Top-Down Proteomics Reveals Novel Protein Forms Expressed in *Methanosarcina acetivorans*

Jonathan T. Ferguson,[a] Craig D. Wenger,[a] William W. Metcalf,[b] and Neil L. Kelleher[a]

[a] Department of Chemistry at University of Illinois at Urbana-Champaign, Urbana, Illinois, USA
[b] Department of Microbiology at University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

Using both automated nanospray and online liquid chromatography mass spectrometry LC-MS strategies, 99 proteins have been newly identified by top-down tandem mass spectrometry (MS/MS) in *Methanosarcina acetivorans*, the methanogen with the largest known genome [5.7 mega base pairs (Mb)] for an Archaeon. Because top-down MS/MS was used, 15 proteins were detected with mispredicted start sites along with an additional five from small open reading frames (SORFs). Beyond characterization of these more common discrepancies in genome annotation, one SORF resulted from a rare start codon (AUA) as the initiation site for translation of this protein. Also, a methylation on a 30S ribosomal protein (MA1259) was localized to Pro59–Val69, contrasting sharply from its homologue in *Escherichia coli* (rp S12) known to harbor an unusual β-thiomethylated aspartic acid residue. (J Am Soc Mass Spectrom 2009, 20, 1743–1750) © 2009 Published by Elsevier Inc. on behalf of American Society for Mass Spectrometry

Comprising one of the three domains of life, the Archaea utilize unique metabolic pathways to survive in seemingly inhospitable environments such as high salt or extremes of temperature and pH [1, 2]. Archaea share many similarities to both bacteria and eukaryotes, but are a separate form of life based on now classic studies of 16S rRNA [3]. The Archaea are an important component of the biosphere, in part due to the production of methane as a byproduct that contributes to the global carbon cycle and has been used in commercial energy-producing applications [2, 4, 5]. *Methanosarcina acetivorans* is an especially diverse methanogen in its ability to use several substrates (mono-, di-, trimethyl amines, methanol, carbon monoxide, and acetate) for energy with methane generated as the byproduct. As long as *M. acetivorans* is supplied with a suitable anaerobic environment, the organism can generate all necessary amino acids solely from its food source. While not the only methanogenic archaeon, it is arguably one of the most metabolically diverse, with a large genome of 5.7 Mb harboring three distinct methanogenesis pathways [5]. Furthermore, *M. acetivorans* also contain pyrrolysine in methyltransferases [6, 7]. This 22nd amino acid was discovered first in a similar methanogen, *Methanosarcina barkeri* [8].

Several approaches have been undertaken to characterize the *M. acetivorans* proteome. A bottom-up two-dimensional (2D) gel approach by Ferry and coworkers resulted in 422 identified proteins from 968 spots [9]. They have furthered their analysis with N-14/N-15 metabolic labeling and microarray data comparing protein expression in acetate versus methanol grown cells [10, 11]. In a previous paper published by our lab [12], we reported 101 proteins identified using a semiautomated offline approach on a 8.5 Tesla (T) quadrupole-Fourier transform (Q-FT) mass spectrometer. Several groups have now reported detecting up to 1000 *intact* proteins from microbial systems, with none of these proteins identified by direct tandem mass spectrometry (MS/MS) of whole proteins [13–15]. The Hunt lab [16] and others [17, 18] have shown MS/MS of ribosomal proteins on a chromatographic time scale using ion trap and electron-based MS/MS methodology. Continuing the advancement of top-down technology, we report the use of two distinct platforms [19], producing 99 unique protein identifications. None of the 99 proteins reported here overlap with the 101 previously reported by our laboratory, and 29 of the 99 overlap with the 2D-gel bottom-up study by Ferry et al. [9]. Included in the identified protein list are 15 mispredicted start sites, five previously unannotated proteins, a few co- or post-translational modifications (PTMs), and the use of a very rare AUA codon for a start methionine (Met) in *lieu* of the standard start AUG.

Experimental

Sample Preparation

M. acetivorans was grown in HS medium with 125 mM methanol and 40 mM acetate in 1 L culture under anaerobic conditions [20]. Cells were harvested at 800 × *g*

Address reprint requests to Dr. N. L. Kelleher, University of Illinois at Urbana-Champaign, 600 S. Mathews, Urbana, IL 61801, USA. E-mail: kelleher@scs.uiuc.edu

1044-0305/09/$32.00
doi:10.1016/j.jasms.2009.05.014

Published online June 6, 2009
Received December 10, 2008
Revised April 29, 2009
Accepted May 23, 2009

for 10 min and stored at −80 °C until needed. Wet cells (0.25 g) were removed from the cell pellet and suspended in 1.5 mL lysis buffer (50 mM Tris buffer at pH 8.0 with inhibitor cocktail, P8340; from Sigma, St. Louis, MO, USA) and sonicated for 2 min at 15 s on/15 s off intervals. The first dimension of separation relied either on chromatofocusing of 10 to 15 mg of protein (PF2D; Beckman Coulter, Fullerton, CA, USA) or an acid precipitation with 0.4 N H_2SO_4. The former strategy separates proteins on the basis of their isoelectric point (pI) from a pH of 8.5 to 4.0. As a straightforward procedure for reducing complexity of the proteome by limiting the number of types of proteins introduced to a HPLC column, acid precipitations were performed. For acid precipitations, 1.5 mL of 0.4 N H_2SO_4 was added to the lysis buffer, and the lysate was allowed to sit on ice for 30 min. The sample was spun at 2500 × *g* for 5 min. The supernatant was transferred to a 1.5 mL tube and spun again at 14,000 × *g*. Twenty μL of this sample was loaded to a 75 μm column or 175 μL was loaded onto a 1 mm column (see below) while the remaining solution was treated with perchloric acid to further precipitate proteins and simplify chromatography. For perchloric acid precipitation, 70% perchloric acid was added to the sulfuric acid supernatant to reach a final concentration of 5% perchloric acid. The sample was spun at 14,000 × *g* for an additional 5 min. The supernatant was collected for MS analysis. The precipitate was redissolved in 6 M urea and saved for MS analysis. The combination of these separation strategies, along with the following LC-MS approaches, allowed for a larger variety of detected proteins, and complements previous work performed with gel electrophoresis [12].

