Supporting Information I

Native proteomics in discovery mode using size exclusion chromatography-capillary zone electrophoresis-tandem mass spectrometry

Xiaojing Shen, Qiang Kou, Ruiqiong Guo, Zhichang Yang, Daoyang Chen, Xiaowen Liu, Heedeok Hong and Liangliang Sun *

* Corresponding author. E-mail: lsun@chemistry.msu.edu

Table of Contents

Experimental Procedures

Materials and chemicals

3-(Trimethoxysilyl)propyl methacrylate, ammonium persulfate, ammonium acetate (NH_4AC) and the Microcon-30kDa centrifugal filter units for buffer exchange were purchased from Sigma-Aldrich (St. Louis, MO). Hydrofluoric acid (HF) and LC/MS grade water were purchased from Fisher Scientific (Pittsburgh, PA). Acrylamide were purchased from Acros Organics (NJ, USA). Bare fused silica capillaries (50-μm i.d., 360-μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ).

Preparation of LPA-coated separation capillary for CZE

The inner wall of the separation capillary (50-μm i.d., 360-μm o.d.) was coated with LPA based on the protocol described in references [1] and [2]. A bare fused silica capillary was successively flushed with 1 M hydrochloric acid, water, 1 M sodium hydroxide, water, and methanol, followed by treatment with 3-(trimethoxysilyl) propyl methacrylate to introduce carbon-carbon double bonds on the inner wall of the capillary. The treated capillary was filled with degassed acrylamide solution in water containing ammonium persulfate, followed by incubation at 50 ˚C water bath for 35 to 40 min with both ends sealed by silica rubber. After that, the capillary was flushed with water to remove the unreacted reagents. Then one end of the LPA-coated capillary was etched with HF based on the protocol in reference [3] to reduce its outer diameter to around 70 μm.

Sample preparation

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

One aliquot of the *E. coli* lysate containing about 600 µg of proteins (~2 mg/mL) was fractionated with size exclusion chromatography (SEC) on an Agilent Infinity II HPLC system. The SEC column (4.6 x 300 mm, 2.7 µm particles, 300 Å pores) was from Agilent. The mobile phase was 100 mM NH_4Ac (pH 7.0), and the flow rate was 0.15 mL/min. 8 fractions were collected from 11-19 min (1 min for each fraction) for relatively small proteins based on our preliminary experiment. Then each fraction was loaded onto a Microcon-30 kDa centrifugal filter unit, respectively, followed by centrifugation to remove the lysis buffer. We washed the membrane with 50 mM NH_4Ac (pH 6.9) for buffer exchange, followed by adding 40 µL of 50 mM NH_4Ac (pH 6.9) into each filter unit to extract the proteins on the membrane. We gently vortexed the filter units for 5 min and took the protein solution from the filter units for native CZE-MS/MS analysis. The use of Microcon-30 kDa centrifugal filter unit for buffer exchange was based on the recent native proteomics work from the Kelleher group.[4]

SDS-PAGE

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

Native CZE-ESI-MS and MS/MS analysis

An ECE-001 capillary electrophoresis autosampler (CMP Scientific, Brooklyn, NY) was used for automated operation of CZE. A commercialized electrokinetically pumped sheath flow interface (CMP Scientific) was used to couple CZE to MS. [5,6] A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) were used for the experiments. The commercialized electrokinetically pumped sheath flow interface (CMP Scientific) was directly attached to the Q-Exactive HF mass spectrometer for experiments. The ESI emitters of the CZE-MS interface were pulled from borosilicate glass capillaries (1.0 mm o.d., 0.75 mm i.d., 10 cm length) with a Sutter P-1000 flaming/brown micropipet puller. The opening size of the ESI emitters was 20 μm. The spray emitter with ~4 cm length was typically used. Voltage for ESI was ~2 kV.

