Protein Science (1998), 7:1796–1801. Cambridge University Press. Printed in the USA.

Efficient sequence analysis of the six gene products (7–74 kDa) from the *Escherichia coli* thiamin biosynthetic operon by tandem high-resolution mass spectrometry

NEIL L. KELLEHER, SEAN V. TAYLOR, DAVID GRANNIS, CYNTHIA KINSLAND, HSIU-JU CHIU, TADHG P. BEGLEY, AND FRED W. McLAFFERTY

Department of Chemistry, Baker Laboratory, Cornell University, Ithaca, New York 14853-1301

(RECEIVED December 30, 1997; ACCEPTED April 14, 1998)

Abstract

The 10^5 resolving power and MS/MS capabilities of Fourier-transform mass spectrometry provide electrospray ionization mass spectra containing >100 molecular and fragment ion mass values of high accuracy. Applying these spectra to the detection and localization of errors and modifications in the DNA-derived sequences of proteins is illustrated with the *thiCEFSGH* thiamin biosynthesis operon from *Escherichia coli*. Direct fragmentation of the multiply-charged intact protein ions produces large fragment ions covering the entire sequence; further dissociation of these fragment ions provides information on their sequences. For ThiE (23 kDa), the entire sequence was verified in a single spectrum with an accurate (0.3 Da) molecular weight (M_r) value, with confirmation from MS/MS fragment masses. Those for ThiH (46 kDa) showed that the M_r value (1 Da error) represented the protein without the start Met residue. For ThiF (27 kDa), MS/MS localized a sequence discrepancy to a 34 residue peptide. The first 107 residues of ThiC (74 kDa) were shown to be correct, with C-terminal heterogeneity indicated. For ThiG (predicted M_r = 34 kDa), ESI/FTMS showed two components of 7,310.74 (ThiS) and 26,896.5 Da (ThiG); MS/MS uncovered three reading frame errors and a stop codon for the first protein. MS/MS ions are consistent with 68 fragments predicted by the corrected ThiS/ThiG DNA sequences.

Keywords: electrospray; *Escherichia coli*; Fourier-transform mass spectrometry; operon; sequence; thiamin biosynthesis; *thiCEFSGH*

With the genome of *Escherichia coli* completely sequenced (Blattner et al., 1997), it may become the first organism with a known function for every gene. For the 4,288 putative proteins encoded by these genes, functional predictions are often made based on sequence homology or distinctive motifs, and without exact knowledge of the mature protein's sequence. Currently, genomic database errors are estimated at 0.4–2.9 errors per 1,000 bases (Krawetz, 1989; Mewes & George, 1994), with corresponding errors and modifications in DNA-predicted protein sequences usually detected by re-sequencing the DNA, Edman sequencing of selected peptides, or by NMR or X-ray crystal structure determination. These methods have significant limitations. Re-sequencing the DNA can often give the same errant result due to G-C compression. Edman sequencing requires isolation of a small peptide containing the sequencing error. Structure determination by NMR or X-ray crystallography requires highly pure protein as well as unpredictable time periods for error detection. Recently, peptide mapping by mass spectrometry (MS) has developed into a facile method for protein microcharacterization by surveying peptides (0.5–3 kDa) from the protein for discrepancies in the molecular weights (M_r) predicted from a DNA-derived sequence (Shevchenko et al., 1996; Blackburn & Anderegg, 1997; Yates, 1998). However, this method relies on extensive protein degradation (often employing peptide chromatography prior to MS), and sequence coverage is only 50–90%, even for 10^{-12} mol of protein (Shevchenko et al., 1996; Blackburn & Anderegg, 1997).

We illustrate here a more specific and faster method for characterization of proteins, even those of marginal purity, by M_r measurement and direct fragmentation of multiply-charged protein ions from electrospray ionization (ESI) (Fenn et al., 1990) combined with tandem (MS/MS) Fourier-transform mass spectrometry (FTMS) (Comisarow & Marshall, 1974; Marshall &

Reprint requests to: Fred W. McLafferty, Department of Chemistry, Baker Laboratory, Cornell University, Ithaca, New York 14853-1301; e-mail: fredwmcl@aol.com.

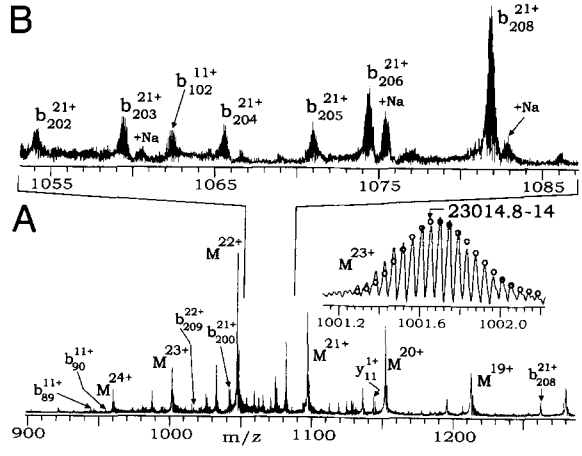

Fig. 1. A: Broadband ESI/FT mass spectrum of the ThiE gene product, single scan, 120 V nozzle skimmer potential difference. Inset, expansion of $(M + 23H)^{23+}$ ions; O, theoretical abundance distribution of the isotopic peaks. **B:** Expansion of m/z 1,055–1,085.

