J|A|C|S
ARTICLES
Published on Web 01/05/2002

Top Down Characterization of Larger Proteins (45 kDa) by Electron Capture Dissociation Mass Spectrometry

Ying Ge, Brian G. Lawhorn, Mariam ElNaggar, Erick Strauss, Joo-Heon Park, Tadhg P. Begley, and Fred W. McLafferty*

Contribution from Baker Laboratory, Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301

Received May 31, 2001

Abstract: The structural characterization of proteins expressed from the genome is a major problem in proteomics. The solution to this problem requires the separation of the protein of interest from a complex mixture, the identification of its DNA-predicted sequence, and the characterization of sequencing errors and posttranslational modifications. For this, the "top down" mass spectrometry (MS) approach, extended by the greatly increased protein fragmentation from electron capture dissociation (ECD), has been applied to characterize proteins involved in the biosynthesis of thiamin, Coenzyme A, and the hydroxylation of proline residues in proteins. With Fourier transform (FT) MS, electrospray ionization (ESI) of a complex mixture from an *E. coli* cell extract gave 102 accurate molecular weight values (2−30 kDa), but none corresponding to the predicted masses of the four desired enzymes for thiamin biosynthesis (GoxB, ThiS, ThiG, and ThiF). MS/MS of one ion species (representing ∼1% of the mixture) identified it with the DNA-predicted sequence of ThiS, although the predicted and measured molecular weights were different. Further purification yielded a 2-component mixture whose ECD spectrum characterized both proteins simultaneously as ThiS and ThiG, showing an additional N-terminal Met on the 8 kDa ThiS and removal of an N-terminal Met and Ser from the 27 kDa ThiG. For a second system, the molecular weight of the 45 kDa phosphopantothenoylcysteine synthetase/decarboxylase (CoaBC), an enzyme involved in Coenzyme A biosynthesis, was 131 Da lower than that of the DNA prediction; the ECD spectrum showed that this is due to the removal of the N-terminal Met. For a third system, viral prolyl 4-hydroxylase (26 kDa), ECD showed that multiple molecular ions (+98, +178, etc.) are due to phosphate noncovalent adducts, and MS/MS pinpointed the overall mass discrepancy of 135 Da to removal of the initiation Met (131 Da) and to formation of disulfide bonds (2 × 2 Da) at $C_{32}-C_{49}$ and $C_{143}-C_{147}$, although 10 S−S positions were possible. In contrast, "bottom up" proteolysis characterization of the CoaBC and the P4H proteins was relatively unsuccessful. The addition of ECD substantially increases the capabilities of top down FTMS for the detailed structural characterization of large proteins.

Introduction

Proteomics, the analysis of the entire complement of proteins expressed by a cell, tissue, or organism, has become of key importance for biology and medicine with the completion of the human genome project.[1] Full proteome analysis requires separation, identification, and sequence characterization steps. In well-developed methodology,[2,3] the complex mixture from a given cell is separated by two-dimensional gel electrophoresis, the protein "spot" subjected to proteolysis, and the resulting peptides examined by mass spectrometry (MS). Their masses, plus MS/MS fragmentation data where necessary, identify the protein by matching against the DNA-predicted sequences. This can require only a small part of the protein sequence data; a "sequence tag" of 6−10 contiguous amino acids is often sufficient.[4]

However, the identified protein must often be characterized further structurally. DNA sequence errors can occur, resulting in protein sequence errors. In contrast to the genome, the proteome is not a fixed feature of an organism; instead, it changes with the state of development, the tissue, or even the environmental conditions. Variation can occur in gene splicing during mRNA processing, and posttranslational modifications can also occur.[5] Again, MS is the preferred technique for sequence characterization;[6] if the molecular weight does not

* Address correspondence to this author. Phone: (607)-255-4699. Fax: (607)-255-7880; E-mail: Fredwmcl@aol.com.

(1) Wilkins, M. R.; Sanchez, J.-C.; Gooley, A. A.; Appel, R. D.; Humphery-Smith, I. *Biotech. Gen. Eng. Rev.* **1995**, *13*, 19−50. Wilkins, M. R.; Williams, K. L.; Appel, R. D.; Hochstrasser, D. F., Eds. *Proteome Research: New Frontiers in Functional Genomics*; Springer-Verlag: Berlin, 1997. Jolles, P.; Jornvall, H., Eds. *Proteomics in Functional Genomics: Protein Structure Analysis*; Birkhauser: Basel, Boston, Berlin, 2000.

(2) (a) Andersen, J. S.; Svensson, B.; Roepstorff, P. *Nature Biotechnol.* **1996**, *14*, 449−457. (b) Qin, J.; Chait, B. T. *Anal. Chem.* **1997**, *69*, 4002−4009. (c) Thomas, J. J.; Bakhtiar, R.; Siuzdak, G. *Acc. Chem. Res.* **2000**, *33*, 179−187. (d) Palmblad, M.; Wetterhall, M.; Markides, K.; Hakansson, P.; Bergquist, J. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1029−1034.

(3) Yates, J. R., III *Trends Genet.* **2000**, *16*, 5−8. Pandey, A.; Mann, M. *Nature* **2000**, *405*, 837−846. Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001**, *101*, 269−295.

(4) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.

(5) Williams, K. L.; Hochstrasser, D. F. *Proteome Research: New Frontiers in Functional Genomics*; Springer-Verlag: Berlin, 1997; pp 1−12. Abbott, A. *Nature* **1999**, *402*, 715−720.

10.1021/ja011335z CCC: $22.00

agree with that predicted from the DNA sequence, the location of the modification(s) causing the mass difference is sought in pieces of the protein. For this, the "bottom up" approach utilizes the masses of the proteolysis products described above. However, these usually represent only 45−95% of the sequence,[2] and identification even of a single mass modification is difficult because many peptides of unpredicted masses are produced (e.g., self-proteolysis). The high-throughput characterization of sequencing errors and posttranslational modifications is still an unsolved problem.[7]

In the alternative "top down" strategy,[8] the conventional protein separation steps can be augmented with the unusually high ($>10^5$) resolving power of Fourier transform (FT) MS[2d,9] to "purify" further the intact protein ions produced by electrospray ionization of the original mixture. For the protein identification step, dissociation (MS/MS) of the selected protein molecular ion usually produces sufficiently unique mass values for successful matching against the DNA-predicted sequences.[10] For the sequence characterization step, the energetic ion dissociation methods such as collisionally activated dissociation (CAD)[11] and infrared multiphoton dissociation (IRMPD)[12] often produce 100% sequence coverage,[8] while the high mass accuracy of FTMS for molecular ions (±1 Da up to 50 kDa) and MS/MS fragment ions (± 0.1 Da up to 5 kDa)[13] gives far greater reliability in assigning multiple modifications to specific parts of the protein. Here the assignment specificity depends on the number of bond cleavages represented by the fragment masses; for the 259 residue carbonic anhydrase, Lys-C cleaved 16 bonds and chymotrypsin cleaved 62 bonds, while CAD cleaved 57 bonds.[8a] Increasing the assignment specificity by further CAD of the selected fragment ions (MS/MS/MS, MS^n) has been of limited help because the remaining bonds have much higher dissociation energies.