A 1-meter LPA coated capillary (50-μm i.d. and 360-μm o.d.) was used for the CZE. The background electrolyte (BGE) for CZE was 50 mM NH_4Ac (pH 6.9), and the sheath buffer was 25 mM NH_4AC (pH 6.9). 15 kV was applied at the sample injection end and 1 psi was applied at the mean time for CZE separation. The *E. coli* sample was injected into the separation capillary for CZE-MS/MS with 5-psi pressure for 20 s.

A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used for all of the experiments. The transfer capillary temperature was 200 ^{o}C, and the S-lens RF level was 50. A top3 data dependent acquisition (DDA) method was used. The number of microscans was 3 for both MS and MS/MS. The resolution for MS and MS/MS was 240,000 and 120,000 (*m/z* 200), respectively. The AGC target was 3E6 for MS and 1E6 for MS/MS. The maximum injection time was 200 ms for MS and 500 ms for MS/MS. The mass range for MS scans was 1000-4000 *m/z*. Three most abundant protein peaks in the mass spectrum were sequentially isolated with isolation window as 4 *m/z* in the quadrupole, followed by fragmentation with normalized collisional energy (NCE) as 35. “Exclude isotopes” was turned on and the dynamic exclusion was 30 s.

Data analysis

The RAW files were first converted into mzML files using Msconvert tool.[7] Then, TopFD (TOP-Down Mass Spectrometry Feature Detection) was used for the spectral deconvolution to produce msalign files. TopFD (http://proteomics.informatics.iupui.edu/software/toppic/) is an improved version of MS-

Deconv. [8] Finally, TopPIC (version 1.1.3) [9] was used for database search with msalign files as input. The *E. coli* UniProt database (UP000000625, 4307 entries) was used for database search. The false discovery rates (FDRs) were estimated using the target-decoy approach. [10] The database search parameters were as follows: the maximum number of unexpected modifications as 2, the precursor and fragment mass error tolerances as 15 ppm, and the mass shift of unknown modifications as -200 to 10000 Da. In order to reduce the redundancy of proteoform identifications (IDs), proteoforms identified by multiple spectra were considered as one proteoform ID if those spectra match the same proteoform feature reported by TopFD or those proteoforms belong to the same protein and have similar precursor masses (within 1.2 Da).

Two rounds of analyses were performed. TopPIC was employed to search each raw file against the *E. coli* database separately, and no filter was applied in this step. Then, all the proteoform spectrum-matches (PrSMs) identified from the 8 SEC fractions were combined and filtered out with a 1% spectrum-level FDR. The identified proteoforms were further filtered with a 5% proteoform-level FDR. The identified proteoforms are listed in **Supporting Information II**.

Workflow for identification of protein complexes

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

Second, we performed a database search of the acquired MS and MS/MS spectra using TopPIC to identify proteoforms.

Third, we believe if one proteoform is a complex with some co-factor, after database search there should be a detected mass shift that matches with the mass of the co-factor. We obtained a potential protein cofactor list from the UniProt *E.coli* database, as shown in **Table S2**. Here we take RNA polymerase-binding transcription factor DksA as

an example. We identified this protein by TopPIC and obtained the proteoform as shown in **Figure S3**. We found it has an unknown modification of ~63.5 Da, which is close to the mass of zinc or copper. We think the proteoform should be a potential protein complex with a zinc ion or a copper ion.

Fourth, in order to confirm this modification (+63.5) is not an unusual covalent modification, we compared the proteoform with our recently published large-scale top-down proteomics dataset of *E. coli* under a denaturing condition. [11] If the proteoform matches with some proteoform identified under the denaturing condition in terms of the mass shift within a 4-Da mass tolerance, we think the modification (+63.5) should be some covalent modification and the proteoform is not a protein-metal complex. If we did not observe any proteoform similar to the proteoform identified in this work, we conclude the proteoform should be a protein complex.

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

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

Table S1. The list of the identified protein complexes with the SEC-CZE-MS/MS from the *E. coli* proteome.

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

Table S2. The names and masses of the major protein co-factors in the UniProt *E. coli* database.

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

Table S3. The metal binding stoichiometry of some identified metalloproteins.

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

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

Table S4. The list of some of the post-translational modifications (PTMs) detected in this work.

