Electrospray Mass Spectrometric Analysis of the Domains of a Large Enzyme: Observation of the Occupied Cobalamin-Binding Domain and Redefinition of the Carboxyl Terminus of Methionine Synthase[†,‡]

James T. Drummond,[§] Rachel R. Ogorzalek Loo,[∥] and Rowena G. Matthews*[,§]

Biophysics Research Division and Department of Biological Chemistry, and Protein and Carbohydrate Structure Facility, The University of Michigan, Ann Arbor, Michigan 48109

Received March 15, 1993; Revised Manuscript Received May 28, 1993[⊗]

ABSTRACT: Cobalamin-dependent methionine synthase from *Escherichia coli* catalyzes the methylation of homocysteine to form methionine, using methyltetrahydrofolate as the primary methyl donor. We have used electrospray mass spectrometry as a powerful tool for characterizing separable fragments obtained by proteolysis of this monomeric 136.1-kDa enzyme. A central 28.0-kDa domain, reported to bind the cobalamin, has been purified to homogeneity in 30% yield. We were able to detect the domain with bound cobalamin by electrospray mass spectrometry at neutral pH. Mass analysis of a 37.2-kDa carboxyl-terminal domain was grossly inconsistent with either of the two amino acid sequences from previously published DNA sequences. We then used electrospray mass spectrometry to analyze peptides generated by a lysyl endoproteolytic digest of a C-terminal fragment, and we have constructed a peptide map that accounts for >95% of the peptide mass derived from this domain. The correct translational end of this protein (27 residues downstream from the previously predicted ultimate residue) has been established, and sequence conflicts within the two published DNA sequences have been resolved (GenBank Accession Number J04975). Resequencing the DNA near the carboxyl terminus ruled out a frameshifted reading of the DNA and suggested that a cytosine had twice been incorrectly inserted late in the reading frame. The strategies reported here for sequence confirmation, localization of coenzyme-binding regions, and identification of chemically modified peptides within a large protein are potentially applicable to the characterization of many other proteins.

Electrospray mass spectrometry is emerging as a powerful analytical technique for the analysis of intact proteins and peptides (Fenn et al., 1989; Smith et al., 1990; Chait & Kent, 1992). Among the applications are the direct confirmation of molecular mass, characterization of posttranslational modification, and the determination of amino acid sequence. Inherent in this approach is the ability to verify that recombinant protein products are consistent with the primary amino acid sequences deduced from sequenced DNA (Van Dorsselaer et al., 1990). In addition to the global confirmation of protein mass, considerably more information can be obtained by digesting proteins with specific endoproteinases (e.g., trypsin or thermolysin) to generate a family of peptides that may again be characterized by electrospray mass spectrometry (Ling et al., 1991; Edmonds et al., 1991; Covey et al., 1991). This allows for the localization of modification sites, and specific examples include the identification of pyruvic acid imine formation with hemoglobin (Prome et al., 1991) and the phosphorylation and acetylation of spinach light-harvesting chlorophyll protein II (Michel et al., 1991). Complete resolution of the peptides is not essential; in fact, identification of virtually all 37 peptides generated by tryptic digestion of human apolipoprotein AI, a 28-kDa[1] protein, have been identified as a mixture in a single mass spectral analysis (Chowdhury et al., 1990).

Cobalamin-dependent methionine synthase from *Escherichia coli* (MetH) is an exceptionally large monomeric enzyme of 136.1 kDa, and because of its large size, this enzyme would be difficult to analyze on a low-resolution, limited m/z (mass/charge) quadrupole mass spectrometer. However, Banerjee et al. (1989) have reported that the native enzyme can be digested with catalytic amounts of trypsin to generate relatively stable fragments, and we have optimized conditions both for limited digestion of this enzyme and for purification of defined fragments to homogeneity in good yields. We have proposed (Drummond et al., 1993) that the enzyme can be viewed as a structural mosaic of domains that each possess specific binding and catalytic properties, and that limited tryptic proteolysis effectively separates individual activities that contribute to the complex turnover scheme. The physical analysis reported here provides part of the structural framework for the assignment of activity to specific protein regions.