Grosshans, 1991). This combination (McLafferty, 1994; Williams, 1998) has proven uniquely powerful for isotopic resolution (resolving power $>10^5$) of protein ions and their ion fragmentation products. Such ion masses can cover the complete sequence of proteins as large as 43 kDa (Kelleher et al., 1995; Wood et al., 1995), even at the 10^{-17} mol level (Valaskovic et al., 1996), achieving mass accuracies of ±0.05 Da up to 5 kDa, ±0.2 Da up to 20 kDa, and ±1 Da to 50 kDa.

The gene products examined here, those predicted by a five gene cluster at 90 min on the *E. coli* genome from two published sequences (Blattner et al., 1993; Vander Horn et al., 1993), are required for the biosynthesis of thiamin (Vander Horn et al., 1993). Of the five initially predicted gene products, ESI/FTMS shows that only two have the DNA-predicted sequence. Further, for one product (originally assigned as ThiG), the identification and localization of a reading frame error predicted an additional gene product (ThiS). This discovery was critical to the identification of ThiS as the immediate sulfur source for the biosynthesis of the thiazole moiety of thiamin (Taylor et al., 1998).

Results and discussion

Five proteins are predicted from the DNA sequence of the *thiCEFGH* operon from *E. coli* (Blattner et al., 1993; Vander Horn et al., 1993; GenBank accession numbers are U00006 and M88701, respectively).

ThiE

This protein has been identified as thiamin phosphate synthase (Backstrom et al., 1995). From the ESI/FTMS spectrum, the measured M_r value of 23,014.8-*14*[1] (Fig. 1A, inset) agrees within 13 ppm of the predicted value of 23,015.1-*14* from both published DNA sequences. This is strong evidence for an entirely correct sequence with retention of a deformylated start methionine and no disulfide bonds. Note the 11 Da width at half-height of this isotopic cluster of peaks; if these were an unresolved envelope (as measured by all instruments except FTMS), ions typically formed by H_2O loss (−18 Da) or Na adduction (+22 Da) would produce similarly broad, overlapping envelopes, shifting the centroid of the resulting broader peaks by an unpredictable amount. Thus, the verification of the DNA-derived sequence would be far less reliable.

However, collisionally activated dissociation (CAD) of these molecular ions caused by the relatively high nozzle skimmer (NS) inlet potential difference of 120 V in the ESI source also produced 14 fragment ions (Fig. 1B), all of whose mass values are assignable (±30 ppm) to *b*- or *y*-type ions (i.e., those containing the N- and C-terminus, respectively) (Roepstorff & Fohlman, 1984) predicted from the sequence. Of these, two ion pairs (b_{200}/y_{11} and b_{206}/y_5) sum to the observed M_r value within 0.0 and 0.2 Da, respectively (Fig. 2, top). Thus the single ESI/FTMS spectrum of Figure 1 provides extensive confirmation of the DNA-predicted sequence.

ThiH

This enzyme is required for the biosynthesis of the thiazole moiety of thiamin (Vander Horn et al., 1993). Both published DNA sequences give a predicted M_r value of 43,320 Da (377 AA), with the 21AA10XHis linker (of amino acid sequence MGHHHHHHHHHHSS GHIEGRH) and the start methionine attached, M_r = 45,841 Da. The poorly resolved spectrum and satellite peaks (Fig. 3A) apparently result from a large number of unknown adducts (or impurities); Ni^{2+} adducts (+59 Da) are possible from using a Ni^{2+} affinity column during the purification. The measured M_r value of 45,709[1] Da for the most abundant low-resolution peak was 132 Da less (ΔM_r = −132) than that predicted from the DNA sequence. Collisionally activated dissociation of ThiH multiply-charged ions generated 53 fragment ions (Fig. 3B); none exhibit satellite peaks like those of the molecular ions, which is consistent with the assignment of this mass heterogeneity to non-covalent adducts that are preferentially lost in fragmentation. Of the fragment ions, 26 as large as y_{199} are assignable as *y*-type ions, verifying the predicted 199 C-terminal residues. Although none of the other fragment ions