Offline LC-MS

Offline separation was performed on a 4.6 × 150 mm C4 reversed-phase liquid chromatography column. Buffer A consisted of 99.9% water + 0.1% trifluoroacetic acid and Buffer B was 99.92% acetonitrile + 0.08% trifluoroacetic acid. The gradient consisted of a 5 min wash with 5% Buffer B, followed by a 5 min ramp to 15% B, and then a linear gradient to 50% B for 50 min, and a 10 min ramp to 95% B with a 5 min wash. The flow rate was set to 1 mL/min, and 1 min fractions were collected. The collected samples were lyophilized to dryness and reconstituted in 10 to 20 μL 49.5:49.5:1 water:methanol: formic acid. The samples were introduced into a 8.5 T custom built Q-FT mass spectrometer, described previously in high detail [12, 21]. Proteins were fragmented with collision-induced dissociation (CID) unless otherwise specified.

LC-MS with Offline Automation

A full paper describing this data acquisition platform has recently been published [19]. The prepared samples were run on a polymer column PLRP-S 1000 Å, 5 μm 1.0 × 150 mm (Higgins Analytical, Mountain View, CA, USA) at 100 μL/min into a TriVersa NanoMate (Advion BioSciences, Ithaca, NY, USA). A split flow allowed 300 nL/min into the mass spectrometer with the remaining eluent collected in a 96-well plate for later analysis. A list of targets of signal-to-noise ratio (S/N) 25:1 or greater were acquired from the LC-MS run. In the LC-MS experiment, no fragmentation was performed. The raw file produced was processed using in-house software [19]. From the list of protein targets, the Advion TriVersa automatically acquired sample from the 96-well plate and interrogated each protein target in direct infusion mode. Upon detecting a normalized signal intensity of 500 or greater in an isolation preview scan, one MS^1 scan, five isolation scans, and either 25, 50, or 100 fragmentation scans were taken depending on signal intensity of the precursor [19].

Nanocapillary-LC-MS

For nanocapillary-LC-MS, 75 μm columns of ~8–12 cm length were fritted with lichrosorb (EM Separations, Gibbstown, NJ, USA) and packed with 10 μm C4 packing material (Y.M.C. Separation Technology, Kyoto, Japan). Twenty micrograms of total protein was loaded onto each column. Each column was connected to an Eksigent 1D+ (Eksigent, Dublin, CA, USA), which used an Advion TriVersa as the electrospray source. Samples were introduced into a 12 T LTQ FT Ultra (Thermo Fisher Scientific, San Jose, CA, USA) at 300 nL/min. Gradients were the same as previously published [19]. Data acquisition for these experiments consisted of "zoom mapping," which used the ion trap to isolate *m/z* segments of the spectrum for FT detection. Spectra were "mapped" in 60 *m/z* windows from 750 to 1100 *m/z* in 50 *m/z* steps to allow for a 10 *m/z* overlap between windows; this ensured no targets would fall at a window boundary. Furthermore, the maximum allowable injection time for each isolation was 8 s, although the automatic gain control (AGC) setting of 10^6 typically limited this time to 1–4 s depending on the protein's abundance. Since AGC counts charges of all ions in the ion trap, performing an isolation scan limited the vast majority of ions to the protein of interest, thereby greatly improving signal-to-noise. AGC also reduces space charge effects, thereby improving mass accuracy. Previously we reported a mass accuracy of 3.5 ppm for MS^2 data [19]. After each isolation FT scan, a data-dependent scan was performed to fragment the most abundant target in that particular 60 *m/z* window. If a target was fragmented twice in a 2 min period, it was placed on an exclusion list for 4 min.

Database Searching

Two different search modes were implemented using the ProSightHT tool within ProSightPC 2.0 (Thermo Fisher Scientific, San Jose, CA, USA). These automated searches were run iteratively to interrogate a *M. acetivorans* database containing 21,658 protein forms created from 4470 basic sequences by considering all N-terminal acetylation and Met on/off combinations. The database was created using the Database Manager tool within ProSightPC 2.0. For all searches, the fragment tolerance was set to ±10 ppm. The first search ran in absolute mass mode set at a ±10 Da precursor tolerance with multiplexing enabled. This quick search allowed for the identification of ~70% proteins reported here. The second search in the iterative search "tree" ran in the same mode but with a tolerance of ±5000 Da with Δm mode and multiplexing enabled; when used iteratively, ProSight performs the second search automatically only if the preceding search fails (i.e., returns no hit or a hit with an expectation value greater than 10^{-4}). The final search was run at a ±1.1 Da precursor tolerance in "biomarker" mode. This mode searches a protein's sequence for subsequences with theoretical monoisotopic masses that match to the observed mass within the specified tolerance. This search considers all protein sequences in the *M. acetivorans* top-down database. If a subsequence matched within the tolerance, the observed fragmentation masses were compared to predicted b- and y-ions for that candidate. To find the unannotated small open reading frames (SORFs), a fourth search was performed in a directed fashion on a database constructed from a simple six-frame translation of the *M. acetivorans* genome. These six-frame translations were annotated in the database as six continuous "megaproteins" and were interrogated using the biomarker search mode described above. Given the larger number of candidate sequences for this search, the time per search was ~5 min and benefited from high-quality fragmentation data to obtain e-values below 10^{-4}, which was considered an identification without the need for further analysis. Identifications with e-values greater than 10^{-4} were manually validated and considered identifications if the intact mass uniquely matched the suspected protein within ±10 ppm, or if the lack of an intact mass was due to poor signal intensity of large proteins (>25 kDa) or sample complexity due to artifactual modifications.