Thus a new MS/MS methodology that appears promising for improving the top down MS/MS approach is electron capture dissociation (ECD), whose fast ($<10^{-12}$ s) nonergotic dissociation of covalent protein backbone bonds generates far more, as well as unique cleavages.[15] Using "activated ion" (AI) dissociation of the ion's noncovalent tertiary structure prior to electron capture, ECD has been demonstrated for larger proteins of established sequences.[16] ECD of carbonic anhydrase cleaves 117 backbone bonds, with a combined total of 138 different bonds from ECD, CAD, and IRMPD spectra.[16]

To demonstrate its problem-solving capabilities, this expanded top down MS/MS approach is applied to characterize novel enzymes on three important biosynthetic pathways. ThiS and ThiG are required for the biosynthesis of the thiazole moiety of thiamin (Vitamin B_1).[17] CoaBC (45 kDa) has recently been identified as a bifunctional enzyme on the Coenzyme A biosynthesis pathway with both phosphopantothenoylcysteine synthetase and decarboxylase activities.[18] A novel viral prolyl 4-hydroxylase (P4H) has recently been identified and cloned from *Paramecium Bursaria Chlorella* Virus-1, a eukaryotic algal virus. This enzyme shows distinct sequence similarity to the C-terminal half of the catalytic α-subunit of mammalian P4H which catalyzes the hydroxylation of prolyl residues at X-Pro-Gly sequences in procollagen, an essential step in the biosynthesis of collagen, the major protein component of connective tissue.[19]

Experimental Section

Materials. Endoproteinase Lys-C, endoproteinase Glu-C, cyanogen bromide (CNBr), dithiothreitol (DTT), Tris, iodoacetic acid, and trifluoroacetic acid (TFA) from Sigma (St. Louis, MO) and TPCK Trypsin from Pierce (Rockford, IL) were used without further purification.

Strains, Plasmids, and Protein Purification. CoaBC was overexpressed in *E. coli* Tuner(DE3) with plasmid pCLK1210 as described previously.[18] The other recombinant proteins were overexpressed in *E. coli*. BL21(DE3) (Novagen). Viral P4H was overexpressed from plasmid pET30b-PBCV-1E36 and purified on a His-Bind Ni^{2+}-chelate affinity column (Novagen).[19c] The protein cluster containing GoxB, ThiS, ThiG, and ThiF was overexpressed from pCLK830, a pET-22b derived plasmid carrying the *B. subtilis goxBthiSGF* genes; the cell free extract was fractionated by precipitation with 60% ammonium sulfate. A portion was separated by one-dimensional sodium dodecyl sulfate polyacrylamide gel electrophoresis (1-D SDS-PAGE) and a second portion was analyzed directly by ESI/FTMS. ThiS and ThiG were also overexpressed with plasmid pCLK821, a pET-22b plasmid carrying the *B. subtilis thiSGF* genes and purified by ammonium sulfate fractionation followed by chromatography on DEAE sepharose and gel filtration.

(6) Kelleher, N. L.; Costello, C. A.; Begley, T. P.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 981−984. Wood, T. D.; Chen, L. H.; Kelleher, N. L.; Little, D. P.; Kenyon, G. L.; McLafferty, F. W. *Biochemistry* **1995**, *34*, 16251−16254. Kelleher, N. L.; Nicewonger, R. B.; Begley, A. P.; McLafferty, F. W. *J. Biol. Chem.* **1997**, *272*, 32215−32220. Roy, R. S.; Belshaw, P. J.; Walsh, C. T. *Biochemistry* **1998**, *37*, 4125−4136. Wilkins, M. R.; Gasteiger, E.; Gooley, A. A.; Herbert, B. R.; Molloy, M. P.; Binz, P. B.; Ou, K.; Sanchez, J.-C.; Bairoch, A.; Williams, K. L.; Hochstrasser, D. F. *J. Mol. Biol.* **1999**, *289*, 645−657. Fridriksson, E. K.; Beavil, A.; Holowka, D.; Gould, H. J.; Baird, B.; McLafferty, F. W. *Biochemistry* **2000**, *39*, 3369−3376. Cheng, X.; Cole, R. N.; Zaia, J.; Hart, G. W. *Biochemistry* **2000**, *39*, 11609−11620. Reilly, J. P.; Arnold, R. J. *Anal. Biochem.* **2000**, *269*, 105−112.

(7) Mann, M. *Nat. Biotechnol.* **1999**, *17*, 954−955.

(8) (a) Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W. *J. Am. Chem. Soc.* **1999**, *121*, 806−812. (b) McLafferty, F. W.; Fridriksson, E. K.; Horn, D. M.; Lewis, M. A.; Zubarev, R. A. *Science* **1999**, *284*, 1289−1290. (c) Kelleher, N. L. *Chem. Biol.* **2000**, *7*, R37−45. (d) Meng, F.; Cargile, B. J.; Miller, L. M.; Forbes, A. J.; Johnson, J. R.; Kelleher. N. L. *Nat. Biotechnol.* Accepted for publication.

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

(10) Mortz, E.; O'Connor, P. B.; Roepstorff, P.; Kelleher, N. L.; Wood, T. D.; McLafferty, F. W.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8264−8267.

(11) Gauthier, J. W.; Trautman, T. R.; Jacobson, D. B. *Anal. Chim. Acta* **1991**, *246*, 211−225. Senko, M. W.; Speir, J. P.; McLafferty, F. W. *Anal. Chem.* **1994**, *66*, 2801−2808.

(12) Little, D. P.; Speir, J. P.; Senko, M. W.; O'Connor, P. B.; McLafferty, F. W. *Anal. Chem.* **1994**, *66*, 2809−2815.

(13) Careful calibration with internal standards using high-field (e.g., 9.4 T) FT instruments can provide far higher mass accuracy, even subppm; this can be achieved at high molecular weights by minimizing the isotopic peak complexity using monoisotopic substrates for protein preparation.[2d,14]

(14) Marshall, A. G.; Senko, M. W.; Li, W.; Li, M.; Dillon, S.; Guan, S.; Logan, T. M. *J. Am. Chem. Soc.* **1997**, *119*, 433−434. Li, W.; Hendrickson, C. L.; Emmett, M. R.; Marshall, A. G. *Anal. Chem.* **1999**, *71*, 4397−4402. Smith, R. D. *Int. J. Mass Spectrom.* **2000**, *200*, 509−544.

(15) (a) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J. Am. Chem. Soc.* **1998**, *120*, 3265−3266. (b) Zubarev, R. A.; Kruger, N. A.; Fridriksson, E. K.; Lewis, M. A.; Horn, D. M.; Carpenter, B. K.; McLafferty, F. W. *J. Am. Chem. Soc.*, **1999**, *121*, 2857−2862. (c) Zubarev, R. A.; Horn, D. M.; Fridriksson, E. K.; Kelleher, N. L.; Kruger, N. A.; Lewis, M. A.; Carpenter, B. K.; McLafferty, F. W. *Anal. Chem.* **2000**, *72*, 563−573. (d) Cerda, B. A.; Horn, D. M.; Breuker, K.; McLafferty, F. W. *J. Am. Chem. Soc.* Submitted for publication.