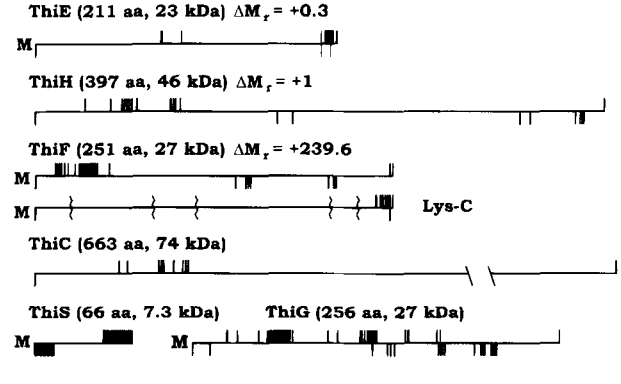

Fig. 2. Ion fragmentation data for ThiE, ThiH, ThiF, ThiC, ThiS, and ThiG. ΔM_r, difference between predicted and measured M_r values; *M*, indicates presence of deformylated methionine at N-terminus; horizontal lines, sequences; vertical bars above and below lines, *b*- and *y*-type ions, respectively; jagged lines, Lys-C cleavages.

[1] Relative molecular weight (M_r) values are reported for the theoretically-predicted most abundant isotope peak (the difference in 1.0034 Da units from the monoisotopic peak is given in italics following each value). M_r values with no italicized number following indicate an average value.

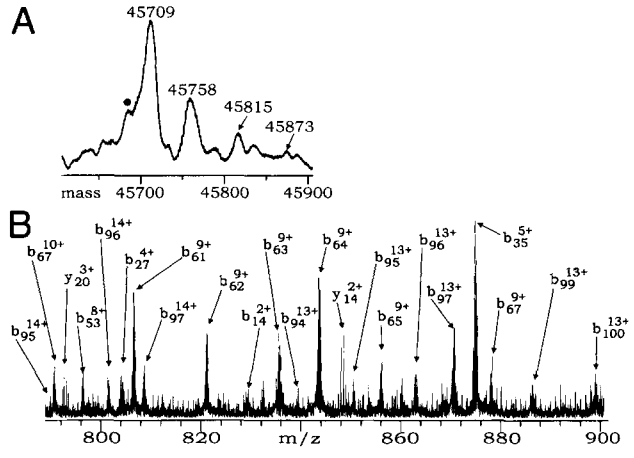

Fig. 3. A: Deconvoluted spectrum of the ThiH gene product, 10 scans; ●, H_2O loss. **B:** Partial collisional dissociation (NS) of ThiH (*m/z* 790–900), 10 scans.

could be assigned as *b*-type products, a string of eight of these ions yielded the partial sequence Asp-Met-Met-Ala-Leu/Ile-Leu/Ile-Ser. Searching the DNA-predicted sequence gave the corresponding Asp^{61}-Met^{62}-Met^{63}-Ala^{64}-Leu^{65}-Leu^{66}-Ser^{67}, consistent with the predicted b_{61}-b_{67} ions if 131 Da is subtracted from each observed mass value. Deleting the start Met residue (*m* = 131 Da) generates a sequence with M_r = 45,710 that predicts *b*-type ions consistent with 19 of the observed fragments (Fig. 3B). It is therefore reasonable to assume (Fig. 2) that there are no errors in the middle region (Tyr^{100}-Met^{198}) of the protein and that the entire ThiH sequence is consistent with the DNA-derived sequence after removal of the amino terminal methionine.

ThiF

This gene product catalyzes the adenylation of the C-terminal glycine of ThiS (Taylor et al., 1998). The two published DNA sequences disagree on the size of the ThiF protein due to differences in the assignment of the start codon but are otherwise identical, 250 AA (Vander Horn et al., 1993) and 245 AA (Blattner et al., 1993). Since the overexpression construct used in this study biases expression toward the 250 AA sequence, the correct start codon for the ThiF protein was not determined here. The ThiF gene product (predicted M_r = 26,730.6-*17*) gave M_r = 26,970.2-*17* (Fig. 4A), ΔM_r = +239.6. Also visible is a 7,310.7-*4* Da component and its +16 derivative, which will be discussed below. IR multiphoton dissociation of the ThiF ions (Fig. 4B) yielded 17 *b*-type ions (b_{16} to b_{52}) assignable to the predicted sequence. This verifies the first 52 amino acids of the sequence, demonstrates that the N-terminus contains the deformylated start methionine, and localizes the +239.6 ΔM_r to the last 199 residues.

By adding this 239.6 Da to the predicted fragment ion mass values, eight more ions were assignable as *y*-type and three as internal ions. The smallest is the $y_{36}b_{248}$ internal ion,[2] localizing ΔM_r to P^{215}-D^{248} and allowing a more accurate ΔM_r determination of 239.11.

[2]Both amide bonds cleaved to form this internal ion are N-terminal to proline (P^{248} and P^{215}), well known to be structurally favorable for inducing fragmentation (Loo et al., 1988; Little et al., 1994; McLafferty, 1994).