Our purpose in this work is to summarize the practical strategies we used to characterize the 136.1-kDa cobalamin-dependent methionine synthase using electrospray mass spectrometry. The primary requirement is the ability to

[†] This research has been supported in part by Research Grant R37 GM24908 (R.G.M.) and Pharmacological Sciences Training Grant T32 GM07767 (J.T.D.) from the National Institute of General Medical Sciences, National Institutes of Health. Additional support was provided by the National Cancer Institute Grant P30 CA46952 to the University of Michigan Comprehensive Cancer Center. J.T.D. has also been supported by an NSF Graduate Fellowship.

[‡] The nucleic acid sequence reported in this paper has been submitted to GenBank under Accession Number J04975.

[§] Biophysics Research Division and Department of Biological Chemistry.

[∥] Protein and Carbohydrate Structure Facility.

[⊗] Abstract published in *Advance ACS Abstracts*, August 15, 1993.

[1] Abbreviations: (k)Da, (kilo)dalton; [α-^{35}S]dATP, deoxyadenosine 5′-[α-^{35}S]thiotriphosphate; FPLC, fast protein liquid chromatography; HPLC, high-pressure liquid chromatography; LysC, lysyl endopeptidase C from *Achromobacter* (protease I); MetH, cobalamin-dependent methionine synthase from *Escherichia coli*; TLCK, N^{α}-(*p*-tosyl)-L-lysine chloromethyl ketonehydrochloride; TPCK, *N*-tosyl-L-phenylalaninechloromethyl ketone.

0006-2960/93/0432-9282$04.00/0

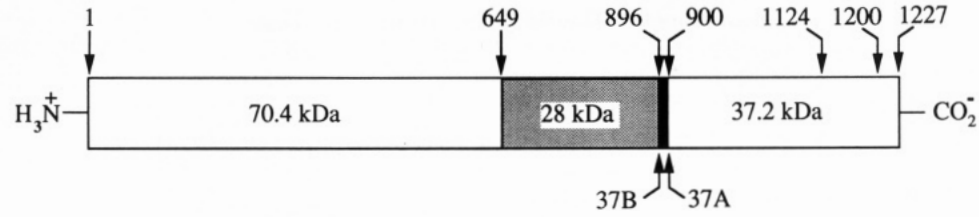

FIGURE 1: Location of domains in the primary sequence of cobalamin-dependent methionine synthase. The overlapping peptides 37A and 37B both extend to the carboxyl terminus, and these fragments bind *S*-adenosylmethionine. The 28-kDa peptide denoted by the gray box binds the cobalamin cofactor, and it is obtained following extensive proteolysis of the enzyme. Methionine synthase is a monomer of 1227 amino acids, and the C-terminus has now been rigorously determined. The carboxyl terminal translational stop sites predicted by Banerjee et al. (1989) and Old et al. (1990) are denoted by the arrows at residues 1124 and 1200, respectively.

subdivide the protein into smaller units, and we have isolated a domain reported to bind the cobalamin prosthetic group (Banerjee et al., 1989) and a domain that binds the essential *S*-adenosylmethionine cofactor (Drummond et al., 1993). The implicit benefits are that analysis of the domains can yield much more accurate information about their sequence and function and that this collected information can be assembled to characterize the intact enzyme. In combination with amino-terminal analyses to establish one end of polypeptides, the masses of domains (and peptides derived from these domains) can be used to deduce the corresponding carboxyl termini and definitively position the domains within a primary amino acid sequence (Figure 1). This task is often difficult using traditional chemical methodologies, since C-terminal determination for polypeptides is not as well developed as N-terminal sequencing (Inglis, 1991).

This work demonstrates detection of a noncovalently bound prosthetic group, complexed with the protein, observed under specific conditions in the mass spectrometer. Recently, Katta and Chait (1991) have shown that equine myoglobin can be observed by mass spectrometry with and without the bound heme prosthetic group, and the heme occupancy was dependent upon the protein existing in a native conformation in solution. Additionally, proteins complexed with substrates, inhibitors, or peptides have recently been observed by mass spectrometry (Baca & Kent, 1992; Ganem et al., 1991a,b; Ogorzalek Loo et al., 1993), and these results suggest that the specific interactions required for noncovalent binding may be preserved under the relatively mild conditions required for electrospray mass spectrometry. We report that a domain of 28.0 kDa, derived from methionine synthase, was characterized with bound cobalamin by electrospray mass spectrometry. The ability to characterize such a product implies that highly useful information may be derived from proteins that are too large to be directly evaluated by mass, assuming that fragments derived from them retain the substrate or coenzyme binding determinants of the intact protein. Indeed, a C-terminally truncated human H-ras protein (amino acids 1–166) possessing noncovalently bound guanosine 5′-diphosphate (GDP) has been observed in this manner (Ganguly et al., 1992).