(16) Horn, D. M.; Ge, Y.; McLafferty, F. W. *Anal. Chem.* **2000**, *72*, 4778−4784.

(17) Begley, T. P.; Downs, D. M.; Ealick, S. E.; McLafferty, F. W.; Van Loon, A. P. G. M.; Taylor, S.; Campobasso, N.; Chiu, H.-J.; Kinsland, C.; Reddick, J. J.; Xi, J. *Arch. Microbiol.* **1999**, *171*, 293−300.

(18) Strauss, E.; Kinsland, C.; Ge, Y.; McLafferty, F. W.; Begley, T. P. *J. Biol. Chem.* **2001**, *276*, 13513−13516.

(19) (a) Prockop, D. J.; Kivirikko, K. I. *Annu. Rev. Biochem.* **1995**, *64*, 403−434. (b) Wu, M.; Moon, H.-S.; Begley, T. P.; Myllyharju, J.; Kivirikko, K. I. *J. Am. Chem. Soc.* **1999**, *121*, 587−588. (c) Ericksson, M.; Myllyharju, J.; Tu, H.; Hellmam, M.; Kivirikko, K. I. *J. Biol. Chem.* **1999**, *274*, 2231−22134.

Proteolysis, Alkylation, and Reduction. For Lys-C or Glu-C proteolysis, 50−100 μg of P4H in 20 mM Tris (pH 7.5) was incubated with an equal volume of Lys-C or Glu-C (5−10 μg in 100 mM Tris, pH 8.5) for 3 h at 37 °C and quenched with 2−3 μL of acetic acid. Trypsin proteolysis used 0.5−1 μg (in 400 mM NH_4HCO_3, pH 8.0) of trypsin. For CNBr chemical cleavage, 70 μL of CNBr in formic acid (~1 g/mL) was added to 30 μL of P4H (2 mg/mL, 20 mM Tris, pH 7.5) at −80 °C, and the reaction mixture was incubated for 3 h at room temperature in the dark, dried in vacuo, and dried again after addition of 200 μL 0.1% TFA solution. For alkylation, 100 μg of P4H (0.6 mg/mL, 20 mM Tris, pH 7.8) was thawed, added to an equal volume of iodoacetic acid (10−30 mM in 50 mM Tris, pH 7.5), incubated at room temperature for 5−30 min, and quenched with 2−3 μL of acetic acid. For the disulfide reduction experiments, aliquots of P4H were reduced with 20 mM DTT at 37 °C for 3 h.

MS Analysis. All samples (20−100 μg) were desalted by ultrafiltration or by using reverse-phase protein/peptide traps (Michrom Bioresources, Auburn, CA), washed with 2:96:2 (MeOH:H_2O:AcOH), and step eluted with 70:26:4 (MeOH:H_2O:AcOH). Mass spectra were acquired on a 6 T modified Finnigan FTMS described previously[20] by using a nanospray[21] of 10^{-11}−10^{-12} mol samples. For MS/MS spectra, specific ions were isolated by using stored waveform inverse Fourier transform (SWIFT),[22] followed by sustained off-resonance irradiation (SORI) CAD,[11] nozzle-skimmer (NS) CAD, IRMPD,[12] blackbody infrared radiative dissociation (BIRD),[23] and "in-beam" AI ECD.[16] AI ECD used electron beam currents sufficient (0.1−0.35 μA) to dissociate almost all precursor ions. MS/MS spectra are averages of 20−100 scans. Assignment of the fragment masses and compositions used the computer program THRASH.[24] After each mass value, the mass difference (in units of 1.00235 Da) between the most abundant isotopic peak and the monoisotopic peak is denoted in italics.

Results and Discussion

In the first step of the top down approach, ESI of the sample produces molecular ions of its components, separated by using the MS high-resolution capability. This is important if conventional separation techniques are inadequate to yield pure samples or if the techniques are of undesirable complexity. As an example of proteins in a complex mixture, two proteins associated with thiamin biosynthesis in *B. subtilis* were characterized.

Identification of ThiS from a Complex Protein Mixture. Overexpression of the *B. subtilis goxBthiSGF* genes[17,25] in *E. coli* followed by a simple ammonium sulfate precipitation of the cell lysate gave a complex mixture of proteins as analyzed by 1-D SDS-PAGE (Figure 1a), demonstrating only a low level of overexpression of GoxB, ThiS, ThiG, and ThiF at the predicted mass of 40937, 7625, 27022, and 36400, respectively.[26] The only band (Figure 1a) that could correspond to ThiS represented ~1% of the total as determined by a densitometer scan. The proteins in the ammonium sulfate precipitate were desalted and introduced into the mass spectrometer to obtain an ESI/MS spectrum that shows 179 well-resolved isotopic clusters representing 102 distinctive masses

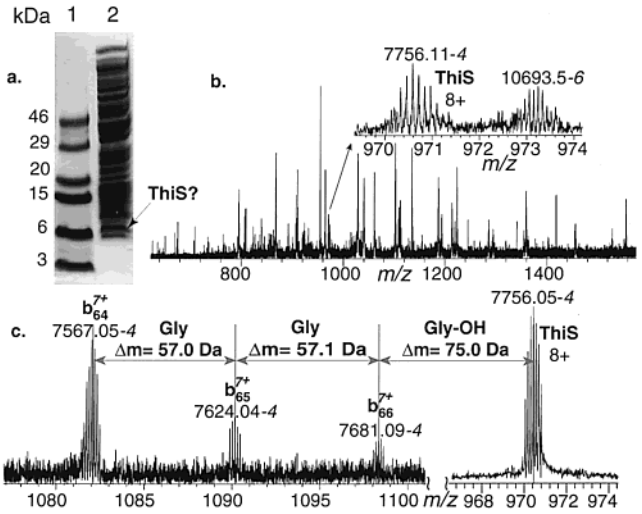

Figure 1. (a) Lane 1, molecular weight markers; lane 2, SDS-PAGE analysis of 60% ammonium sulfate fractionation of the cell lysate from overexpressed *B. subtilis goxBthiSGF* genes. (b) Partial ESI/FTMS spectrum of the mixture. (c) Partial SORI-CAD spectrum of the SWIFT-isolated 7756.11-*4* ions.

of 2−30 kDa (Figure 1b); the selective loss of larger ions could be due to preferential retention of smaller proteins by the peptide trap used for desalting. The GoxB, ThiF, ThiS, and ThiG molecular weights predicted from the *B. subtilis* genome sequence matched none of these 102 mass values.[26] For ThiS, 10 species were within ±0.5 kDa of the predicted mass (M_r = 7625); the three most abundant were SWIFT isolated and subjected to CAD. After trying the more abundant 7273.20-*4* and 7957.46-*4* ions, the 7756.11-*4* Da ions yielded a CAD spectrum clearly indicating the C-terminal sequence, Gly-Gly-Gly-OH (Figure 1c), that uniquely matches[10] the predicted sequence for ThiS from the genomic database[26] (note that SWIFT isolation substantially improves signal/noise, Figure 1c, right). Further characterization of ThiS to locate the 131 Da molecular weight discrepancy is described below. Thus minimal fractionation of the cell free extract and simple top down MS characterization have demonstrated the presence of a modified form of the ThiS protein in a mixture in which it is present at ~1% of the total protein and ~2% of the total molecular ions in the spectrum. The ECD spectrum of the 7756 Da ions was less definitive because of poor signal/noise, although useful ECD spectra have been obtained from ~1% mixture components.[15d]

ThiS and ThiG Protein Mixture. By using a different plasmid, the *B. subtilis thiSGF* genes were overexpressed in *E. coli*. After three stages of purification, the fraction whose ESI/MS spectrum is shown in Figure 2 gives two species of M_r = 7756.34-*4*, 131 Da higher than predicted for ThiS (vide supra), and 26803.6-*17*.