Supporting this, an endoproteinase Lys-C digest of ThiF yielded six peptides; five corresponded to the predicted Met^1-Lys^{25}, Leu^{26}-Lys^{83}, Ser^{84}-Lys^{113}, Asp^{114}-Lys^{207}, and Leu^{208}-Lys^{225}, with one 239.3 Da higher than that predicted for Ser^{226}-Val^{250}. Its fragmentation by collisions with N_2 gas ejected the predicted C-terminal PV residues (−214.13 Da); MS^3 of these ions localizes the ΔM_r to residues 226–242 (Fig. 2). Only then was it noted that a resequencing of the DNA (Blattner et al., 1993) had resulted in replacement of Cys^{242} with Arg and Trp, which results in ΔM_r = +239.17. This refined DNA-derived sequence now predicts M_r = 26,969.8-*17*, with 27 fragment ions consistent within 25 ppm of their predicted values.

ThiC

This gene product is required for the biosynthesis of the pyrimidine moiety of thiamin (Vander Horn et al., 1993). The two published DNA sequences, 627 AA (Vander Horn et al., 1993) and 631 AA (Blattner et al., 1993), show differences in the last 30 nucleotides. ThiC is the largest of the proteins in the *thiCEFSGH* operon with predicted M_r = 70,270 Da (wild-type, using the 627 AA predicted sequence (Vander Horn et al., 1993)). The fusion protein utilized in this study was predicted to be 664 AA, with M_r = 74,528 Da (627 AA sequence (Vander Horn et al., 1993) plus the 37 AA N-terminal histidine tag of sequence MRGSHHHHHHGMA SMTGGQQMGRDLYDDDDKDRWGSE). SDS-PAGE analysis of the partially purified protein (chromatography on a Ni^{2+} affinity column) indicated several minor contaminants and heterogeneity of ThiC possibly due to proteolysis. In the mass spectrum, molecular ions for the affinity-purified sample were poorly defined, with only broad regions of signal discernible (not shown). Combined collisional (NS, 110 V) and IR multiphoton dissociation (70 ms) mass spectra yielded 14 fragment ions (Fig. 5); if the start Met is removed, nine of these are consistent with b_{52}-b_{107}, confirming the N-terminal 107 residues, as well as the removal of the N-terminal Met. The five remaining fragment ions were not identified; none was consistent with a predicted *y*-ion. C-terminal heterogeneity

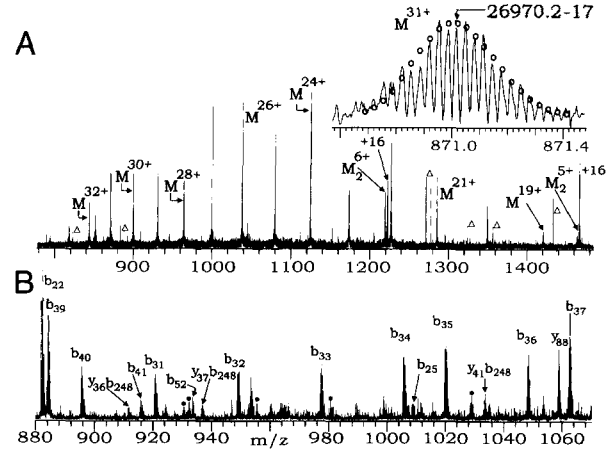

Fig. 4. A: Broadband molecular ion spectrum of ThiF, 40 scans; M_2^{n+}, molecular ions for a 7,310.7-*4* protein that co-purified with ThiF; Δ, noise spike. Inset, expansion of 31$^+$ ions; O, theoretical isotopic abundance distribution. **B:** Partial IR multiphoton dissociation spectrum of ThiF molecular ions in (A); ●, H_2O– or NH_3– loss peak.

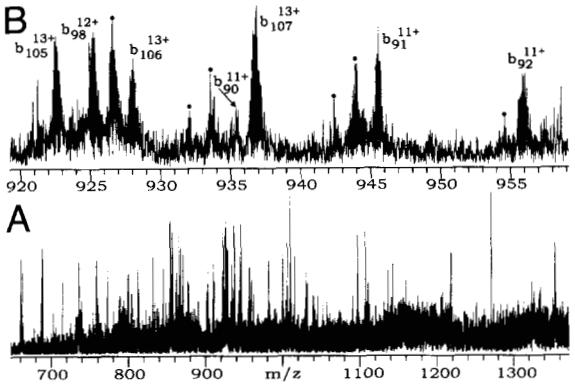

Fig. 5. A: Collisional dissociation (NS) of ThiC ions. **B:** Expansion of m/z 920–960; ●, H_2O– or NH_3– loss peak.