During the course of this work, we were able to evaluate the carboxyl terminus of methionine synthase by electrospray mass spectrometry. The mass of a C-terminal 37.2-kDa domain was not consistent with either of the reported protein sequences deduced from DNA sequences (Banerjee et al., 1989; Old et al., 1990). In order to resolve this dilemma, we applied the same general strategy to this domain that we applied to the intact enzyme: When a polypeptide is so large as to conceal desired information, divide it with a specific protease and characterize the products by their mass. Not only was the correct reading frame for methionine synthase at the C-terminus identified, but some of the conflicts in nucleotide assignments between the two reported DNA sequences were resolved. We propose that the technique of peptide mapping by mass (see references above) can be extended to the characterization of chemical modification of large proteins, and it is our intent to report the modifications to methionine synthase that occur following oxidative inactivation by nitrous oxide that we have identified with this technique (J. T. Drummond and R. G. Matthews, manuscript in preparation).

EXPERIMENTAL PROCEDURES

Materials. Bovine serum albumin, trypsin (TPCK treated), TLCK, and other reagents were purchased from the Sigma Chemical Co. [α-^{35}S]-dATP (400 Ci/mmol) was supplied by Amersham, and LysC was purchased from Waco Bio-Products.

Prediction of Peptide Masses and Isoelectric Points. Algorithms contained in the MacVector sequence analysis software package (International Biotechnologies, Inc., of the Kodak Company) were used to calculate the masses and isoelectric points of the enzyme, domains, and peptides. The isoelectric point determination in this package was described by Wood et al. (1974).

Purification of Methionine Synthase and Its Domains. Recombinant methionine synthase (MetH) from *E. coli* K-12 strain DH5αF/p4B6.3 was overproduced and purified as previously reported (Banerjee et al., 1989). Generation of the 28.0-kDa cobalamin-binding domain was accomplished by incubating the enzyme (1.0 mg, 7.3 nmol) in 2.0 mL of Tris·HCl buffer (pH 7.2, 1 mM EDTA) with trypsin (50 μg, 5% w/w) for 45 min at ambient temperature. The reaction was quenched by the addition of bovine pancreatic trypsin inhibitor (65 μL of a 2 mg/mL solution in the same buffer), and the solution was concentrated in a Centricon 10 micro-concentrator (Amicon) at 4 °C. The buffer was replaced with 25 mM potassium phosphate buffer, pH 7.2, at 4 °C, and the mixture was loaded onto a MonoQ HR 5/5 anion-exchange FPLC column (Pharmacia LKB Biotechnology) equilibrated in 25 mM potassium phosphate buffer. At a flow rate of 1 mL/min, the column was washed for 20 min with 25 mM phosphate buffer. The fragments were then eluted with a linear gradient from 25 to 260 mM phosphate buffer over a period of 25 min. The purified domain, readily identified by its red color, was concentrated in a Centricon 10 micro-concentrator and stored in 50 mM phosphate buffer. The domain was recovered in a 30% yield (2.2 nmol from 7.3 nmol of intact protein). For mass spectrometric analysis, the storage buffer was replaced with water by ultrafiltration in a Centricon 10. The protein solution was diluted to 2.0 mL with water at 4 °C and reconcentrated, and the process was repeated through three cycles to reduce the buffer concentration below 1 mM. All protein concentrations were determined using the BioRad protein assay, based on the method of Bradford (1976), using bovine serum albumin as a standard.

Generation of the 37.2-kDa C-Terminal Domain. Purified methionine synthase (4 mg, 29 mmol) was incubated with trypsin (4 μg, or 0.1% w/w) in 750 μL of 50 mM potassium phosphate buffer at pH 7.2 for 30 min at room temperature. The digest was quenched by the addition of TLCK (7 μL of an 11 mM stock in water), and after 10 minutes the mixture was loaded onto a MonoQ HR 5/5 anion-exchange FPLC column (Pharmacia LKB Biotechnology). At a flow rate of 1 mL/min, the column was washed for 20 min with 50 mM potassium phosphate buffer at pH 7.2. Two overlapping fragments of ~37 kDa, designated 37A and 37B (see Figure 2) were then eluted with a linear gradient from 50 to 275 mM potassium phosphate buffer (pH 7.2) over a period of 25 min. The purified 37-kDa domains (750 μg of a 1.5:1 mixture of 37B and 37A; 68% overall yield) were concentrated at 4 °C in Centricon 10 microconcentrators and stored in 50 mM phosphate buffer. The homogeneous 98.4-kDa domain with bound cobalamin eluted after the 37.2-kDa domain, and it was concentrated and stored in a similar manner.