Glu-C Digestion of MA1259

Glu-C (Endoproteinase V8) sequencing grade from *Staphylococcus aureus* was acquired from Roche (Penzberg, Germany) and digestion was performed according to manufacturer's instructions in 25 mM ammonium carbonate buffer, pH 7.8 for 18 h at 25 °C.

DNA Extraction, PCR, and DNA-Sequencing

DNA extraction was performed using Qiagen's DNeasy Tissue Handbook 03/2004: Purification of Total DNA from Cultured Animal Cells. Primers were designed using OligoPerfect Designer (http://www.invitrogen.com/). Two μL of template DNA was added to the following recipe: 10 μL of HF buffer (Finnzymes), 36 μL of 0.2 μm filtered water, 1 μL of 10 mM dNTP mix, 0.3 μL forward primer (100 μM), 0.3 μL reverse primer (100 μM), and 2 μL Phusion High-Fidelity DNA polymerase (2 U/μL). Polymerase chain reaction (PCR) of MA1259 ribosomal protein S12P used the following primers to produce a 539 bp product: Forward: 5′-TTACCTCTACTTCCGTCGGGT-3′. Reverse: 5′-TTAATGCCACAATAGGAAACTGTG-3′. PCR of 8552.18 Da unannotated protein: Forward: 5′-TCGTTTGCTTCCTCAACTTTG-3′. Reverse: 5′-CGGAGAAGGCGAGGTCTG-3′. The PCR program for MA1259 occurred on a Px2 thermal cycler (Thermo Electron Corp., Waltham, MA, USA) as follows: 98 °C for 3 min followed by 98 °C for 30 s, 63.9 °C for 30 s, 72 °C for 15 s (35 cycles). The program concluded with: 72 °C for 5 min and 4 °C hold temperature until manual termination. The PCR program for targeting the unannotated protein's DNA sequence was equivalent to the above except an annealing temperature of 64.4 °C was substituted for the 63.9 °C. PCR products were agarose gel purified, extracted, and subjected to a second round of PCR. The gel extraction was performed with QIAquick's gel extraction kit, and the PCR purification used QIAquick's PCR purification kit. Capillary sequencing was performed with an ABI 3730XL capillary sequencer (Applied Biosystems, Foster City, CA, USA) at the W. M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana-Champaign.

Results and Discussion

Given our recent development of improved data acquisition using offline automation [19], we employed this platform along with nanocapillary-LC-MS/MS [17] to raise the number of intact proteins detected and identified versus those already found by top-down MS [12]. In one offline automation run, 147 protein targets were selected for analysis, leading to 42 successful MS/MS experiments representing 18 new and unique protein identifications. Another run resulted in an additional 24 unique identifications while avoiding most of the signals arising from past identifications [21]. Ultimately, a total of 99 new and unique identifications were obtained in this study (including some manual and LC-MS/MS experimentation, described below), as compared to the 101 intact proteins previously identified for this organism [12].

Some proteins were only detected by LC-MS/MS, which allowed for faster identification (just 2–4 FTMS/MS scans are possible on a chromatographic time scale). Using a 1 mm i.d. column, one LC-MS run of proteins isolated via sulfuric acid precipitation provided 90 protein forms (i.e., targets) after manual validation of intact mass values, with 29 of these actually identified by data-dependent MS/MS and ProSightHT. Five of these 29 were only detected via the LC-MS approach (all LC-MS identifications are listed in Supplemental Table 1). With the remaining samples from the above run, a secondary precipitation was performed using perchloric acid. The dissolved precipitate was loaded onto a 75 μm C4 column, and this injection produced 30 identifications based on MS^2 data. Thirty μL of the supernatant were loaded onto an equivalent column, which resulted in six additional identifications, including a putative mercury-ion binding protein with a C-X-X-C motif that was detected with a disulfide bond intact, experimentally determined as, $\Delta m = -2.02$ Da (Figure 1c). For top-down to become more widespread, proteome coverage needs to improve, and the experiments outlined above indicate the potential of using project-wide exclusion lists to direct more instrument time toward lower abundance species. The C-X-X-C motif has been shown to play a critical importance in thioredoxin in *Bacillus subtilis* [22]. Therefore, we wished to identify and fully characterize thioredoxin in *M. acetivorans* and used a slower but data-rich offline electron capture dissociation (ECD) methodology. In Figure 1d, ECD unambiguously identified thioredoxin MA3212, but with a negative mass shift localized near the N-terminus. The removal of the first four amino acids near the N-terminus resulted in the matching of 10 c-type ions. Therefore, this protein must begin at the AUG start codon immediately downstream of that predicted during automated genome annotation. Previously, three mispredicted start sites were reported for the methanogen, *Methanococcus jannaschii* [23], and Patrie et al. [12] reported four mispredicted start sites in *M. acetivorans*. In this study, we added an additional 11 mispredicted start sites for *M. acetivorans*, correcting 7.5% of the 200 total proteins identified by top-down MS (Table 1). In all cases, the mispredicted start site caused a negative mass discrepancy between the observed and the theoretical mass. During gene annotation, the methionine furthest upstream was automatically annotated as the start methionine. In such cases, a top-down dataset contained matching fragment ions that incorporated the N-terminus (i.e., b- or c-type ions) and contained a common mass shift. Algorithms for automated genome annotation in the Archaea could base future predictions on the cases detected here to more accurately predict translational start sites.