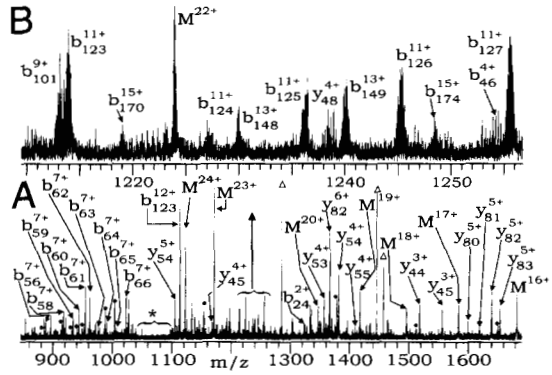

Fig. 7. A: Collisional dissociation (SORI) fragmentation of $(ThiG)^{25+}$ ions, 10 scans; M^{n+}, molecular ions: ●, H_2O– or NH_3– loss peak; *, SORI blindspot; Δ, noise spike. **B:** Expansion of m/z 1209–1257.

would explain the inability to observe ions representing either the intact protein or C-terminal fragments, and is consistent with the four residue difference in the DNA-predicted sequences. However, the largest protein for which this 6T FTMS instrument has produced significant isotopically-resolved molecular ions is porcine albumin, which is 67 kDa (Speir et al., 1995). Isotopically-resolved molecular ions have been obtained for 112 kDa proteins using a 9.4T FTMS instrument (Kelleher et al., 1997).

ThiG

This protein is required for the biosynthesis of the thiazole moiety of thiamin (Vander Horn et al., 1993). A large discrepancy was observed between the two DNA-predicted sequences for ThiG, 324 AA, M_r = 34 kDa (Vander Horn et al., 1993) and 281 AA, M_r = 30 kDa (Blattner et al., 1993) and the value estimated using SDS-PAGE analysis (24 kDa). The mass spectrum of the purified protein shows two components (Fig. 6), 7,310.74-*4* (to which we have assigned the name ThiS) and 26,896.5-*17* (which retains the name ThiG), that sum to within 202 Da of the DNA-predicted value using the Vander Horn DNA sequence (Vander Horn et al., 1993). These components co-purify and probably associate noncovalently. ThiS would also account for the presence of the 7,310 Da component in the ThiF sample (Fig. 4). Collisional dissociation of the $(M + 25H)^{25+}$ ThiG ions

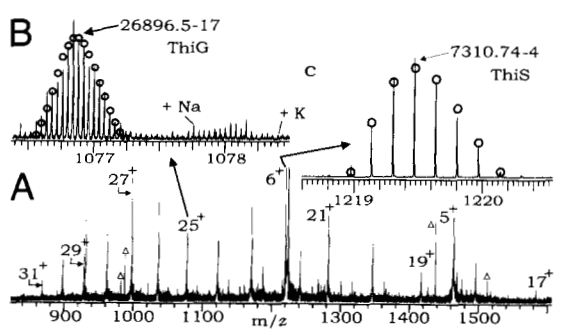

Fig. 6. A: Broadband spectrum (single scan) of the ThiS and ThiG gene products. **B:** Expansion of $(ThiG)^{25+}$. **C:** Heterodyne spectrum of $(ThiS)^{6+}$ (resolving power ~250,000). O, theoretical isotopic abundance distributions; Δ, noise spike.

yielded 108 isotopic distributions of 51 mass values, with 27 *b*- and 13 *y*-type ions identified (Fig. 7). The *y*-ions are consistent with the predicted C-terminus at Ala^{324}, and the *b*-ions position the N-terminus at Met^{69} (Fig. 8), with a predicted M_r = 26,896.1-*17* (ΔM_r = +0.4). A different protein sample showed a component of 25,512.9-*16* Da of ~50% relative abundance (data not shown), a 1,382.6 Da loss consistent with removal of the 14 C-terminal residues (predicted 1,382.7 Da); this was confirmed by MS/MS fragmentation (not shown). The use of protease inhibitors during the purification prevented this C-terminal truncation.

ThiS

The M_r value of the protein accompanying ThiG is not consistent with the remaining DNA-derived sequence (Fig. 8). However, MS/MS of the ThiS ions (Fig. 9) generated two terminal fragment ladders (Fig. 2, schematic on bottom) consistent with the predicted 12 N-terminal residues with the deformylated start methionine still attached (Fig. 8). The other ladder of *y*-ions indicated 15 C-terminal residues ending in Gly-Gly, but to assign these correctly, a stop codon must be present on the N-terminal side of the start codon of the ThiG protein (previously Met^{69} (Fig. 8)). For ThiS, a TGA stop codon can be formed at this site by a reading frame shift of −1 (adenosine deletion), that then predicts all the 15 C-terminal residues determined by the b_{51}-b_{65} fragments from MS/MS.[3] The re-sequenced DNA confirmed this adenosine deletion, detected two other extra bases in GC rich areas (Fig. 8), and predicted two gene products for the originally assigned ThiG. The new ThiS is predicted as a 66 residue protein of 7,310.70-*4* Da (Fig. 2); the sequences also match 68 *b*- and *y*-type fragment ions from collisional and IR multiphoton dissociation of the two subunits.