LysC Digestion of the C-Terminal 37.2-kDa Domain. The protein to be digested (90 μL, containing 11 nmol of the 37.2-kDa C-terminal domain) was diluted 1:1 with 9 M urea in 100 mM Tris buffer, pH 8.0, and incubated with LysC (0.55 nmol) at 37 °C overnight. Peptides from LysC digests were separated on a Rainin HPLC instrument, Model HPXL, using the Dynamax software package from Rainin to program the gradients. Peptides were detected at 214 nm with an absorbance detector (Model 759A) from Applied Biosystems. Separation was achieved on a 0.41 × 25-cm C_{18} reverse-phase column (Vydac) at a flow rate of 0.7 mL/min. A two-solvent system was used, where solvent A was 0.1% trifluoroacetic acid in water and solvent B was 0.1% trifluoroacetic acid in a 75:25 (v/v) mixture of acetonitrile and water. After the column was equilibrated with 5% B, 90 μL of the digest was injected, and a gradient was run as follows: 0–10 min, 5% B; 10–20 min, 5–30% B; and 20–85 min, 30–85% B; the bulk of the peptides eluted over the last range. Individual fractions (100–500 μL) were lyophilized to dryness *in vacuo* on a SpeedVac concentrator (Savant Instruments), and the dried peptides were redissolved in 50 μL of glass-distilled water for mass spectral analysis. Specifically, 200 μg (5.4 nmol) of the digested 37.2-kDa domain was loaded onto the column; theoretically, 2.8–36.6 μg of each peptide was present. This allowed for multiple mass spectral analyses and complete amino acid sequencing of each peptide where required.

Amino Acid Sequence Analysis. N-terminal sequence analysis of the HPLC-purified peptides and tryptically generated domains was performed in a Model 473 Applied Biosystems liquid-phase sequencer at the University of Michigan Protein and Carbohydrate Core Facility.

Mass Spectrometry. Electrospray ionization mass spectra were obtained using a Vestec electrospray source and a Model 201 single-quadrupole mass spectrometer (Vestec Corp., Houston, TX) fitted with a 2000-m/z range (Allen & Vestal, 1992; Andrews et al., 1992). For routine analysis, aqueous samples were diluted to 4% acetic acid/50% acetonitrile and delivered to the source from a 10-μL injection loop at a flow rate of 5 μL/min. The electrospray interface was heated to 55 °C, and an 18-V repeller voltage was employed (Allen & Vestal, 1992). The 28.0-kDa domain with bound cobalamin was examined by spraying a 0.1 mg/mL solution of the complex in distilled water at an interface temperature of 30 °C, while the apodomain mass analysis was performed by spraying a 0.1 mg/mL solution in 4% acetic acid/50% acetonitrile at an interface temperature of 50 °C.

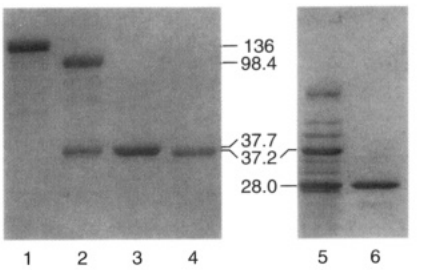

FIGURE 2: Isolable domains from methionine synthase separated by polyacrylamide gel electrophoresis in the presence of sodium dodecyl sulfate. Lane 1, intact methionine synthase; lane 2, enzyme digested with 0.1% (w/w) trypsin as described in the Experimental Procedures section; lane 3, purified 37.7-kDa C-terminal domain (37B); lane 4, purified 37.2-kDa C-terminal domain analyzed in this work (37A); lane 5, enzyme digested with 5% (w/w) trypsin; lane 6, purified cobalamin-binding domain.