(20) Beu, S. C.; Senko, M. W.; Quinn, J. P.; Wampler, F. M., III; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 557−565.
(21) Wilm, M.; Mann, M. *Anal. Chem.* **1996**, *68*, 1−8.
(22) Marshall, A. G.; Wang, T. C. L.; Ricca, T. L. *J. Am. Chem. Soc.* **1985**, *107*, 7893−7897.
(23) Price, W. D.; Schnier, P. D.; Williams, E. R. *Anal. Chem.* **1996**, *68*, 859−866. Ge, Y.; Horn, D. M.; McLafferty, F. W. *Int. J. Mass Spectrom.* **2001**, *210/211*, 203−214.
(24) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass. Spectrom.* **2000**, *11*, 320−332.
(25) Taylor, S. V.; Kelleher, N. L.; Kinsland, C.; Chiu, H.-J.; Costello, C. A.; Backstrom, A. D.; McLafferty, F. W.; Begley, T. P. *J. Biol. Chem.* **1998**, *273*, 16555−16560.

(26) The report of the genome sequence of *B. subtilis* also contained the predicted protein sequences: Kunst, F.; Ogasawara, N.; Moszer, I.; Albertini, A. M.; Alloni, G.; Danchin, A.; et al. *Nature* **1997**, *390*, 249−256. There the sequence annotation and verification used (1) the GeneMark coding-sequence prediction, with search for coding-sequences preceded by a typical translation initiation signal (5′-AAGGAGGTG-3′), located 4−13 bases upstream of the putative start codons (ATG, TTG or GTG); (2) a BLAST2X analysis on the entire *B. subtilis* genome against the nonredundant protein databank at the NCBI; and (3) the distribution of nonoverlapping trinucleotides or hexanucleotides in the three frames of an open reading frame. Both proteins of Figures 3 and 4 have two methionines near the N-terminus, so that these methods can easily choose the wrong set of nucleotides encoding a methionine (AUG), as the neighboring sequences can affect the efficiency of translation initiation by the ribosome.[27]

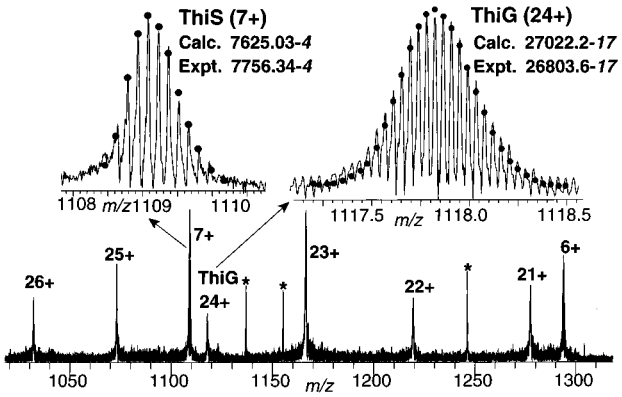

Figure 2. Broadband spectrum of purified ThiS/ThiG, 8 scans. Dots on the expanded portions represent the theoretical abundance distribution of the isotopic peaks corresponding to the assigned mass.

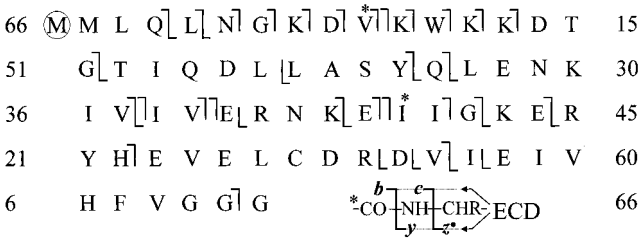

Figure 3. Product map from the in-beam AI ECD spectrum of the Figure 2 ions for assignments to the DNA-predicted sequence for ThiS, but adding an extra Met to the N-terminus. Asterisk: ***a***• ions; ***b*** and some ***y*** ions can arise from adventitious CAD.

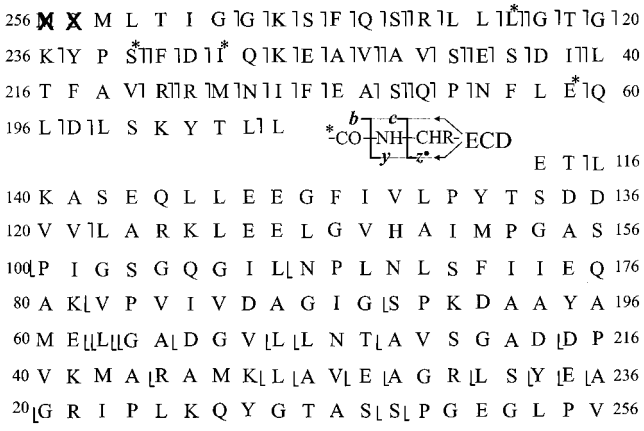

Figure 4. Product map, as in Figure 3, for ThiG with the removal of Met and Ser from the N-terminus.

To characterize this mixture, all ions were subjected simultaneously to MS/MS using AI ECD. Masses derived from the resulting spectrum can be assigned to fragments (or fragment mass differences) expected from the predicted sequence of ThiS; 24 ***c***, 4 ***b***, 13 ***z***•, and 1 ***y*** ions (Figure 3) result from 30 cleavages of its 66 bonds. All of the N-terminal ***c*** and ***b*** type ions contain the extra 131 Da mass discrepancy including the smallest fragment, c_3, indicating another Met at the N-terminus. The remaining ECD products of the 7756/26804 Da ion mixture can form extensive sequence tags that are consistent with the DNA derived sequence of ThiG (Figure 4). The 39 ***c***, 13 ***b***, 13 ***z***•, and 12 ***y*** ions correspond to cleavage of 68 out of its 253 bonds and can be assigned if the 218.6 Da mass discrepancy corresponds to loss of the amino terminal Met and Ser (131.2 + 87.1 Da). Thus a single AI ECD spectrum has characterized ThiS and ThiG simultaneously.

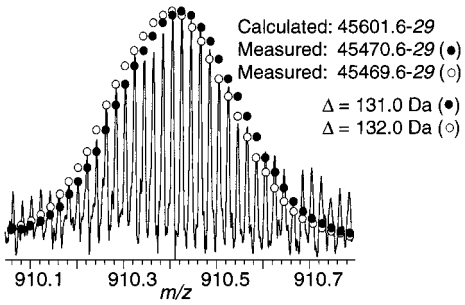

Figure 5. $(M + 50H)^{50+}$ ions in the ESI/FTMS spectrum of CoaBC, 25 scans.