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

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

Table S5. The list of proteins with unreported signal peptide cleavage and initial methionine excision.

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

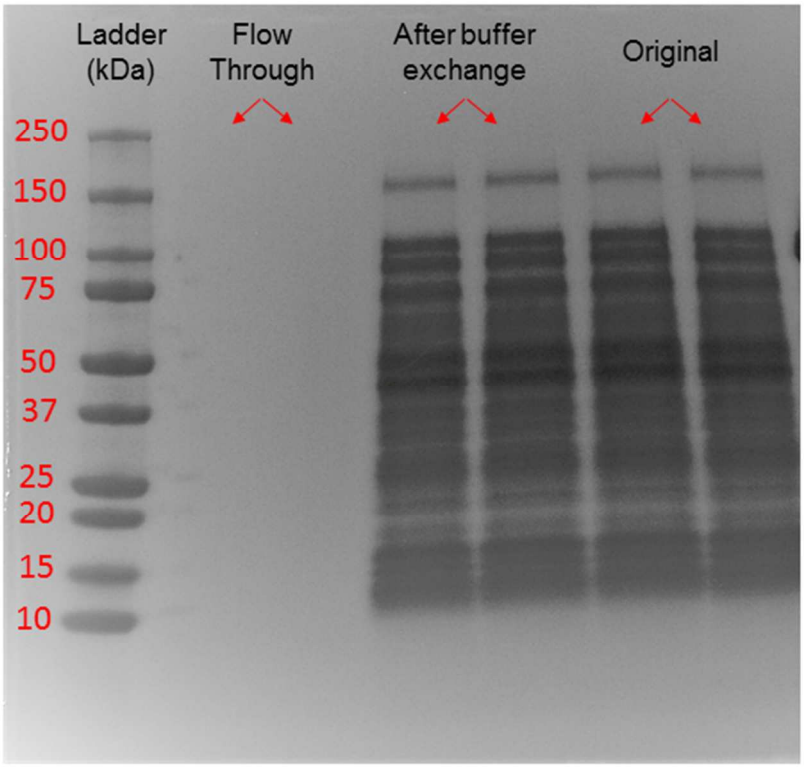

Figure S1. Image of the SDS-PAGE result. *E. coli* cell lysate before (Original) and after the buffer exchange with Microcon-30 kDa centrifugal filter units were analyzed by SDS-PAGE. About 16 μg of proteins in theory were loaded. The flow through during buffer exchange was also analyzed. We performed the buffer exchange experiment in technical duplicate and the data were shown as the two channels.