DNA Sequencing. Double-stranded DNA sequence determination was performed using [α-^{35}S]-dATP and the dideoxynucleoside termination method described in the Sequenase kit (version 2.0) from United States Biochemical. Plasmid purification was carried out by anion-exchange chromatography as recommended by QIAGEN (QIAGEN, Inc.) on lysates of *E. coli* K-12 strain DH5αF′/p4B6.3. Two 17-mer sequencing primers, 5′-AATCTTTCGCCATGTGG-3′ and 5′-CAGATTCGGTGCCAGCC-3′, were synthesized by the DNA Synthesis Core Facility at the University of Michigan to be complementary to the nucleotide sequence of Old *et al.* (1990) over the regions 3722–3738 and 3872–3888, respectively. These primers were chosen to optimize nucleotide assignment in the region where the peptide sequence diverged from that predicted by the DNA sequence.

Sequence Numbering Conversion. Numbering of the published DNA sequences for MetH begins at the translational start site. The scheme used here is consistent with that of Old et al., except that the reading frame is shifted following removal of cytosine 3584 (see the Results Section).

RESULTS

Characterization of the Cobalamin-Binding Domain. Inherent to the intact protein and the 98-kDa domain is the ability to bind the cobalamin prosthetic group, but two smaller, overlapping fragments of ~28 kDa from the enzyme that retain the cobalamin have also been reported. Luschinsky et al. (1992) have crystallized a 28-kDa domain and are attempting to solve the structure by X-ray crystallography, and Banerjee et al. (1989) reported on the amino terminus of a closely overlapping domain generated by tryptic proteolysis. As shown in Figure 2, a relatively stable fragment of 28 kDa is a product of extensive tryptic digestion that we have purified to homogeneity in a single step by anion-exchange chromatography. The ease of purification of this fragment stems from the fact that it binds less tightly to the column than any other fragment, consistent with the calculation that this protein region is more basic (pI = 5.1) than either the intact enzyme (pI = 4.8) or the 37.2-kDa C-terminal domain (pI = 4.7).

When the domain of ~28 kDa was first subjected to electrospray mass spectrometry, a mixed spectrum reflecting masses of 28 002 ± 20 Da and 29 332 ± 15 Da was observed (Figure 3). These values were calculated from the characteristic family of charged states seen in the electrospray mass spectrum as described in the legend for Figure 3, and the domain masses represent a mean value plus or minus the standard deviation determined from the individual charged