The addition of Met to ThiS and the deletion of both Met and Ser from ThiG are not known posttranslational modifications,[28] so they are more likely due to the misassignment of the initiation codon predicted from the genomic sequence using automated annotation programs.[26,27] The more efficient sequencing of the N-terminus of a protein, for which ECD appears to be especially effective,[16,29] will facilitate the identification of such misassignments.

CoaBC (45 kDa). This *E. coli* protein has phosphopantothenoylcysteine synthetase and phosphopantothenoylcysteine decarboxylase activities.[18] The DNA sequence predicts M_r = 45602. The ESI mass spectrum shows 30+ to 62+ molecular ion peaks, with M^{50+} yielding M_r = 45470.6-*29* (Figure 5; note that 45469.6 fits nearly as well), 131 lower than the predicted values. For comparison of the bottom up and top down approaches, proteolysis was first used to localize this mass difference, limiting the reaction time to produce larger peptides and more extensive sequence coverage. The products from Lys-C digestion, without separation, gave an ESI/MS spectrum showing 168 discernible isotopic clusters representing 45 distinct mass values (Table 1). Of these, nine correspond to the M_r values of an expected Lys-C digested peptide and one to a value 131 lower than that predicted for the peptide M^1-K^{207}, a sequence coverage of all but 13 of the 426 residues. Of the remaining 35 fragment masses, 10 could represent H_2O loss or artificial adducts potentially formed from methanol, while peptides from self-proteolysis or without a C-terminal Lys are possible. Thus, this step of the bottom up approach mainly restricted the error to the first 207 amino acids.

In the alternative top down approach, both CAD and IRMPD were applied, yielding a total of 16 bond cleavages. The smallest fragment containing the 131 Da error is $\boldsymbol{b}_{50}$, localizing this error further to within the first 50 amino acids from the N-terminus (Figure 6). A single AI ECD spectrum, in contrast, shows cleavages of 63 bonds; all 39 ***c*** ions, including the lowest mass $\boldsymbol{c}_6$, were identified as 131 Da lower than the predicted mass values, localizing the error to the first six N-terminal residues. As there is only one Met (131 Da) in the first 6 residues, the 131 Da error can be assigned as the removal of the N-terminal Met, the most common posttranslational modification.[28] Obviously, the ECD spectrum would have also supplied detailed information on modifications in the first 60−80 residues of either terminus. Further, phosphopantothenoylcysteine syn-

(27) Mann, M.; Pandey, A. *Trends Biochem. Sci.* **2001**, *26*, 54−60.
(28) Krishna, R. G.; Wold, F. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1993**, *67*, 265−298.
(29) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10313−10317.

Table 1. Peptide Assignments for Lys-C Proteolysis Products of 426 Residue CoaBC

mass	assignment	error, Da	mass	assignment	error, Da
21440.9-*13*	M^{1}-K^{207}	−131.4	1166.67-*0*[b]	N^{327}-K^{338}	0.00
3103.59-*1*[a]	H^{208}-K^{235}	−0.01	2693.31-*1*[a,b]	D^{339}-K^{361}	+0.01
3166.74-*1*[a,b]	M^{236}-K^{267}	−0.01	3690.63-*2*[b]	N^{366}-K^{398}	−0.05
4452.25-*2*[b]	R^{268}-K^{309}	+0.05	853.54-*0*	R^{399}-K^{405}	0.00
1202.60-*0*	Q^{313}-K^{323}	−0.01	2147.12-*1*[a,b]	E^{406}-K^{423}	−0.03

No match:[c] 522.25-*0*, 608.27-*0*, 635.34-*0*, 1141.51-*0*, 1185.60-*0*, 1379.69-*0*, 1492.75-*0*, 1527.80-*0*, 1605.86-*0*, 1742.86-*1*, 1842.01-*1*, 1986.05-*1*, 2210.08-*1*, 2269.20-*1*, 2382.09-*1*, 4483.37-*2*, 7002.53-*4*, 7640.52-*4*, 7810.27-*4*, 7898.33-*4*, 8119.31-*5*, 11485.6-*7*, 13643.8-*8*, 20420.2-*13*, 21471.8-*13*

[a] Ion of 18 Da lower mass also found. [b] Ion of 14 Da higher mass also found; however, no corresponding +14 Da molecular ions were observed. [c] Assignments were only attempted for specific Lys-C cleavages.

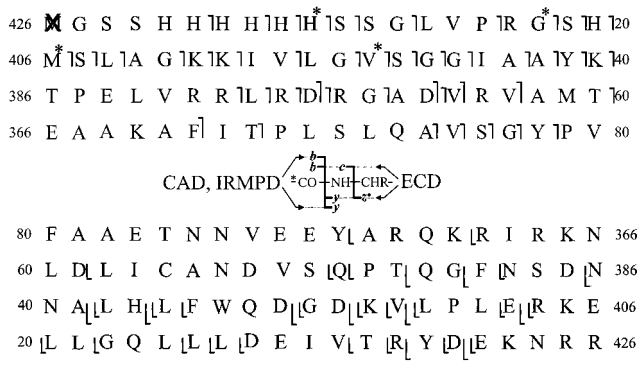

Figure 6. Product map from nozzle-skimmer CAD (30 scans), IRMPD (46 scans), and in-beam AI ECD (96 scans) spectra with assignments to the DNA-predicted sequence of CoaBC with removal of N-terminal Met.

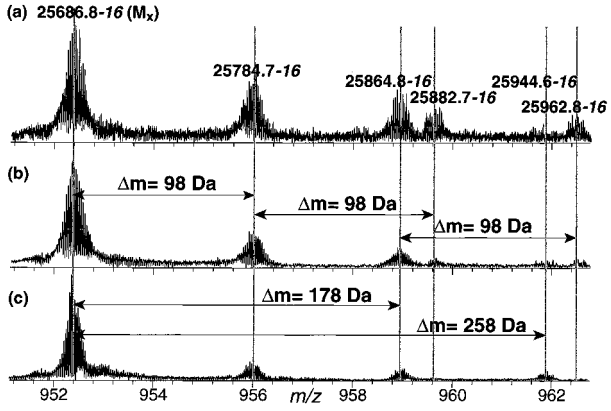

Figure 7. Partial ESI/FTMS spectrum of P4H (27+) overexpressed for (a) 12 h, (b) 3 h, and (c) 1 h.

thetase/decarboxylase was previously described as a pyruvoyl containing protein;[30] the top down data also demonstrate that this cannot be correct.