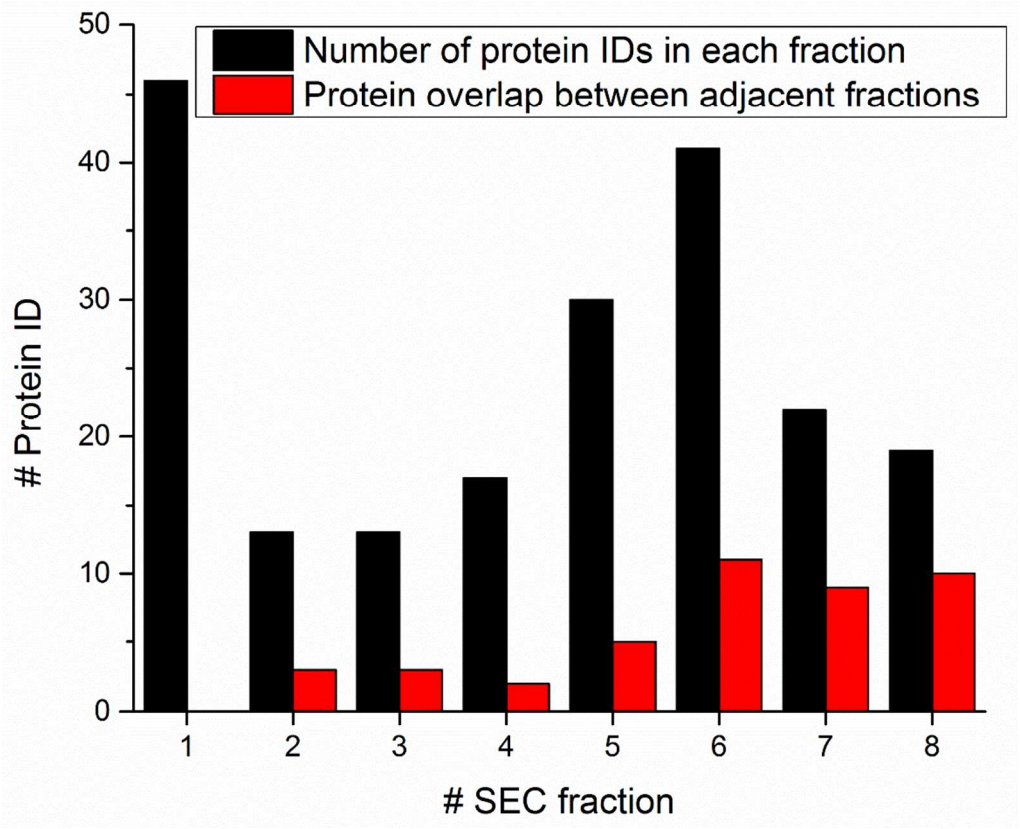

Figure S2. The number of protein identifications (IDs) from each SEC fraction and protein overlaps between adjacent SEC fractions.

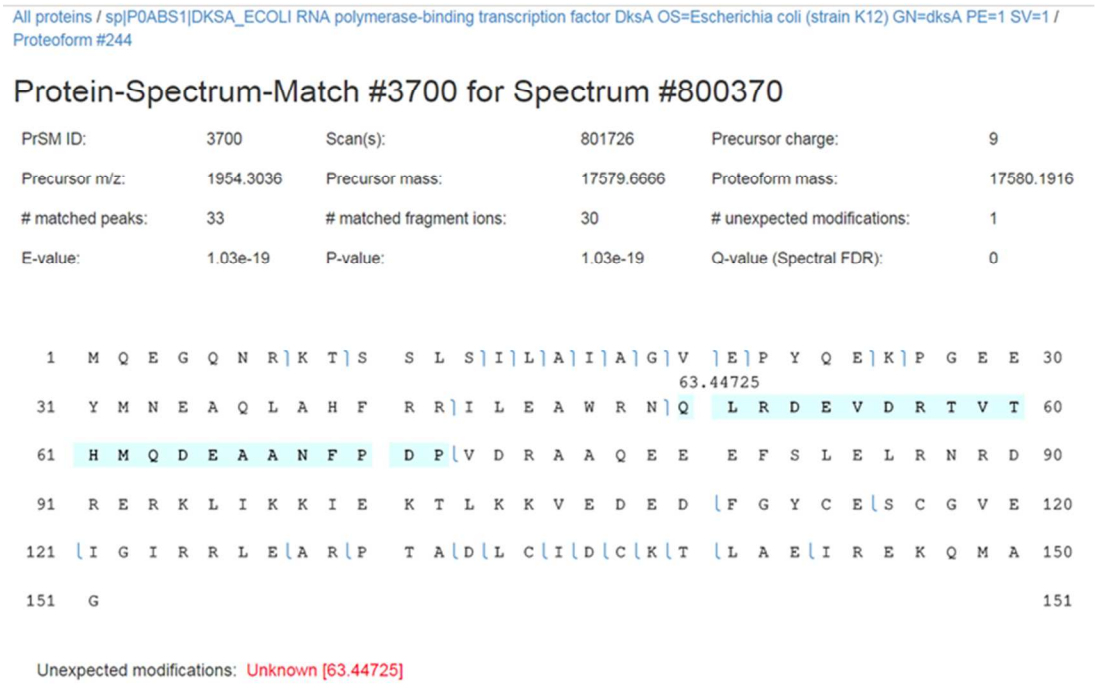

Figure S3. The sequence of the RNA polymerase-binding transcription factor DksA, the observed fragmentation pattern, and the mass shift detected through the database search.