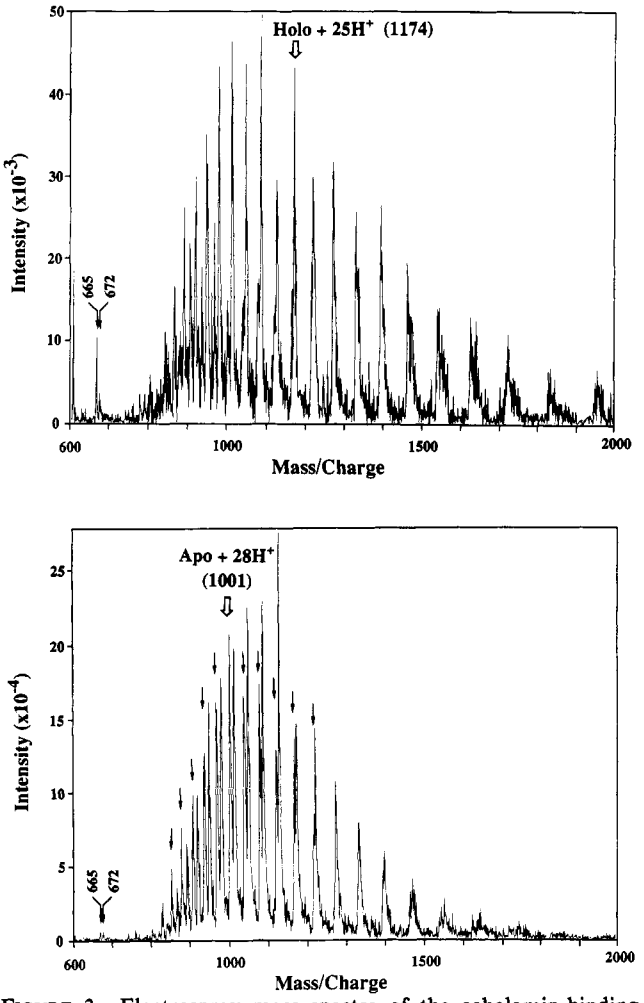

FIGURE 3: Electrospray mass spectra of the cobalamin-binding domain. Panel A (top) shows a spectrum of the holodomain injected under native conditions. The individual peaks reflect the domain mass (M = 29 332 Da) possessing integer cationic charges contributed by the protonation of basic residues in the domain. For example, the mass/charge (m/z) value for the domain containing 25 protons is predicted to be (29 332 + 25)/25 = 1174. Note that some complex dissociation to give free cobalamin and the apodomain is evident at the lower m/z values. Masses for doubly charged methylcobalamin (672 Da) and cobalamin lacking an upper axial ligand (665 Da) are also labeled. Panel B (bottom) shows a spectrum of the domain injected under acidic, denaturing conditions. The arrows denote members of the family of peaks derived from the apodomain, centered around the labeled peak (apodomain + 28 H^+). The remaining peaks are mainly derived from the holodomain and possess masses consistent with those in panel A.

states. The difference of 1330 ± 25 Da is very close to the mass of the cobalamin prosthetic group, and this result implies that these two species differ by the presence of noncovalently bound cobalamin. Taken together with the amino-terminal determination obtained for this fragment, the mass determined for the apodomain of 28 002 ± 20 Da is in good agreement with the prediction that this domain runs from threonine 643 through arginine 896, a hypersensitive tryptic site that initially separates the enzyme into fragments of 98.4 and 37.7 kDa (Figures 1 and 2). This fragment has a deduced mass of 27 992 Da derived from the DNA sequence of Old et al. (1990).

We found that by varying the conditions under which the sample was introduced into the mass spectrometer, more conclusive evidence for the assignment of masses could be obtained. Under mild conditions that allow the domain to retain native structure in solution, the larger mass representing the domain with bound prosthetic group was predominant (Figure 3A). Following harsher treatment (dilution to 50% acetonitrile and 4% acetic acid), the family of masses representing the apodomain appeared prominently in the spectrum, but the holodomain persisted in a slight excess (Figure 3B). Although the cobalamin is noncovalently bound to the enzyme and can be resolved by urea denaturation (Taylor, 1970), the cobalamin remains bound when the holoenzyme is precipitated under acidic conditions (Taylor & Weissbach, 1967). The structural basis for this result is not clear, but the inability to completely remove the prosthetic group under the acidic conditions commonly used for electrospray mass spectrometric analysis is not surprising. One possible explanation is that the domain is contributing a ligand to the cobalamin under acidic conditions, since the dimethylbenzimidazole is likely to be protonated (pK_{obs} = 2.7 for methylcobalamin (Pratt, 1982)) and hence no longer able to serve as a ligand for the cobalt.

Upon careful analysis, these spectra contain considerably more information than the individual masses for the domains. Because the apodomain was generated *in situ* from the holodomain, peaks of varying intensity for each of the cobalamin forms characteristic of the holodomain were detected in all the mass spectra recorded. Note that the peaks represent mass/charge (m/z) ratios, and cobalamins possessing two positive charges appear as m/z peaks of 665 for cob(II)alamin and 672 for methylcobalamin. The domain was initially purified with bound methylcobalamin, but the cobalt–methyl bonds is photolytically unstable and slowly degrades to cob(II)alamin on the domain, as it does on the holoenzyme (data not shown). Alkylcobalamins free in aqueous solution are considerably more sensitive to photolysis (Hogenkamp, 1982), and the loss of the methyl group from the cobalamin could also occur following release from the holodomain. It should also be noted that the overall distribution of masses was altered when the conditions were changed. As seen in Figure 3B, the center of the family of apodomain masses is shifted to lower mass/charge values (higher occupancy of basic residues by protons) following acidic treatment, consistent with an increased accessibility of additional basic residues to protonation. Finally, the apodomain is predicted to have a total of 33 potential cationic residues (including the N-terminus), and a mass/charge value of 850 for the fully protonated species is apparent and labeled in the spectrum.

Identification of the Carboxyl Terminus. When native methionine synthase was digested with catalytic amounts of trypsin, we observed a fragment of ~98 kDa and two similar fragments of ~37 kDa (Figure 2). The smaller fragments were readily purified on an anion-exchange FPLC column, and amino-terminal sequence determinations were performed to localize them along the primary structure deduced from the DNA sequence. The longer fragment begins with lysine 897 (refer to Figure 1), and the shorter fragment has lost four additional amino acid residues from the N-terminus and begins with threonine 901, as previously demonstrated (Banerjee et al., 1989). When these fragments were evaluated by electrospray mass spectrometry, they proved to be over 3 kDa larger than the peptides predicted from the DNA sequence of Old et al. (1990). The masses for these fragments (Table I) directly confirm that the two peptides differ only by four amino acid residues (KKPR) at the amino terminus, since the difference in mass is that of the four residues.