Viral Prolyl 4-Hydroxylase (26 kDa). The ESI/MS spectrum of this enzyme, purified after 12 h of overexpression in *E. coli*, showed mass values of 25687 (M_x), 25785, 25865, 25883, 25945, and 25963 (Figure 7a); none agree with the DNA-predicted value of 25821.9. Moreover, samples purified after shorter periods of overexpression show lower abundances of the higher mass ions (Figure 7b,c), suggesting adduct peaks. Such 98 Da adducts observed previously[31] in ESI of certain proteins can arise from the noncovalent attachments of H_2SO_4 or H_3PO_4,[32] while covalent phosphorylation of proteins[33] yields

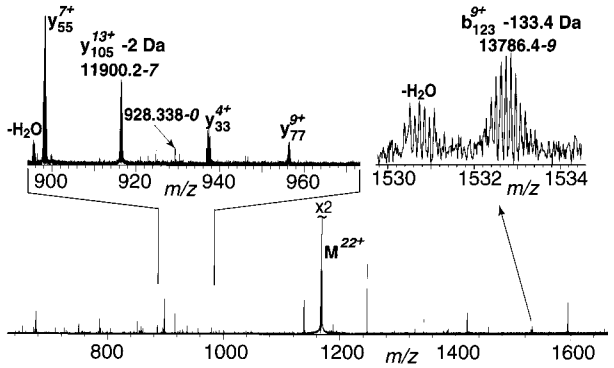

Figure 8. SORI-CAD spectrum (50 scans) of 25687 Da ions of the Figure 7a spectrum.

a mass difference of 80 Da. MS/MS spectra of the Figure 7 ions measured by BIRD, NS-CAD, IRMPD, and AI ECD show extensive backbone fragmentation. Although none of these products show +80 or +98 counterpart ions, covalent modifications such as phosphorylation can be cleaved off by the energetic methods CAD, IRMPD, and BIRD.[33] In contrast, ECD causes negligible loss of side chain posttranslational modifications,[33,34] so that the +98, +178, and +258 Da peaks of Figure 7 must arise from strong noncovalent attachments,[35] presumably to the 25687 Da (M_x) species, of H_3PO_4, $H_3P_2O_7$ (pyrophosphate), and $H_3P_3O_{10}$ (triphosphate).

The 25686.8 Da value (Figure 7a) of the P4H molecular ions is 135 Da lower than the DNA predicted mass. Deletion of neither His (137 Da) nor Met (131 Da) would yield this value within the ±1 Da accuracy of FTMS for such proteins,[8] so that the predicted protein sequence must contain multiple modifications and/or errors. Using the top down approach, MS/MS of the 22+ molecular ion gave a CAD spectrum (Figure 8) in which y_{77} is the largest of many C-terminal fragments, all of whose masses correspond to the DNA-predicted sequence (Figure 9); this localizes the errors to the first 151 amino acids from the N-terminus. The CAD spectrum shows a complemen-

(30) Yang, H.; Abeles, R. H. *Biochemistry* **1987**, *26*, 4076−4081.
(31) Henry, K. D.; Williams, E. R.; Wang, B. H.; McLafferty, F. W.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 9075−9078. Loo, J. A.; Edmonds, C. G.; Smith, R. D.; Lacey, M. P.; Keough, T. *Biomed. Environ. Mass Spectrom.* **1990**, *19*, 286−294.
(32) Chowdhury, S. K.; Katta, V.; Beavis, R. C.; Chait, B. T. *J. Am. Soc. Mass Spectrom.* **1990**, *1*, 382−388. Bruce, J. E.; Hofstadler, S. A.; Winger, B. E.; Smith, R. D. *Int. J. Mass Spectrom. Ion Processes* **1994**, *132*, 97−107. Purves, R. W.; Barnett, D. A.; Guevremont, R. *Int. J. Mass Spectrom.* **2000**, *197*, 163−177.
(33) Stensballe, A.; Jensen, O. N.; Olsen, J. V.; Haselmann, K. F.; Zubarev, R. A. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1793−1800. Shi, S. D.-H.; Hemling, M. E.; Carr, S. A.; Horn, D. M.; Lindh, I.; McLafferty, F. W. *Anal. Chem.* **2001**, *73*, 19−22.
(34) Kelleher, N. L.; Zubarev, R. A.; Bush, K.; Furie, B.; Furie, B. C.; McLafferty, F. W.; Walsh, C. T. *Anal. Chem.* **1999**, *71*, 4250−4253. Mirgorodskaya, E.; Roepstorff, P.; Zubarev, R. A. *Anal. Chem.* **1999**, *71*, 4431−4436.
(35) The increased adduct abundance with the duration of overexpression (Figure 7) could result from increased disulfide bond formation. The proportion of adducts is greatly reduced by reduction of the disulfide bonds (vide infra).

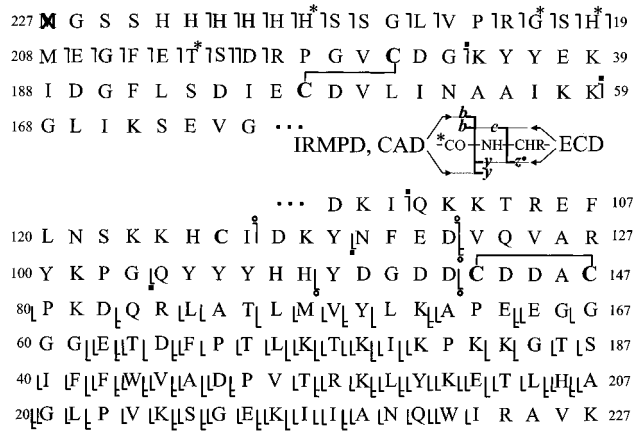

Figure 9. Product map from MS/MS of P4H ions (Figure 7a) using nozzle-skimmer CAD, SORI-CAD, IRMPD, and in-beam AI ECD (Figure 10) spectra, with assignments based on the removal of N-terminal Met and addition of two S−S bonds to modify the DNA-derived sequence: ■ and ○, experimental values 1 and 2 Da, respectively, lower than those predicted.

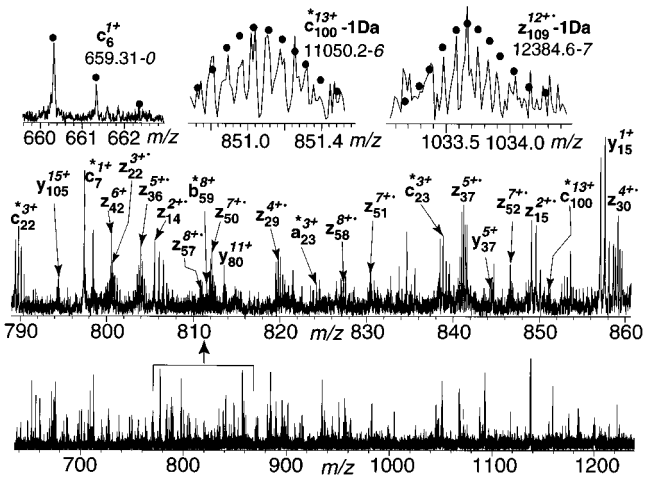

Figure 10. In-beam AI ECD spectrum of P4H overexpressed for 12- h, 100 scans. Asterisk: assigned as N-terminal Met removal.

tary product pair at 11900.2 and 13786.4 Da (sum = 25686.6, M_r), for which the only possible assignments appear to be ($\boldsymbol{y}_{105}$ − 2 Da) and ($\boldsymbol{b}_{123}$ − 133.4 Da), respectively. The missing 2 Da could indicate formation of a disulfide bond (internal 2SH → S−S + 2H). However, this is unexpected in a protein that is overexpressed in *E. coli* because stable disulfide bonds rarely form in the bacterial cytosol,[36] which is known to be more reducing than its eukaryotic counterpart and is not a good environment for the production of properly folded, multiply disulfide-bonded proteins.[37] For the $\boldsymbol{y}_{105}$ fragment, the DNA predicted sequence indicates only two cysteines, with neither in the region covered by $\boldsymbol{y}_{77}$, so that the disulfide bond must be between C_{143} and C_{147} (Figure 9). An IRMPD spectrum of the 20+ to 26+ molecular ions confirmed these data and further localizes the remaining −133.4 Da to the first 116 residues from the N-terminus.