Conclusions

Of the six proteins involved in thiamin biosynthesis examined here, "top down" ESI/FTMS methodology on partially purified proteins (>50%) efficiently identified and localized errors in four of the sequences, confirmed the correct sequence of the other two genes, and determined if post-translational modification of the

[3]GC compression could have caused two of the three DNA sequencing errors for the ThiG gene.

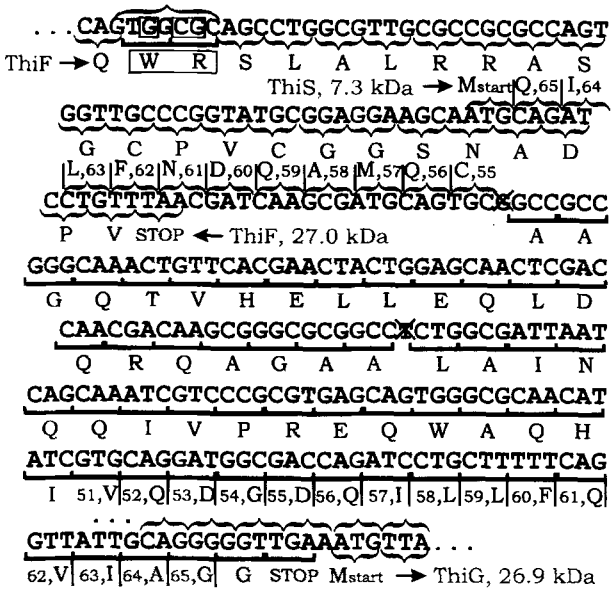

Fig. 8. DNA sequences (Blattner et al., 1993; Vander Horn et al., 1993) with amino acid sequence predicted and revised from mass spectral data. —, —, original open reading frame; ⌴, revised open reading frame; $\overset{}{\underset{n,x}{\lfloor\rfloor}}$, N-terminal (*b*-type) fragment ion containing *n* residues with amino acid *x* at internal terminus; $\overset{x,n}{\rfloor\lfloor}$, C-terminal (*y*-type) fragment ion containing *n* residues with amino acid *x* at internal terminus. Single letter residues in rectangle were added on resequencing to change prediction from Cys to Trp-Arg.

amino terminus of each protein had occurred. The MS/MS-based sequence correction of the previously assigned ThiG protein resulted in the discovery of a new gene product (ThiS) with a critical role in the sulfur transfer chemistry involved in the biosynthesis of the thiazole moiety of thiamin. ThiS was also found to noncovalently associate with ThiF and ThiG.

Measurement of the mass spectra described in this paper required less than 30 min each, utilizing 10^{-11} to 10^{-13} mol samples. This was a small fraction of the amount available; measuring 10^{-17} mol samples is possible with unusual care and skill (Valaskovic et al., 1996). Sequence error detection and localization was a straightforward process of following the molecular mass error (ΔM_r) from larger to smaller fragments; here these were mainly from ion dissociation, but peptide fragments from proteolysis provide complementary information. Note that an ESI/FTMS high-resolution spectrum can provide hundreds of fragment mass values of uniquely high accuracy. With this methodology, sequence confirmation and error localization of the gene products of entire operons can be accurately and efficiently accomplished.

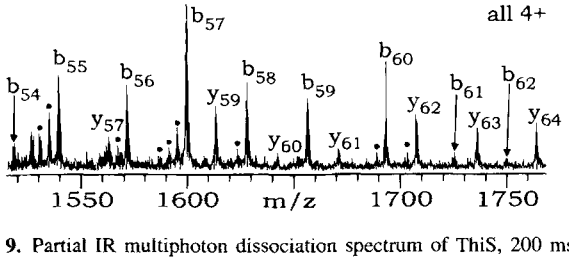

Fig. 9. Partial IR multiphoton dissociation spectrum of ThiS, 200 ms irradiation, 10 scans; ●, H_2O- or NH_3- loss peaks.

Materials and methods

Materials

HPLC grade MeOH and 99.999% HOAc were obtained from Sigma (St. Louis, Missouri). H_2O (18 MΩ) was obtained from a Millipore water purification system and stored in a Teflon bottle. Reversed-phase protein and peptide traps (1 × 10 mm and 3 × 10 mm) were obtained from Michrom BioResources, Inc. (Auburn, California). Nanospray tips (Wilm & Mann, 1994, 1996) were manufactured from borosilicate glass tubes (o.d. = 1.5 mm, i.d. = 0.86 mm) from Sutter Instrument Co. (Novato, California) on a Sutter P-87 puller, and PicoTips (Valaskovic et al., 1995) were obtained from New Objective (Boston, Massachusetts).