In order to account for the observed masses, two possibilities were considered: either the protein was covalently modified or translation extended beyond the predicted stop codon. Because none of the commonly observed modifications (e.g.,

Table I: C-Terminal Domain Analysis

fragment (Figure 1) label/N-terminus	calcd mass to the C-terminus (Da)[a] A	B	C	mass found (Da)
37A/Thr 901	25 415	33 873	37 191	37 206 ± 7[b]
37B/Arg 896	25 925	34 383	37 701	37 724 ± 8
difference[c]	510	510	510	518 ± 11

[a] The domain masses from the indicated residues to the deduced C-terminus were calculated for the sequences from Banerjee et al. (A), Old et al. (B), and this work (C). [b] The fact that the experimentally measured mass is slightly larger than the predicted mass may be rationalized as the partial oxidation of sensitive residues (e.g., methionine), since no attempt was made to protect the domain from oxidation during purification or analysis. Partially oxidized peptides were detected along with the parent peptide masses (Table II) following proteolytic digestion of the domain with LysC. [c] Mass of KKPR–water, the sequence components by which the N-termini of these domains differ.

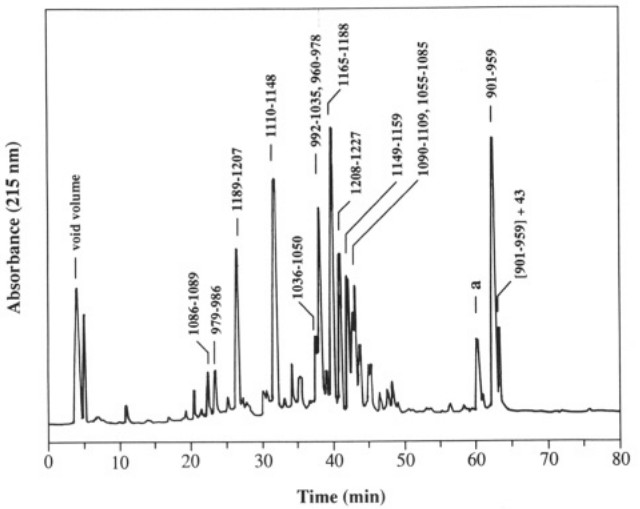

FIGURE 4: Reverse-phase HPLC of peptides generated by LysC digestion of the 37.2-kDa domain. The peak labels correspond to the residue numbers in the primary amino acid sequence, and the masses detected in the isolated fractions are presented in Table II. Not all the minor peaks were evaluated. The mass obtained for the peak labeled "a" (1507 Da) is not readily assigned to a peptide in this domain. The peak eluting at 64 min has a mass consistent with the preceding peak mass of 6774 plus 43 Da and may reflect carbamoylation of a nucleophilic residue in this long peptide by free cyanate during the overnight LysC digestion in 6 M urea (Stark et al., 1960; Stark, 1965).

phosphorylation or lipid attachment) could account for a 3.3-kDa mass increase, the published DNA sequences were reevaluated. By a process of trial and error, we found that by discounting a single nucleotide just prior to the stop codon (nucleotide 3603 from the Old et al. sequence), a new reading frame was established and an additional 3.3 kDa came into frame. This raised the possibility that the expressed protein product reflected a frameshifted reading of the nucleotide sequence, either by the translational machinery of the organisms or by the human readers of the sequence. At least two examples of *in vivo* translational frameshifting have been reported for proteins in *E. coli*: both the γ-subunit of DNA polymerase III (Tsuchihashi & Kornberg, 1990; Flower & McHenry, 1990) and release factor 2 (Craigen et al., 1985) require a specific, frameshifted reading of the messenger RNA.