During data acquisition using automated nanospray, MA0059 was identified with both methionine-on and methionine-off at an ~1:1 ratio. A separate analysis using manual ECD instead of automated CID unambiguously localized a mispredicted start site (data not shown). Therefore, in rare cases, *M. acetivorans* was not precise in its removal of the start methionine. While the above data were acquired by selecting individual protein forms, multiple protein identifications were routinely observed using automated modes for both data acquisition [24] and data processing via ProSightHT. In a LC-MS/MS run, MA0059 and MA0056 were identified as the sole protein in their respective ion trap isolation window, Figure 2b and e. However, two adjacent proteins on the chromosome, MA0057 and MA0058, were identified in parallel using the multiplexing option in the ProSight search algorithm, Figure 2c and d. A careful examination of the spectrum, however, revealed three protein forms, with one of these 16.3 Da heavier than MA0057, too large for an oxidation (16.0 Da). Manual analysis revealed this pro-

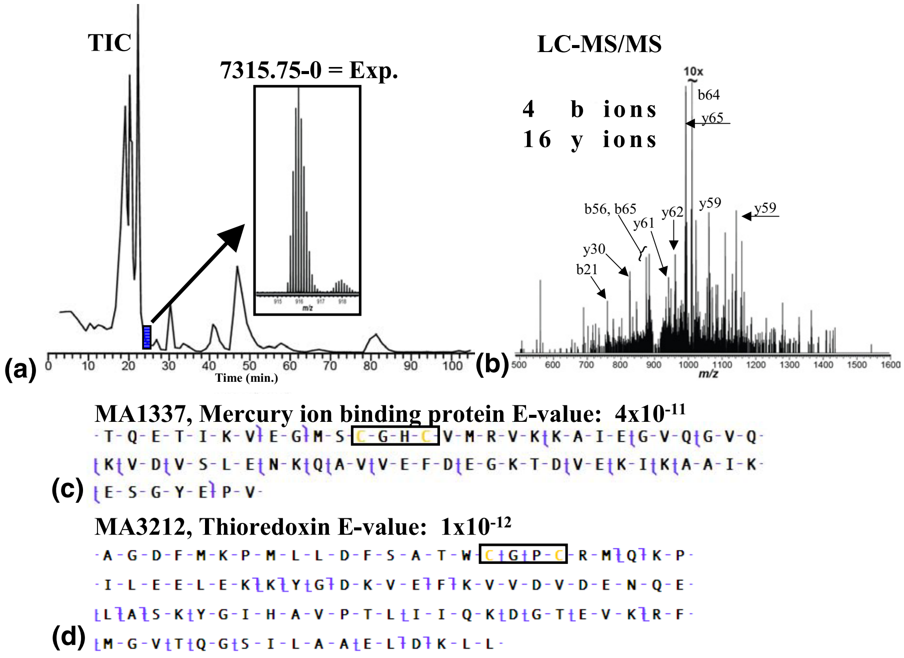

Figure 1. Mercury ion binding protein exhibiting a C-X-X-C motif was detected eluting at ~25 min (**a**). LC-MS/MS identified the protein (**b**) with an e-value of 4×10^{-11} (**c**). A disulfide bond was observed producing a −2 Da mass discrepancy. Thioredoxin has a C-X-X-C motif likely for facilitating reduction. Furthermore, this protein has a mispredicted start site (**d**), where the first four amino acids have been incorrectly annotated. The most abundant fragment ions have been labeled in (**b**).

Table 1. List of mispredicted proteins with predicted and experimentally determined start codons

Protein open reading frame identifier	Nucleotide sequence*
MA0056	ttcataagctggttaataaagtcaaaggaatgaacctg**atg**
MA0057	atttttaactaaagtaaaattaaggaatgattttta**atg**
MA0058	aattaattcaagtaaatttaaggaatggttttc**atg**
MA0059	attctactaaattaaaggaatgatttta**atg**
MA0596	ggatattccgctgaggtgttgaag**atg**
MA1083	attattgtctgaggtgattgca**atg**
MA1089	ctacatagtgcagaaggcctgaaggtgattatg**atg**
MA1108	cgacaggggctacaggatgctgtatgcactccaatggaggaagtat**atg**
MA1522	cttaaataatcctgataacatggggtgctgatt**atg**
MA1964	aggtgaaattccgattaaactgaaaacttaaaattgaggcttacaaag**atg**
MA3212	gacgtttcaagtattgcaggtgatttt**atg**
MA3896	ccagaggagatgaaaacg**atg**
MA4024	ccgaatgaaatcaaagttttaatagtttagttgaaaaagataatcgtttggttgaaaaggatagtgaaagtt**atg**
MA4111	ctgttccagaactaagttacagcaagagagtggtaatg**atg**
MA4116	ccagaaaaccggtttttcggtgattgcc**atg**

*Observed start codon is shown in bold. Theoretical start codon is underlined.

tein to be the Met-on form of MA0058. Furthermore, MA0058 and MA0057, along with MA0056 and MA0059, all had mispredicted start sites. The genes for these proteins lie next to each other on the chromosome in an apparent operon, and are all quite abundant, as shown in Figure 2a. Therefore, our LC-MS/MS high-resolution data acquisition strategy allowed for unambiguous identification of multiple mispredicted proteins (some in a multiplexed fashion), exhibiting variable amounts of N-terminal processing.