The AI ECD spectrum contains 277 isotopic clusters representing 151 distinctive masses (Figure 10). These are consistent with 54 $\boldsymbol{z}^{\bullet}$ and 28 $\boldsymbol{y}$ ions, all of which fit the first 80 C-terminal residues of the DNA predicted sequence including a 26-residue sequence tag, 198−223 (Figure 9). The masses of the ECD peaks designated as $\boldsymbol{z}^{\bullet}_{96}$ and $\boldsymbol{z}^{\bullet}_{109}$ are 1 Da higher than the masses predicted by the sequence with the $C_{143}-C_{147}$ S−S bond modification; a separate (and favored) electron capture can cleave the S−S bond to add an H$^{\bullet}$ atom (−S−S− + H$^{\bullet}$ → −SH + $^{\bullet}$S−) at one of the Cys residues.[15b]

For the N-terminal assignments from the remaining ECD data, the sequence tags HHHHS and RGSHMEGFETSD obviously can be fit to the DNA-predicted sequence by removal of the N-terminal Met (Figure 9), accounting for all but 2 Da of the remaining mass discrepancy. That this arises from another S−S bond is also indicated by the +1 Da mass discrepancies for the ECD products $\boldsymbol{c}_{34}$, $\boldsymbol{b}_{59}$, and $\boldsymbol{c}_{100}$.[38] All of these could result from the extra H$^{\bullet}$ on Cys-32; thus the ECD data are consistent with either the $C_{32}-C_{49}$ or the $C_{32}-C_{114}$ S−S bonds. To distinguish between these final two possibilities, it was necessary to use "internal" (***i***) product ions that result from further dissociation of ***b*** or ***y*** product ions. Multiple assignments are more common for these because all possible ***i*** products resulting from two bond cleavages must be considered.[8,11,12] These internal products are only consistent with no S−S bonds to C_{114}; in the IRMPD spectrum, mass values of 3355.73-*2* and 8619.16-*5* Da correspond to $\boldsymbol{i}_{114-140}$ and $\boldsymbol{i}_{101-174}$ (errors +0.22 and +0.12 Da, respectively), and a nozzle-skimmer CAD spectrum shows 5366.84-*3* Da corresponding to $\boldsymbol{i}_{70-140}$ (6531.29-*4*; $\boldsymbol{i}_{115-171}$, confirms $C_{143}-C_{147}$). For internal ions with multiple assignments, none was consistent with an S−S bond to C_{114}.

For this modified P4H structure, 143 of the 151 ECD mass values could be assigned, including 4 ***a***, 2 ***b***, 23 ***c***, 54 $\boldsymbol{z}^{\bullet}$, 28 ***y***, and 20 ***i*** ions. Another 12 masses corresponded to the common losses of 17, 45, and 60 Da from side chain dissociation of **c**, $\boldsymbol{z}^{\bullet}$ fragments. Overall, the ECD spectrum shows 85 of the 226 possible interresidue cleavages, versus 44 for the combined CAD and IRMPD spectra, a great advantage for characterizing further modifications of this protein.

Several conventional sample treatments were used to compare their utility to locate 2 of the 10 possible disulfide bond positions. Treatment of P4H with iodoacetic acid gave ∼20% of an $(M + 58)^{n+}$ ESI product that should arise from alkylation of the free thiol on the one Cys not involved in a disulfide bond. IRMPD and ECD spectra of $(M + 58)^{n+}$ (data not shown) contain peaks corresponding to unmodified C_{32}, C_{49}, C_{143}, and C_{147}, but none with the +58 Da modification. Therefore, the C_{114} cysteine, or some more reactive site of the protein, has been alkylated.

The ESI spectrum (not shown) of a proteolytic Glu-C digest indicates four disulfide bonds, $C_{32}-C_{49}$, $C_{143}-C_{147}$, $C_{32}-C_{114}$, and $C_{49}-C_{114}$. The spectrum of a trypsin digest shows major disulfide patterns for $C_{32}-C_{49}$ and $C_{143}-C_{147}$, but with minor peaks for $C_{32/49}-C_{143/147}$ and $C_{32/49}-C_{114}$. Lys-C proteolysis indicates $C_{32}-C_{49}$, $C_{143}-C_{147}$, $C_{32}-C_{114}$, $C_{49}-C_{114}$, and traces of $C_{32/49}-C_{143/147}$. Apparently the one of the five cysteines not involved in forming the disulfide bonds is sufficiently nucleophilic under the basic proteolysis conditions to attack the other disulfide bonds and cause disulfide scrambling. CNBr chemical

(36) Locker, J. K.; Griffiths, G. *J. Cell. Biol.* **1999**, *144*, 267−279.
(37) Woycechowsky, K. J.; Raines, R. T. *Curr. Opin. Chem. Biol.* **2000**, *4*, 533−539.

(38) Only one of the two complementary fragment ions will thus add the H$^{\bullet}$ atom, so that a fragment ion containing only one of the S−S bond Cys residues could be formed either with or without the extra 1 Da. Such an ion mixture will affect the relative abundances of the resulting isotopic peaks, but not their masses; the THRASH program[24] thus will give the mass value of the more abundant component, not a value intermediate between their two mass values (see the example in Figure 5).

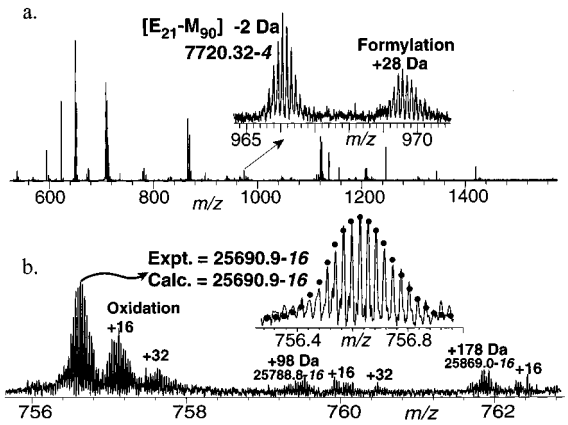

Figure 11. (a) ESI spectrum of products from CNBr chemical cleavage of P4H. (b) Partial ESI spectrum of reduced P4H showing 34+ ion products.

cleavage utilizes acidic conditions that should prevent this attack on S−S by the free thiol. The ESI spectrum of the CNBr digest shows a peptide of M_r = 7720.32-4 that corresponds to E_{21}-M_{90} with a disulfide bond between C_{32} and C_{49} (Figure 11a), while CAD of the 15676.3-9 Da peptide ions gave a D_{122}-M_{157} fragment that confirmed the other disulfide bond, C_{143}−C_{147} (data not shown).

