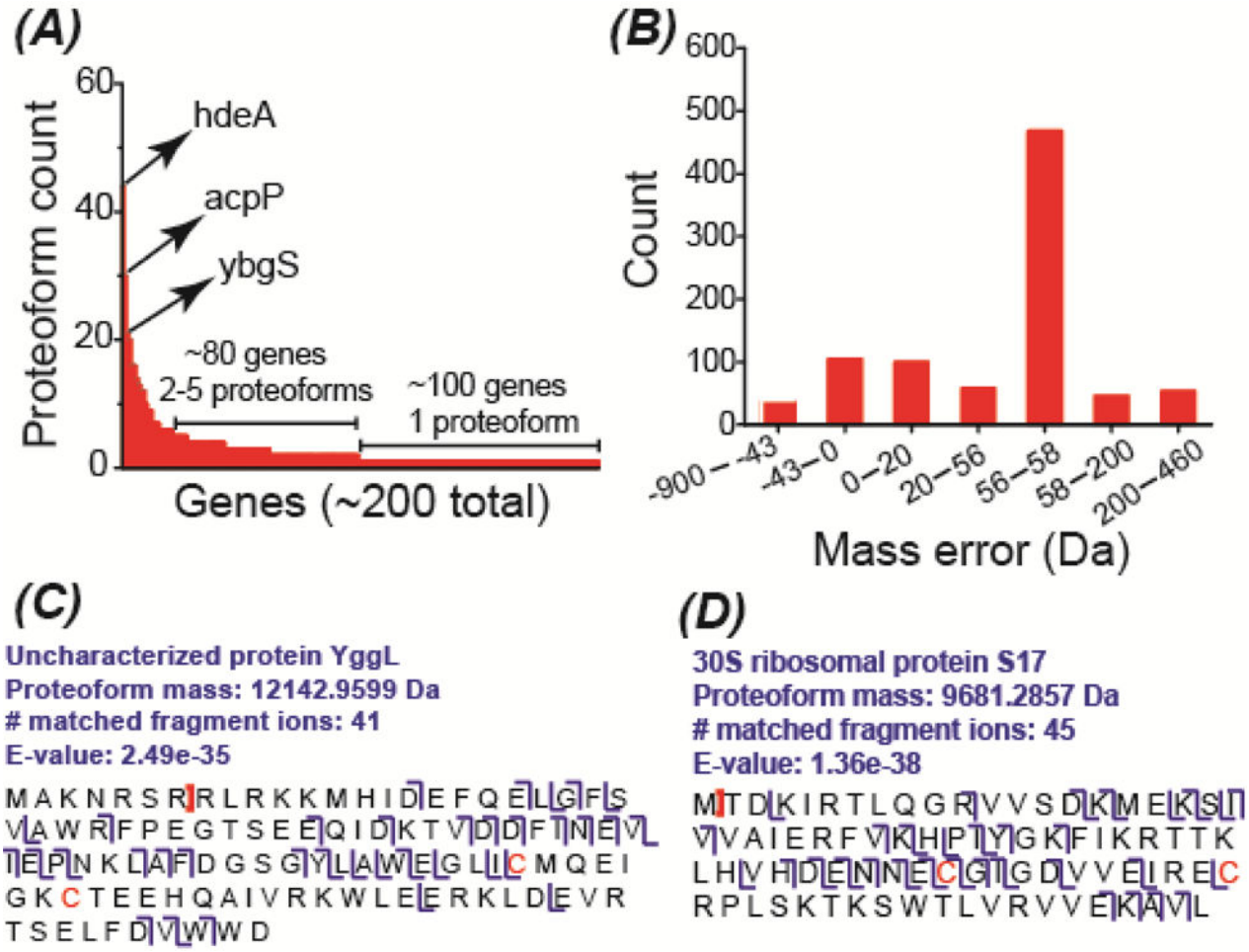


Figure 4.

(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. The single-shot *E.coli* data in Figure 3B was used for these analyses.