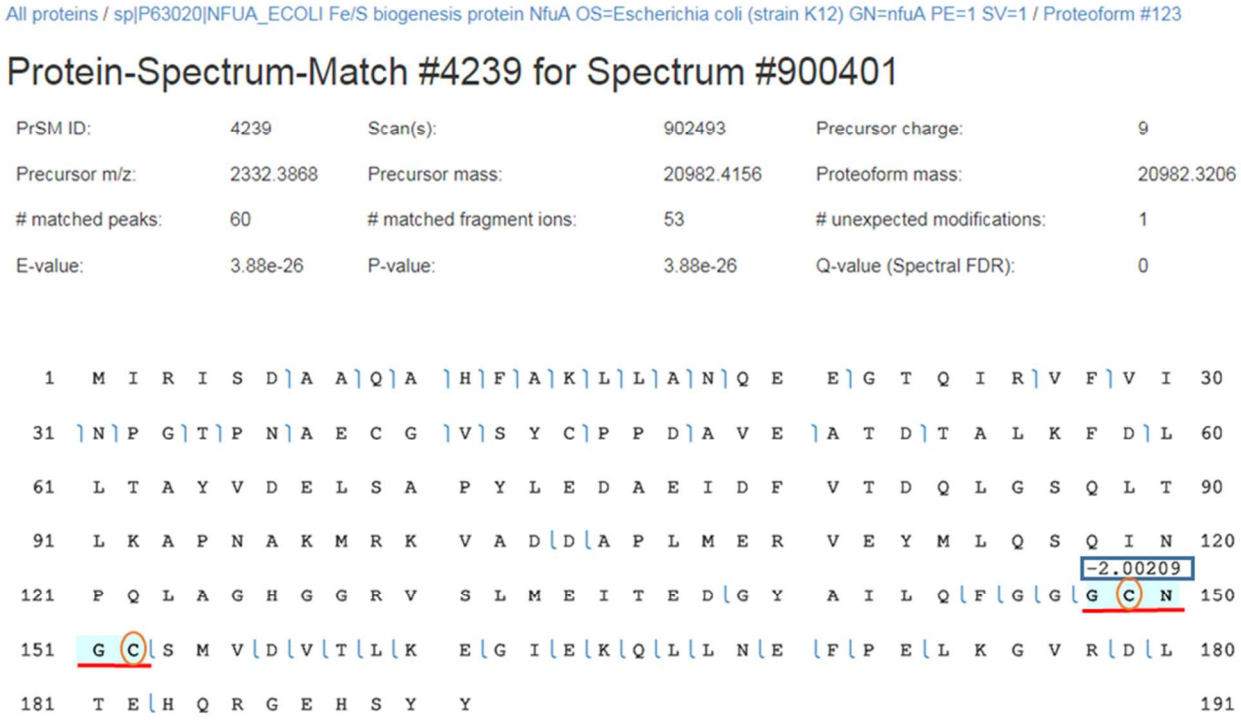

Figure S4. The sequence of the Fe/S biogenesis protein NfuA, the observed fragmentation pattern, and the mass shift detected through the database search. The mass shift, location of the mass shift, and the cysteine amino acids were highlighted.