Strains, plasmids, and protein purification

All proteins used in this study were overexpressed in *E. coli* BL21(DE3). ThiE was overexpressed from plasmid pADB154 and purified as previously described (Backstrom, 1996). ThiH was overexpressed from pCLK402 with a 21 AA N-terminal fusion containing a 10X-histidine tag (C.L. Kinsland & T.P. Begley, unpubl. results). ThiH formed inclusion bodies and was purified as the denatured protein using the protocols described in the pET system manual (Novagen, 1996). ThiF was overexpressed from plasmid pCAC111 (Costello, 1996) and purified as described elsewhere (Taylor et al., 1998). ThiC was overexpressed from pThiC90, a pRSETA-derived vector that appended a 37 AA N-terminal leader containing a 6X-histidine tag (Vander Horn et al., 1993). It was purified as described in the pET system manual (Novagen, 1996). ThiG/ThiS was overexpressed from pADB151 and purified as described elsewhere (Taylor et al., 1998).

ESI/FTMS

Proteins (1–100 μg) were desalted by ultrafiltration (10 kDa MWCO, Amicon) or by step elution from reversed-phase traps (Michrom BioResources) with 70:28:2 (MeOH/H_2O/HOAc; ~1–10 μM solution); 2–5 μL was loaded into either a Nanospray or PicoTip emitter with a 1–4 μm tip for low flow rate (1–50 nL/min) ESI; typically, M^{n+} and MS/MS data were obtained using one 10^{-11} to 10^{-13} mol sample loading. MS/MS sequence data have been obtained on a 10^{-17} mol sample using ESI/FTMS (Valaskovic et al., 1996), but the sample handling procedures are far more difficult. The ESI spray was sampled by a heated metal capillary (110°C) and guided through five stages of differential pumping by dc and rf ion guides into the trapped ion cell (~10^{-9} Torr) of a modified 6T Finnigan FTMS (Beu et al., 1993). Data were stored as 64–512K sets and analyzed with Odyssey v4.0 software. Geometric deconvolutions (Mann et al., 1989) were performed in PV-WAVE v4.01. Theoretical isotopic distributions were generated using Isopro v2.0 and fit to those observed by least squares (Senko et al., 1995a). Spectra were calibrated externally using bovine ubiquitin (8,564.64-5) and analyzed with algorithms written in PV-WAVE (Senko et al., 1995b). Fragmentation was achieved by nozzle-skimmer (NS) dissociation (Loo et al., 1988) (110–230 V) for ions entering the instrument or by isolating specific ions using stored-waveform inverse Fourier-transform (SWIFT) (Marshall et al., 1985), followed by infrared (IR) multiphoton dissociation (Little et al., 1994) for 80–300 ms or collisional activation using sustained off-resonance irradiation (SORI) (1.2–1.4 kHz off-resonance

for 1.5 s during a pulsed N_2 pressure of $\sim 3 \times 10^{-6}$ Torr) (Gauthier et al., 1991).

Acknowledgments

The authors thank Allyson Backstrom, Colleen Costello, and Peter Vander Horn for overexpression plasmids. We also thank the National Institutes of Health (F.W.M., GM16609; T.P.B., DK44083; S.V.T. and N.L.K., BMCB Training Grant), the National Science Foundation (C.L.K., fellowship), and the American Chemical Society (N.L.K., Division of Analytical Chemistry full fellowship sponsored by Perkin-Elmer Corporation) for financial support.

References

Backstrom AD. 1996. Biosynthesis of thiamin in *E. coli* [Ph.D. thesis]. Ithaca, New York: Cornell University, Department of Chemistry.

Backstrom AD, McMordie RAS, Begley TP. 1995. Biosynthesis of thiamin I: The function of the ThiE gene product. *J Am Chem Soc 117*:2351–2352.

Beu SC, Senko MW, Quinn JP, Wampler FM III, McLafferty FW. 1993. Fourier-transform electrospray instrumentation for tandem high-resolution mass spectrometry of large molecules. *J Am Soc Mass Spectrom 4*:557–565.

Blackburn RK, Anderegg RJ. 1997. Characterization of femtomole levels of proteins in solution using rapid proteolysis and nanoelectrospray ionization mass spectrometry. *J Am Soc Mass Spectrom 8*:483–494.

Blattner FR, Burland V, Plunkett G, Sofia HJ, Daniels DL. 1993. Analysis of the *Escherichia coli* genome. IV. DNA sequence of the region from 89.2 to 92.8 minutes. *Nucleic Acids Res 21*:5408–5417.

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science 277*:1453–1462.

Comisarow MB, Marshall AG. 1974. Fourier-transform ion cyclotron resonance spectroscopy. *Chem Phys Lett 25*:282–293.