As seen in Figure 3a, 25 scans of information-rich CID data were acquired but failed all search modes in ProSightPC 2.0. However, searching through a six-frame translation of the *M. acetivorans* genome in biomarker mode identified an 8.5 kDa protein with an especially good expectation value of 10^{-104}. Examination of the gene encoding this protein via PCR confirmed the theoretical nucleotide sequence, suggesting that a predicted isoleucine (AUA) was in fact translated as the start methionine for this gene. A ribosome-binding sequence was detected 8 bp upstream. Though the AUG is the preferred start codon, GUG, UUG, and AUU are used sometimes in a variety of archaea and bacteria [25–29]. A basic local slignment search tool (BLAST) search reveals this "hypothetical" protein was actually annotated in the methanogens *Methanosarcina*

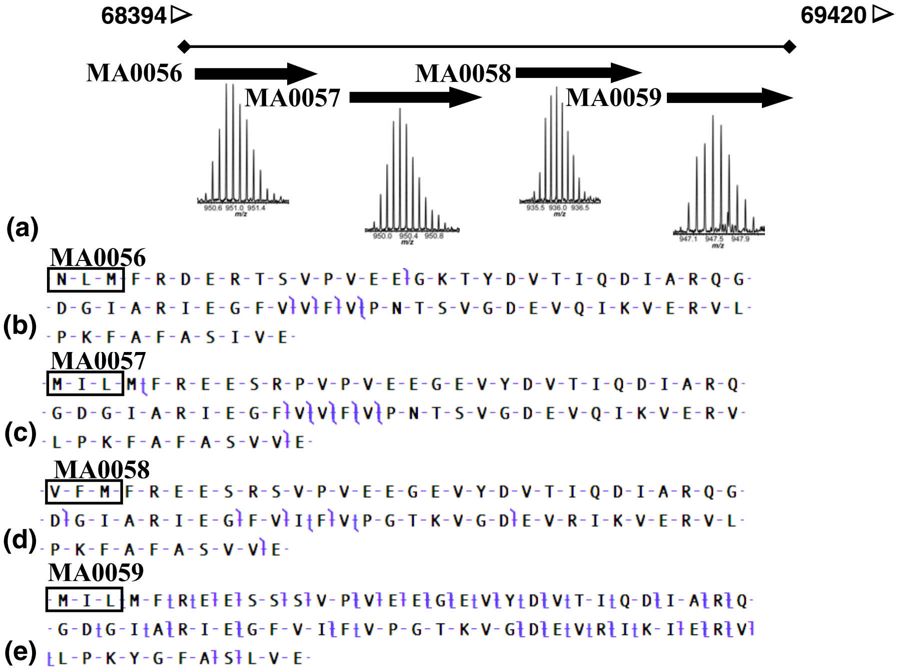
68394
69420
MA0056
MA0057
MA0058
MA0059
950.6 951.0 951.4
m/z
950.0 950.4 950.8
m/z
935.5 936.0 936.5
m/z
947.1 947.5 947.9
m/z
(a)
MA0056
N-L-M F-R-D-E-R-T-S-V-P-V-E-E G-K-T-Y-D-V-T-I-Q-D-I-A-R-Q-G-
-D-G-I-A-R-I-E-G-F-V V F V P-N-T-S-V-G-D-E-V-Q-I-K-V-E-R-V-L-
-P-K-F-A-F-A-S-I-V-E-
(b)
MA0057
-M-I-L-M F-R-E-E-S-R-P-V-P-V-E-E-G-E-V-Y-D-V-T-I-Q-D-I-A-R-Q-
-G-D-G-I-A-R-I-E-G-F V V F V P-N-T-S-V-G-D-E-V-Q-I-K-V-E-R-V-
-L-P-K-F-A-F-A-S-V-V E-
(c)
MA0058
-V-F-M-F-R-E-E-S-R-S-V-P-V-E-E-G-E-V-Y-D-V-T-I-Q-D-I-A-R-Q-G-
-D G-I-A-R-I-E-G F-V I F V P-G-T-K-V-G-D E-V-R-I-K-V-E-R-V-L-
-P-K-F-A-F-A-S-V-V E-
(d)
MA0059
-M-I-L M-F R E E S-S S V-P V E E G E V Y D V T-I Q-D I-A R Q-
-G-D G-I A R-I-E G-F-V-I F V-P-G-T-K-V-G D E V R I K-I E R V
L-P-K-Y-G-F-A S L-V-E-
(e)

Figure 2. Several proteins were detected whose genes are present on the chromosome in a putative operon in an LC-MS run (**a**). Using CID to produce b- and y-ions, each protein was identified with a mispredicted start site, (**b**)–(**e**).

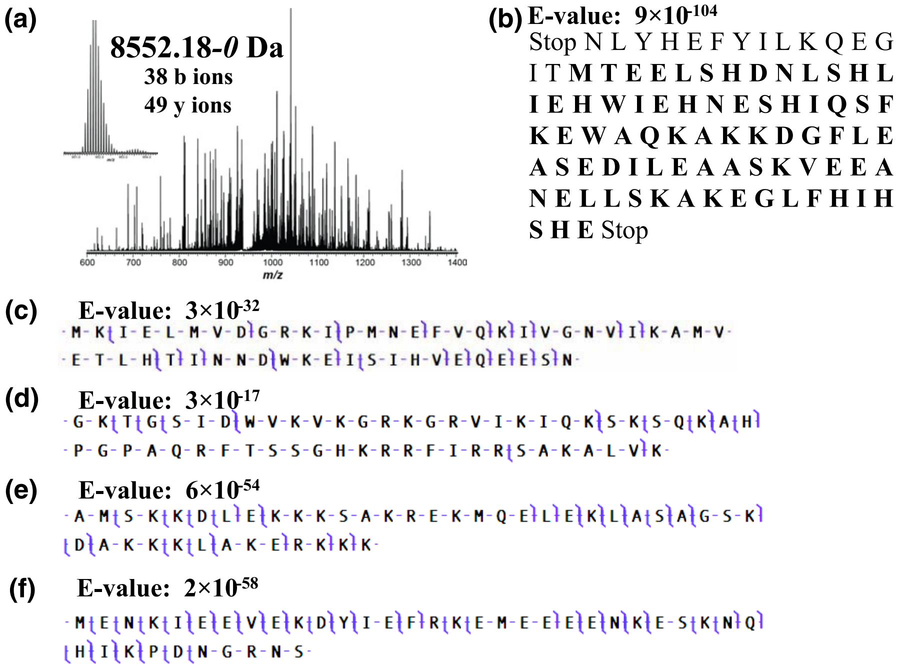

Figure 3. A protein was identified despite not being annotated in the *M. acetivorans* database. In (**a**), 25 CID fragmentation scans of the LC-MS target were acquired *via* automated nanospray. A biomarker search of a database created from simple six-frame translations revealed a single hit, (**b**). DNA sequencing revealed that a Met (AUA) was the start codon for translation initiation. Four more unannotated proteins were identified using a biomarker search of a six-frame translated database, (**c**)–(**f**).

barkeri and *Methanosarcina mazei* but with utilization of normally-encoded start methionines in both cases (data not shown). In *M. barkeri*, the predicted start site is UUG, although the likely start codon is AUU based off its proximity to its ribosome-binding sequence. Furthermore, the *M. mazei* protein likely has an AUA-encoded start site, which has been incorrectly annotated in the database.