After treatment of P4H with dithiothreitol, the ESI spectrum gave M_r = 25690.9-*16*, 4 Da higher than the native form, as expected for the reduction of two disulfide bonds (Figure 11b). Much lower degrees of phosphate and pyrophosphate adducts were observed after the DTT treatment, although oxidation (+16, +32 Da) is indicated. The IRMPD spectrum of these ions showed additional bond cleavages yielding products $\boldsymbol{b}_{19}$, $\boldsymbol{b}_{24}$, $\boldsymbol{b}_{28}$, $\boldsymbol{b}_{33}$, $\boldsymbol{b}_{37}$, $\boldsymbol{b}_{62}$, $\boldsymbol{b}_{73}$, $\boldsymbol{b}_{78}$, $\boldsymbol{b}_{84}$, $\boldsymbol{y}_{129}$, $\boldsymbol{b}_{150}$, while the ECD spectrum showed additional bond cleavages yielding products $\boldsymbol{c}_{33}$, $\boldsymbol{c}_{35}$, $\boldsymbol{c}_{36}$, $\boldsymbol{c}_{37}$, $\boldsymbol{c}_{38}$, $\boldsymbol{c}_{39}$, $\boldsymbol{c}_{40}$, $\boldsymbol{c}_{41}$, $\boldsymbol{a}^{\bullet}_{43}$, $\boldsymbol{z}^{\bullet}_{61}$, $\boldsymbol{c}_{62}$, $\boldsymbol{c}_{63}$, $\boldsymbol{b}_{66}$, $\boldsymbol{c}_{72}$, $\boldsymbol{c}_{77}$, $\boldsymbol{b}_{78}$, $\boldsymbol{z}^{\bullet}_{104}$, $\boldsymbol{y}_{100}$, $\boldsymbol{z}^{\bullet}_{94}$, $\boldsymbol{y}_{80}$, $\boldsymbol{z}^{\bullet}_{60}$, $\boldsymbol{z}^{\bullet}_{59}$, $\boldsymbol{z}^{\bullet}_{56}$, $\boldsymbol{z}^{\bullet}_{43}$, $\boldsymbol{z}^{\bullet}_{32}$, $\boldsymbol{y}_{16}$, reflecting the unfolding in the disulfide regions.[15b] This is also consistent with the dramatic decrease of the phosphate adducts (Figure 11b). Improved methods for S−S bond cleavage in MS/MS alone are under investigation to acquire this extra sequence information without sample reduction; for some proteins, ECD cleavage of S−S bonds is favored over $\boldsymbol{c}$, $\boldsymbol{z}^{\bullet}$ formation.[15b]

Despite extensive phosphate, pyrophosphate, and triphosphate adducts formed in ESI, the CAD and ECD spectra have completely characterized three unexpected sequence modifications in this viral prolyl-4-hydroxylase, two disulfide bonds and removal of the amino terminal methionine. The two disulfide bonds should reflect the native structure, rather than a misfolded form of the enzyme, because the overexpressed enzyme is catalytically active. Because of the reducing environment in the bacterial cytosol, these disulfide bonds are probably formed after the protein was released from the cell. These two identified disulfide bonds appear to be stabilized, as expected,[39] by the formation of the native structure, since (i) the other eight possible disulfide bonds are not detected and (ii) 2 mM DTT, which can readily reduce any disulfide bond in an unstructured species, is not sufficient to reduce these bonds; 20 mM DTT gave total reduction. Proteolysis scrambled the disulfide bonds; with this destruction of the native structure, the intermolecular free Cys can initiate fast intermolecular reshuffling of the S−S bonds.[40]

The identification of the cysteine partners involved in a disulfide bond is important not only for the characterization of native proteins but also for the identification of disulfide intermediates during protein folding.[41] This identification usually involves blocking of free cysteines to avoid disulfide interchange,[39,42] followed by chemical[43] and proteolytic[44] cleavage of the protein, and subsequent analysis of the resulting peptides. The top down methodology replaces these steps with MS/MS measurements, greatly simplifying the procedure.

Conclusions

Top down MS identification can be applied directly to relatively complex protein mixtures by utilizing the unusually high resolving power of FTMS, a valuable complement to conventional separation techniques. The high mass accuracy of FTMS can supply specific characterization of multiple sequence modifications, even identifying the loss of 2 Da in a 11 900 Da fragment ion that shows unexpected S−S bond formation in a viral protein. The original top down MS/MS localization of mass modifications (CAD, IRMPD, etc.)[8] is made far more precise and efficient with the addition of ECD, as it can cleave several times as many backbone bonds. Further, its $\boldsymbol{c}$, $\boldsymbol{z}^{\bullet}$, and $\boldsymbol{a}^{\bullet}$ terminal fragments from cleavage of a single bond do not produce measurable secondary $\boldsymbol{i}$ products, so that structural assignments are more specific than for other MS/MS or bottom up (proteolysis) methods in which a combination of two bonds must be considered. For proteolysis of the CoaBC protein, only a fraction of its product masses could be assigned, and proteolysis of P4H led to scrambling of the S−S bonds.

The possibilities for automation of the top down approach have been demonstrated in the complete de novo sequencing of ubiquitin (8.6 kDa),[29] and previous CAD MS/MS identifications with 10^{-17} mol protein samples[45] promise high sensitivity for this approach. With the addition of the far more definitive MS/MS data of ECD, the top down methodology is a promising alternative to the conventional bottom up approach for efficient and accurate protein structural characterization as well as identification.

Acknowledgment. The authors gratefully acknowledge Cynthia Kinsland, Johanna Myllyharju, and Kari Kivirikko for the overexpression plasmids; Ervin Welker, David Horn, Kathrin Breuker, Neil Kelleher, Blas Cerda, Julian Whitelegge, Aaron Frank, Han Bin Oh, Newman Sze, and Harold Scheraga for helpful discussions; and the National Institutes of Health (grants GM16609 and DK44083) for generous financial support.

JA011335Z

(39) Welker, E.; Narayan, M.; Wedemeyer, W. J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2312−2316.
(40) Welker, E.; Wedemeyer, W. J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4334−4336.
(41) Welker, E.; Narayan, M.; Volles, M. J.; Scheraga, H. A. *FEBS Lett.* **1999**, *460*, 477−479.
(42) Thannhauser, T. W.; Sherwood, R. W.; Scheraga, H. A. J. *Protein Chem.* **1998**, *17*, 37−43.
(43) England, P. M.; Lester, H. A.; Dougherty, D. A. *Biochemistry* **1999**, *38*, 14409−14415. Qi, J.; Wu, J.; Somkuti, G. A.; Watson, J. T. *Biochemistry* **2001**, *40*, 4531−4538.
(44) Lu, H. S.; Jones, M. D.; Patterson, S. D.; *Anal. Chem.* **1998**, *70*, 136−143. Marie, G.; Serani, L.; Laprevote, O. *Anal. Chem.* **2000**, *72*, 5423−5430.
(45) Valaskovic, G. A.; Kelleher, N. L.; McLafferty, F. W. *Science* **1996**, *273*, 1199−1202.