Costello CA. 1996. Mechanistic studies on thiamin biosynthesis in *Escherichia coli* and thiamin degradation in *Bacillus thiaminolyticus* [Ph.D. thesis]. Ithaca, New York: Cornell University, Department of Chemistry.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1990. Electrospray ionization-principles and practice. *Mass Spectrom Rev 9*:37–70.

Gauthier JW, Trautman TR, Jacobsen DB. 1991. Sustained off-resonance irradiation for CAD involving FTMS. CAD technique that emulates infrared multiphoton dissociation. *Analytica Chimica Acta 246*:211–225.

Kelleher NL, Costello CA, Begley TP, McLafferty FW. 1995. Thiaminase I (42 kDa) heterogeneity, sequence refinement, and active site location from high-resolution tandem mass spectrometry. *J Am Soc Mass Spectrom 6*:981–984.

Kelleher NL, Senko MW, Siegel MW, McLafferty FW. 1997. Unit resolution mass spectra of 112 kDa molecules with 3 Da accuracy. *J Am Soc Mass Spectrom 8*:380–383.

Krawetz SA. 1989. Sequence errors in Genbank: A means to determine the accuracy of DNA sequence interpretation. *Nucleic Acids Res 17*:3951–3958.

Little DP, Speir JP, Senko MW, O'Connor PB, McLafferty FW. 1994. Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. *Anal Chem 66*:2809–2815.

Loo JA, Udseth HR, Smith RD. 1988. Collisional effects on the charge state distribution of ions from large molecules formed by electrospray-ionization mass spectrometry. *Rapid Commun Mass Spectrom 2*:207–210.

Mann M, Meng CK, Fenn JB. 1989. Interpreting mass spectra of multiply charged ions. *Anal Chem 61*:1702–1708.

Marshall AG, Grosshans PB. 1991. Fourier-transform ion cyclotron resonance mass spectrometry: The teenage years. *Anal Chem 63*:215A–229A.

Marshall AG, Wang TCL, Ricca TL. 1985. Tailored excitation for Fourier-transform ion cyclotron resonance mass spectrometry. *J Am Chem Soc 107*:7893–7897.

McLafferty FW. 1994. High resolution tandem FT mass spectrometry above 10 kDa. *Acc Chem Res 27*:379–386.

Mewes HW, George DG. 1994. Protein sequences and sequence databases. In: Kellner R, Lottspeich F, Meyer HE, eds. *Microcharacterization of proteins*. Weinheim, Germany: Verlagsgesellschaft mbH. p 220.

Novagen. 1996. pET System Manual. Madison, Wisconsin: Novagen.

Roepstorff P, Fohlman J. 1984. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom 11*:601.

Senko MW, Beu SC, McLafferty FW. 1995a. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J Am Soc Mass Spectrom 6*:229–233.

Senko MW, Beu SC, McLafferty FW. 1995b. Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *J Am Soc Mass Spectrom 6*:52–56.

Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortensen P, Shevchenko AN, Boucherie H, Mann M. 1996. Linking genome and proteome by mass spectrometry: Large scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci USA 93*:14440–14445.

Speir JP, Senko MW, Little DP, Loo JA, McLafferty FW. 1995. High-resolution tandem mass spectra of 37–67 kDa proteins. *J Mass Spectrom 30*:39–42.

Taylor SV, Kelleher NL, Kinsland CL, Chiu H-J, Costello CA, Backstrom AD, McLafferty FW, Begley TP. 1998. Thiamin biosynthesis in *Escherichia coli*: Identification of ThiS thiocarboxylate as the immediate sulfur donor in the thiazole formation. *J Biol Chem 273*:16555–16560.

Valaskovic GA, Kelleher NL, Little DP, Aaserud DJ, McLafferty FW. 1995. Attomole sensitivity electrospray source for large molecule mass spectrometry. *Anal Chem 67*:3802–3805.

Valaskovic GA, Kelleher NL, McLafferty FW. 1996. Attomole protein characterization by capillary electrophoresis-mass spectrometry. *Science 273*:1199–1202.

Vander Horn PB, Backstrom AD, Stewart V, Begley TP. 1993. Structural genes for thiamine biosynthetic enzymes (*thiCEFGH*) in *Escherichia coli* K-12. *J Bact 175*:982–992.

Williams ER. 1998. Tandem FTMS of large biomolecules. *Anal Chem 70*:179A–185A.

Wilm M, Mann M. 1994. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *Int J Mass Spectrom Ion Proc 136*:167–180.

Wilm M, Mann M. 1996. Analytical properties of the nanoelectrospray ion source. *Anal Chem 68*:1–8.

Wood TD, Chen LH, White CB, Babbitt PC, Kenyon GL, McLafferty FW. 1995. Sequence verification of human creatine kinase (43 kDa) isozymes by high-resolution tandem mass spectrometry. *Proc Natl Acad Sci USA 92*:11451–11455.

Yates JR III. 1998. Mass spectrometry and the age of the proteome. *J Mass Spectrom 33*:1–19.