Four more such cases of SORFs have been identified by searching the database created from a raw six-frame translation of genomic DNA (Figure 3c–f). There is no sequencing error or alternative initiation codon involved in these additional four proteins. When annotating a genome, it is common practice to ignore small open reading frames with less than 100 codons unless there is high sequence homology to other *bona fide* proteins [5]. The four proteins in Figure 3c–f are less than 100 residues, but two of the proteins (Figure 3c and d) are homologous to proteins in the related methanogens, *M. mazei* and *M. barkeri*. However, at time of publication of *M. acetivorans*'s sequence, the *M. mazei and M. barkeri* genome sequences were not available and, therefore, not used for comparative purposes [5]. The final two unannotated proteins have intact masses <5000 Da, making them the smallest unannotated proteins detected, Figure 3e and f. A BLAST search revealed little similarity in archaea, bacteria, or eukaryotes (data not shown). The closest BLAST hit for the protein in Figure 3e (e-value 3×10^{-4}) was to a putative uncharacterized protein in *Methanoregula boonei* (strain 6A8; accession number A7I9X5). A translation of the DNA sequence for both proteins, Figures 3e and f, suggested they have an AUG start codon and a UAA stop codon for translation termination. Therefore, it was likely that the detected species were full-length proteins and not fragments or proteolysis products. In an alternative strategy, the Yates lab used a 2D gel approach to enrich for SORFs and identified 22 using 2D gels and bottom-up MS [30]. A mass error list for the protein in Figure 3a has been included in Supplemental Table 1, which can be found in the electronic version of this article.

Of the 59 predicted ribosomal proteins, 43 were identified, including MA1259 that harbored a +14 Da mass discrepancy (Figure 4). In *Escherichia coli*, ribosomal protein S12 is the homologue to MA1259 and harbors a unique β-methyl-thioaspartic acid modification on aspartic acid 88 [31]. A sequence alignment positioned the analogous aspartic acid to position 112 on the *M. acetivorans* homologue. However, ECD data on an 8.5 T Q-FT mass spectrometer clearly showed this was not the site of the +14 Da modification (Figure 4a), and localized it to the Pro59–Asp85 region from these data. Therefore, the fraction containing MA1259 was subjected to a Glu-C digestion to produce a peptide containing the site of modification, Figure 4b. The desired peptide was detected in the sample at low intensity and MS/MS of it only narrowed the site of the

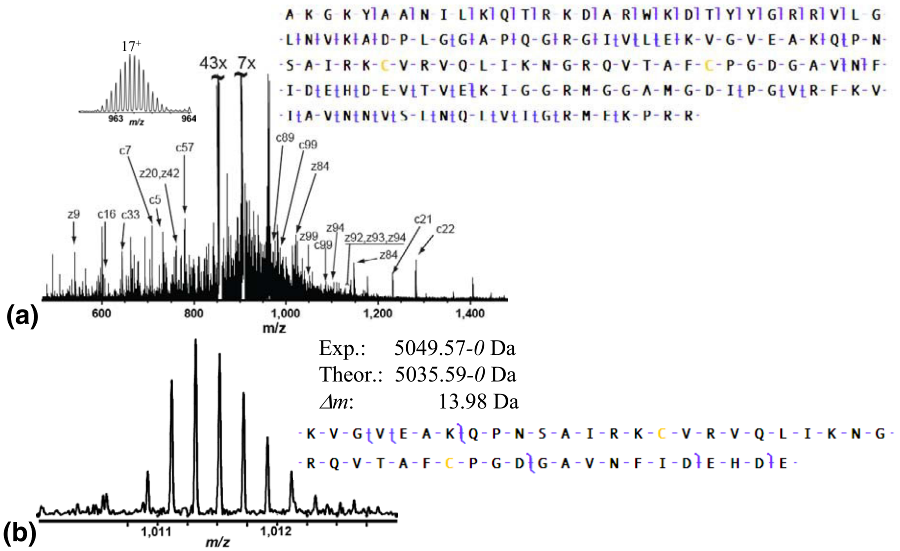

Figure 4. Intact protein MA1259 ribosomal protein S12P exhibits a +13.98 Da mass shift. Presumably, the uncalibrated data from the Q-FT mass spectrometer, which lacks automatic gain control, produced a −0.04 Da (−7.92 ppm) discrepancy from the theoretical +14.02 Da methylation. The protein was fragmented with ECD to localize the modification to a stretch of 30 amino acids (**a**). Offline analysis of a Glu-C digest detected a peptide with two missed cleavages incorporating the +14.02 Da modification (**b**). Fragmentation data further narrowed down possible sites of modification to 28 amino acids.

+14.00 Da shift (consistent with methylation, +14.02 Da) from a 30- to 27-residue window. CID on a 12 T LTQ FT Ultra proved more fruitful, reducing the possible sites to Pro59–Val69. PCR from genomic DNA was performed followed by DNA-sequencing to confidently assign the +14 Da shift to a PTM and not a DNA-sequencing error.