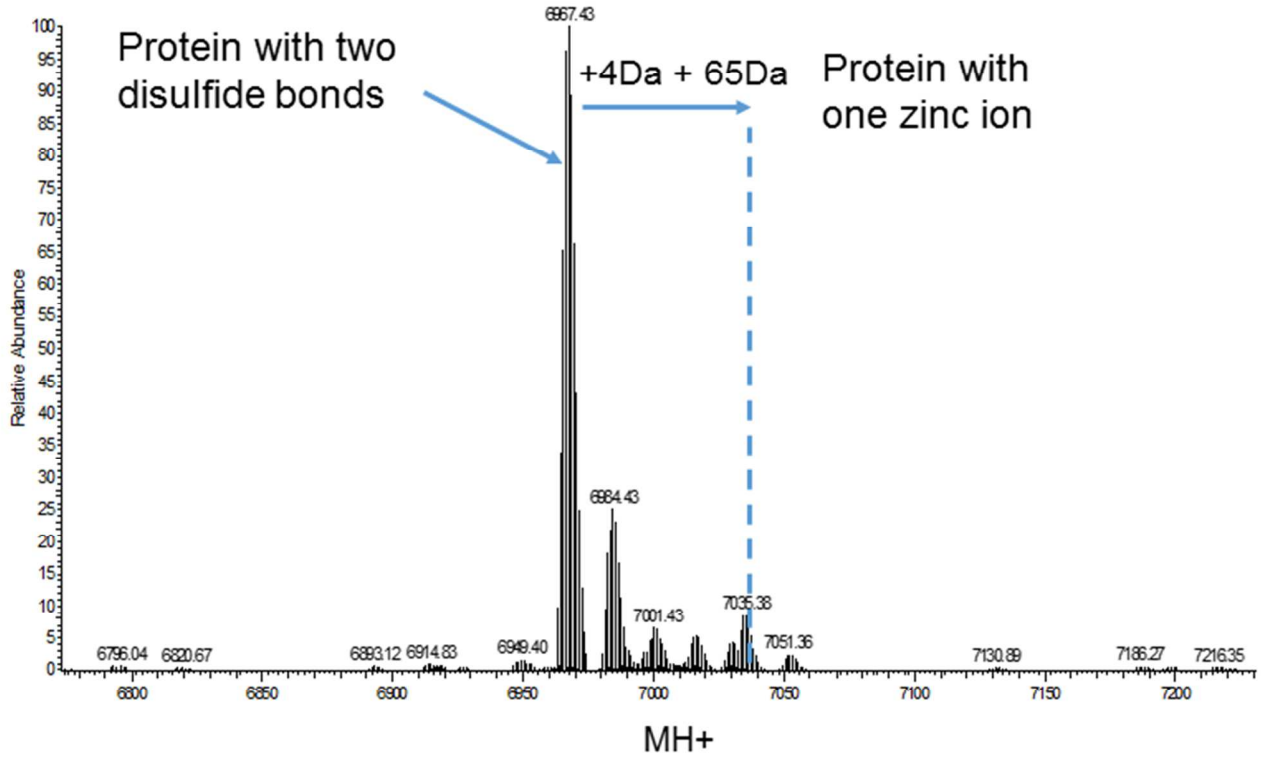

Figure S5. The deconvoluted spectrum from the averaged mass spectrum across the peak of the identified 50S ribosomal protein L31 proteoform without the zinc cofactor. The Xtract software from Thermo Fisher Scientific was used for the mass deconvolution with the default settings.