Summary

Combining our previous work with current platforms presented here that identified 99 new proteins (Supplemental Table 1), a total of 200 protein identities in *M. acetivorans* have been confirmed. Of these, five proteins were unannotated, 15 were mispredicted, two exhibited variable removal of the start methionine, and one was a proteolysis fragment from the identical C-terminus of two possible proteins. The incorporation of LC-MS/MS and offline automation speeds up the processing of samples through automation of the top-down process. For five data files that were automatically processed and iteratively searched, ~2 h were required to complete analysis. In these, a total of 835 identifications were detected by the search algorithm in this time period. The expanded use of top-down will use both online and offline strategies, with increasingly sophisticated data acquisition strategies tailored for the challenges of top-down fragmentation [16, 18, 19, 32, 33].

Acknowledgments

The authors are grateful to Dr. Gary Olsen for fruitful discussions, and Jun Kai Zhang for culturing *M. acetivorans* cells. N.L.K. acknowledges the National Institutes of Health (NIH GM 067193-06) and the University of Illinois Institute for Genomic Biology. W.W.M. acknowledges support in part by NSF MCB Division of Cellular and Biosciences grant no. MCB12466 and by DOE Energy Biosciences Program grant no. DEFG02-02-ER15296.

Appendix A Supplementary Material

Supplementary material associated with this article may be found in the online version at doi:10.1016/j.jasms.2009.05.014.

References

1. Eichler, J. Facing extremes: Archaeal surface-layer (glyco)proteins. *Microbiology* **2003,** *149,* 3347–3351.
2. Schiraldi, C.; Giuliano, M.; De Rosa, M. Perspectives on biotechnological applications of archaea. *Archaea* **2002,** *1,* 75–86.
3. Fox, G. E.; Magrum, L. J.; Balch, W. E.; Wolfe, R. S.; Woese, C. R. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proc. Natl. Acad. Sci. U.S.A.* **1977,** *74,* 4537–4541.
4. Daniels, L. Biotechnological potential of methanogens. *Biochem. Soc. Symp.* **1992,** *58,* 181–193.
5. Galagan, J. E.; Nusbaum, C.; Roy, A.; Endrizzi, M. G.; Macdonald, P.; FitzHugh, W.; Calvo, S.; Engels, R.; Smirnov, S.; Atnoor, D.; Brown, A.; Allen, N.; Naylor, J.; Stange-Thomann, N.; DeArellano, K.; Johnson, R.; Linton, L.; McEwan, P.; McKernan, K.; Talamas, J.; Tirrell, A.; Ye, W.; Zimmer, A.; Barber, R. D.; Cann, I.; Graham, D. E.; Grahame, D. A.; Guss, A. M.; Hedderich, R.; Ingram-Smith, C.; Kuettner, H. C.; Krzycki, J. A.; Leigh, J. A.; Li, W.; Liu, J.; Mukhopadhyay, B.; Reeve, J. N.; Smith, K.; Springer, T. A.; Umayam, L. A.; White, O.; White, R. H.; Conway de Macario, E.; Ferry, J. G.; Jarrell, K. F.; Jing, H.; Macario, A. J.; Paulsen, I.; Pritchett, M.; Sowers, K. R.; Swanson, R. V.; Zinder, S. H.; Lander, E.; Metcalf, W. W.; Birren, B. The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Res.* **2002,** *12,* 532–542.
6. Mahapatra, A.; Srinivasan, G.; Richter, K. B.; Meyer, A.; Lienard, T.; Zhang, J. K.; Zhao, G.; Kang, P. T.; Chan, M.; Gottschalk, G.; Metcalf, W. W.; Krzycki, J. A. Class I and class II lysyl-tRNA synthetase mutants and the genetic encoding of pyrrolysine in *Methanosarcina* spp. *Mol. Microbiol.* **2007,** *64,* 1306–1318.
7. Mahapatra, A.; Patel, A.; Soares, J. A.; Larue, R. C.; Zhang, J. K.; Metcalf, W. W.; Krzycki, J. A. Characterization of a *Methanosarcina acetivorans*

mutant unable to translate UAG as pyrrolysine. *Mol. Microbiol*. **2006,** *59,* 56–66.

8. Hao, B.; Gong, W.; Ferguson, T. K.; James, C. M.; Krzycki, J. A.; Chan, M. K. A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science* **2002,** *296,* 1462–1466.
9. Li, Q.; Li, L.; Rejtar, T.; Karger, B. L.; Ferry, J. G. Proteome of Methanosarcina acetivorans Part I: An expanded view of the biology of the cell. *J. Proteome Res.* **2005**, *4,* 112–128.
10. Li, Q.; Li, L.; Rejtar, T.; Karger, B. L.; Ferry, J. G. Proteome of Methanosarcina acetivorans Part II: Comparison of protein levels in acetate- and methanol-grown cells. *J. Proteome Res*. **2005,** *4,* 129–135.
11. Li, L.; Li, Q.; Rohlin, L.; Kim, U.; Salmon, K.; Rejtar, T.; Gunsalus, R. P.; Karger, B. L.; Ferry, J. G. Quantitative proteomic and microarray analysis of the archaeon *Methanosarcina acetivorans* grown with acetate versus methanol. *J. Proteome Res*. **2007,** *6,* 759–771.
12. Patrie, S. M.; Ferguson, J. T.; Robinson, D. E.; Whipple, D.; Rother, M.; Metcalf, W. W.; Kelleher, N. L. Top down mass spectrometry of <60-kDa proteins from *Methanosarcina acetivorans* using quadrupole FTMS with automated octopole collisionally activated dissociation. *Mol. Cell Proteom*. **2006,** *5,* 14–25.
13. Smith, R. D. Advanced mass spectrometric methods for the rapid and quantitative characterization of proteomes. *Comp. Funct. Genomics* **2002,** *3,* 143–150.
14. Shen, Y.; Tolic, N.; Hixson, K. K.; Purvine, S. O.; Pasa-Tolic, L.; Qian, W. J.; Adkins, J. N.; Moore, R. J.; Smith, R. D. Proteome-wide identification of proteins and their modifications with decreased ambiguities and improved false discovery rates using unique sequence tags. *Anal. Chem.* **2008,** *80,* 1871–1882.
15. Krishnamurthy, T.; Hewel, J.; Bonzagni, N. J.; Dabbs, J.; Bull, R. L.; Yates, J. R. III. Simultaneous identification and verification of *Bacillus anthracis Rapid Commun*. *Mass Spectrom.* **2006,** *20,* 2053–2056.
16. Chi, A.; Bai, D. L.; Geer, L. Y.; Shabanowitz, J.; Hunt, D. F. Analysis of intact proteins on a chromatographic time scale by electron transfer dissociation tandem mass spectrometry. *Int. J. Mass Spectrom.* **2007,** *259,* 197–203.
17. Collier, T. S.; Hawkridge, A. M.; Georgianna, D. R.; Payne, G. A.; Muddiman, D. C. Top-down identification and quantification of stable isotope labeled proteins from *Aspergillus flavus* using online nano-flow reversed-phase liquid chromatography coupled to a LTQ-FTICR mass spectrometer. *Anal. Chem*. **2008,** *80,* 4994–5001.
18. Parks, B. A.; Jiang, L.; Thomas, P. M.; Wenger, C. D.; Roth, M. J.; Boyne, M. T. II; Burke, P. V.; Kwast, K. E.; Kelleher, N. L. Top-down proteomics on a chromatographic time scale using linear ion trap Fourier transform hybrid mass spectrometers. *Anal. Chem.* **2007,** *79,* 7984–7991.
19. Wenger, C. D.; Boyne, M. T. II; Ferguson, J. T.; Robinson, D. E.; Kelleher, N. L. A versatile online-offline engine for automated acquisition of high-resolution tandem mass spectra. *Anal. Chem.* **2008,** *80,* 8055–8063.
20. Metcalf, W. W.; Zhang, J. K.; Shi, X.; Wolfe, R. S. Molecular, genetic, and biochemical characterization of the serC gene of *Methanosarcina barkeri* Fusaro. *J. Bacteriol*. **1996,** *178,* 5797–5802.
21. Patrie, S. M.; Robinson, D. E.; Meng, F.; Du, Y.; Kelleher, N. L. Strategies for automating top-down protein analysis with Q-FTICR MS. *Int. J. Mass Spectrom.* **2004,** *234,* 175–184.
22. Lewin, A.; Crow, A.; Hodson, C. T.; Hederstedt, L.; Le Brun, N. E. Effects of substitutions in the CXXC active site motif of the extracytoplasmic thioredoxin ResA. *Biochem. J.* **2008,** *414,* 81–91.
23. Forbes, A. J.; Patrie, S. M.; Taylor, G. K.; Kim, Y. B.; Jiang, L.; Kelleher, N. L. Targeted analysis and discovery of posttranslational modifications in proteins from methanogenic archaea by top-down MS. *Proc. Natl. Acad. Sci. U.S.A.* **2004,** *101,* 2678–2683.
24. Meng, F.; Cargile, B. J.; Miller, L. M.; Forbes, A. J.; Johnson, J. R.; Kelleher, N. L. Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat. Biotechnol*. **2001,** *19,* 952–957.
25. Nolling, J.; Pihl, T. D.; Vriesema, A.; Reeve, J. N. Organization and growth phase-dependent transcription of methane genes in two regions of the *Methanobacterium thermoautotrophicum* genome. *J. Bacteriol*. **1995,** *177,* 2460–2468.
26. Polard, P.; Prere, M. F.; Chandler, M.; Fayet, O. Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS911. *J. Mol. Biol*. **1991,** *222,* 465–477.
27. Sacerdot, C.; Fayat, G.; Dessen, P.; Springer, M.; Plumbridge, J. A.; Grunberg-Manago, M.; Blanquet, S. Sequence of a 1.26-kb DNA fragment containing the structural gene for *E. coli* initiation factor IF3: Presence of an AUU initiator codon. *EMBO J*. **1982,** *1,* 311–315.
28. Sazuka, T.; Ohara, O. Sequence features surrounding the translation initiation sites assigned on the genome sequence of *Synechocystis* sp. strain PCC6803 by amino-terminal protein sequencing. *DNA Res*. **1996,** *3,* 225–232.
29. Wang, G.; Nie, L.; Tan, H. Cloning and characterization of sanO, a gene involved in nikkomycin biosynthesis in *Streptomyces ansochromogenes*. Lett. Appl. Microbiol. **2003,** *37,* 452–457.
30. Oshiro, G.; Wodicka, L. M.; Washburn, M. P.; Yates, J. R. III; Lockhart, D. J.; Winzeler, E. A. Parallel identification of new genes in *Saccharomyces cerevisiae*. *Genome Res*. **2002,** *12,* 1210–1220.
31. Kowalak, J. A.; Walsh, K. A. β-Methylthio-aspartic acid: Identification of a novel posttranslational modification in ribosomal protein S12 from *Escherichia coli*. *Protein. Sci*. **1996,** *5,* 1625–1632.
32. Roth, M. J.; Parks, B. A.; Ferguson, J. T.; Boyne, M. T. 2nd; Kelleher, N. L. "Proteotyping": Population proteomics of human leukocytes using top down mass spectrometry. *Anal. Chem*. **2008,** *80,* 2857–2866.
33. Waanders, L. F.; Hanke, S.; M., M. Top-down quantitation and characterization of SILAC-labeled proteins. *J. Am. Soc. Mass Spectrom*. **2007,** *18,* 2058–2064.