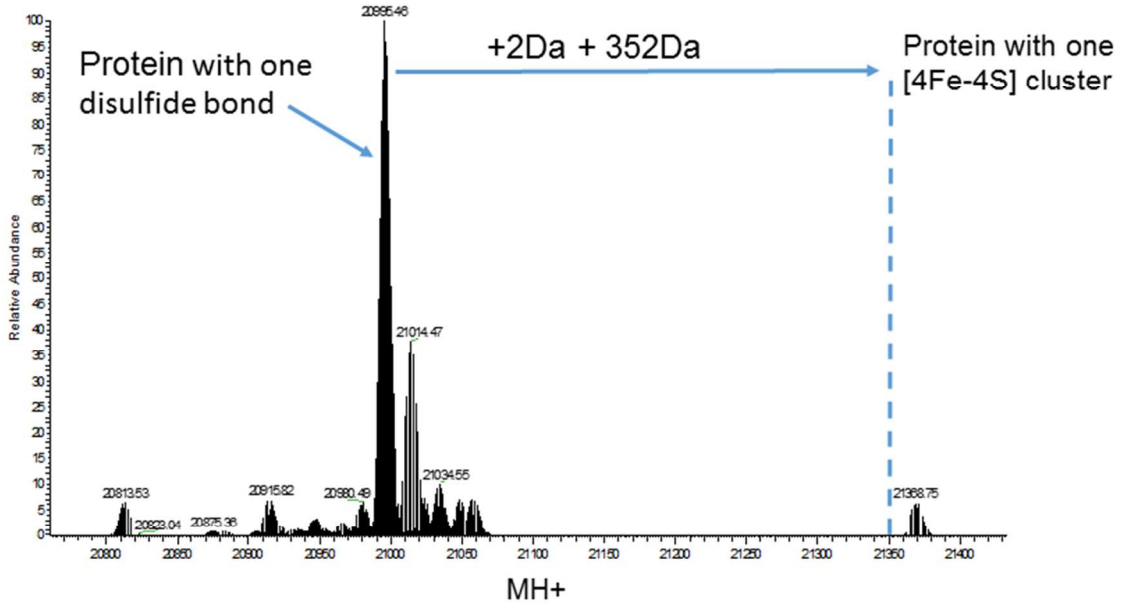

Figure S6. The deconvoluted spectrum from the averaged mass spectrum across the peak of the identified Fe/S biogenesis protein NfuA proteoform without the [4Fe-4S] cofactor. The Xtract software from Thermo Fisher Scientific was used for the mass deconvolution with the default settings.

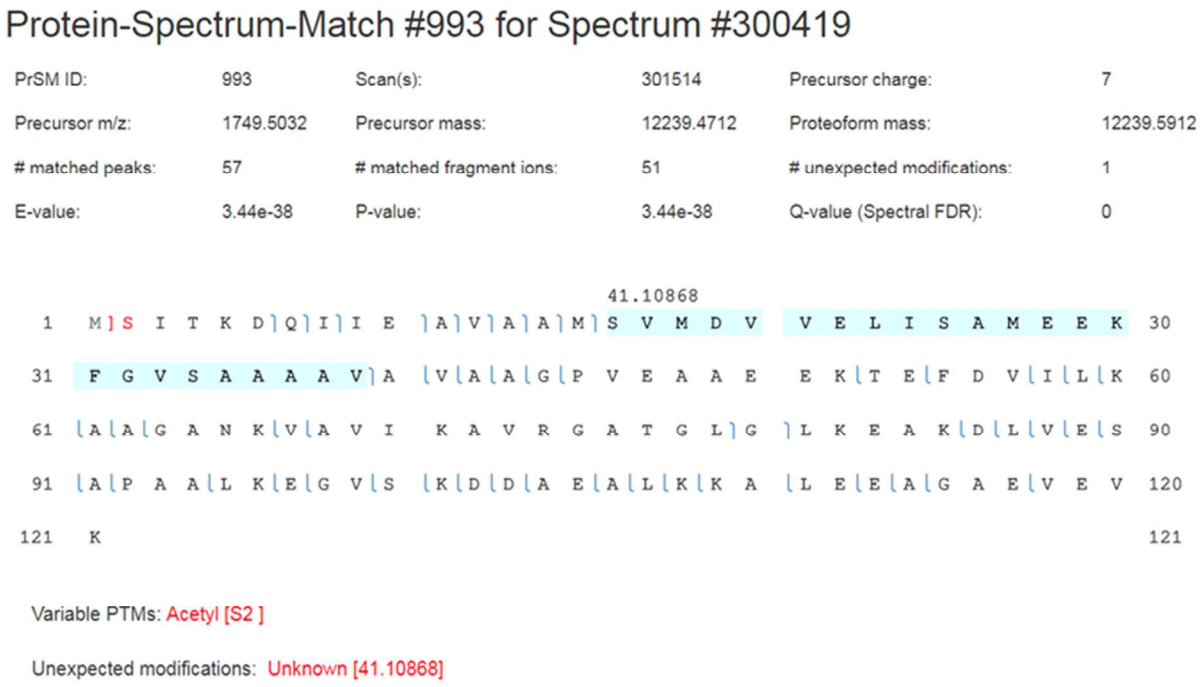

Figure S7. The sequence of the 50S ribosomal protein L7/L12, the observed fragmentation pattern, and the modifications through the database search. The initial methionine excision, N-terminal acetylation, and one +41 Da modification were labelled.

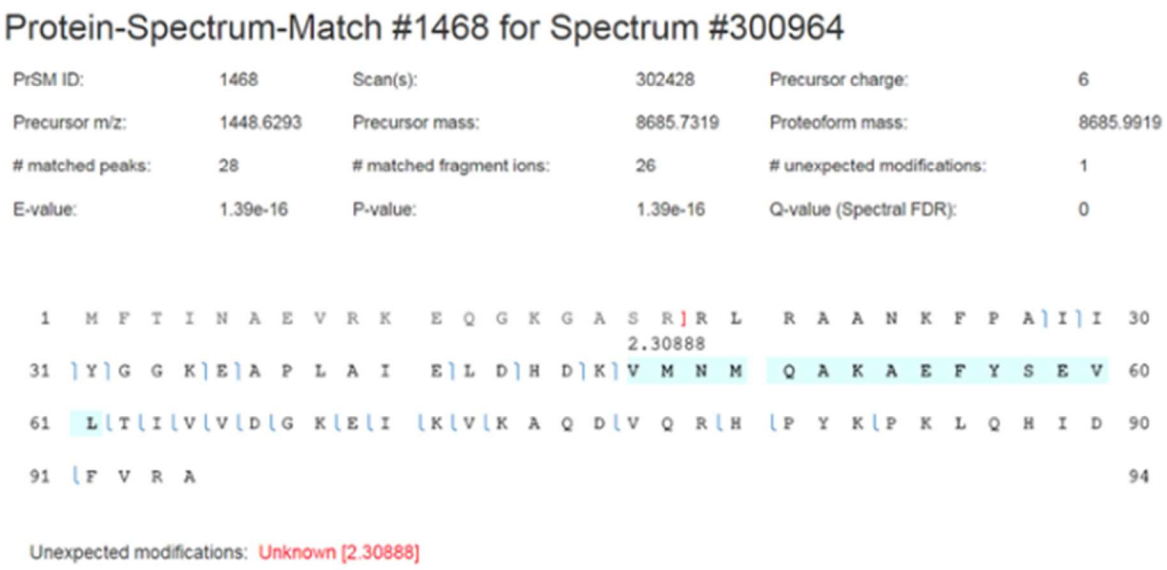

Figure S8. The sequence of the 50S ribosomal protein L25, the observed fragmentation pattern, and the modifications through the database search. The first 18 amino acids are cleaved as a signal peptide, which has not been reported in the literature. The signal peptide cleavage and one +2.3 Da modification were labelled.

References

(1) Zhu, G.; Sun, L.; Dovichi, N.J. *Talanta* **2016**, *146*, 839-843.
(2) Chen, D.; Shen, X.; Sun, L. *Analyst* **2017**, *142*, 2118-2127.
(3) Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J. *Angew. Chem. Int. Ed.* **2013**, *52*, 13661-13664.
(4) Skinner, O. S.; Haverland, N. A.; Fornelli, L.; Melani, R. D.; Do Vale, L. H.; Seckler, H. S.; Doubleday, P. F.; Schachner, L. F.; Srzentić, K.; Kelleher, N. L.; Compton, P. D. *Nat. Chem. Biol.* **2018**, *14*, 36-41.
(5) Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2554-2560.
(6) Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J. *J. Proteome Res.* **2015**, *14*, 2312-2321.
(7) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534– 2536.
(8) Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. *Mol. Cell. Proteomics* **2010**, *9*, 2772– 2782.
(9) Kou, Q.; Xun, L.; Liu, X. *Bioinformatics* **2016**, *32*, 3495-3497.
(10) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207–214.
(11) McCool, E. N.; Lubeckyj, R. A.; Shen, X; Chen, D; Kou, Q; Liu, X; Sun, L. *Anal Chem.* **2018**, *90*, 5529-5533.