Construction of a Proteolytic Map. To characterize the expressed protein product, the entire 37.2-kDa domain was digested with a lysine-specific endoproteinase (LysC from *Achromobacter*) to generate a family of peptides. The peptides were then separated by reverse-phase HPLC (Figure 4), and each fraction was characterized by electrospray mass spectrometry; these data are summarized in Table II. Over 95% of the domain mass was accounted for from a single separation run, and a peptide map (Figure 5) was constructed for this region of the protein. This analysis does not require complete chromatographic resolution of the peptides, since mixtures gave readily characterized families of masses. Masses for the two peptides predicted to come into frame by deletion of cytosine 3584 were directly observed, including the peptide predicted by the new translational stop site. This is the only peptide generated by LysC digestion that does not end in

Table II: Masses Detected for HPLC-Purified Peptides

sequence range	mass predicted	mass found	mass/charge values obsd[a]
901–959	6771.2	6774	1355, 1694.5
960–978	2261.5	2262	754, 1132
979–986	920.9	922	923
987–991	546.6		
992–1035	5082.4	5085	726, 849, 1018, 1271.5, 1696
1036–1050	1617.2 (3232.4)[b,c]	3232	809, 1078, 1617.5
1051–1054	403.4		
1055–1085	3317.3[c]	3318	1660
1086–1089	489.6	490	491
1090–1109	2358.5	2358	590, 786.5, 1180.5
1110–1148	4506.7 (9011.4)[b]	9014	1128, 1288.5, 1503.5, 1804
1149–1159	1330.5	1330	666.5, 1331
1160–1164	572.6		
1165–1188[d]	2687.8	2688	898, 1345
1189–1207[d,e]	2372.5	2372	593, 791.5, 1187.5
1208–1227[d,e]	2223.3	2224	1113

[a] The observed masses reflect mass/charge ratios, and a family of masses was observed for each peptide that represents the occupancy of protons on basic residues (including the amino terminus) of the peptide. A peptide of mass M bearing a single positive charge (proton) is expected to yield an $(M + 1)/1$ signal. Peptides bearing n positive charges yield masses of $(M + n)/n$. LysC is a good protease for this analysis, because it ensures that all the peptides formed (except the C-terminal peptide) will contain at least two positive charges, the amino-terminal nitrogen and the C-terminal lysine ε-amino group. [b] Peptides that contain cysteine were located as the oxidized cystinyl disulfides. The observed mass is therefore 2 times the predicted mass less 2 protons $(2M - 2)$. [c] Peptides containing a disagreement between the published DNA sequences. [d] Completely sequenced by Edman degradation to verify the predicted C-terminal sequence. [e] New C-terminal peptides predicted by this work. The mass of the ultimate peptide agrees with the mass predicted from the stop codon in this reading frame.

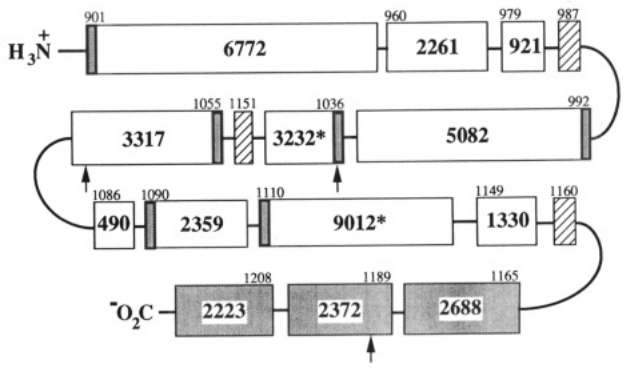

FIGURE 5: Peptide map for the C-terminal domain of methionine synthase. The numbers within the boxes give the masses for the peptides produced by LysC digestion that are deduced from the DNA sequence, and the boxes are labeled externally with the residue number for the first amino acid in the peptide. These data correspond to those presented in Table II. An asterisk indicates that the peptide contains a cysteine and was found as the oxidized disulfide. Arrows indicate the positions where published DNA sequence differences would affect the expressed peptide mass; in the penultimate peptide, the error predicts premature termination of translation. Small gray boxes indicate that four amino acids were sequenced from the amino terminus to confirm peptide identity, and completely filled boxes were completely sequenced. Hatched boxes represent the three peptides that were not found in this analysis.

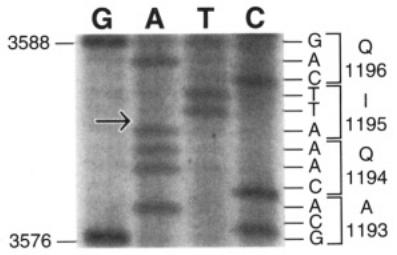

FIGURE 6: Resequencing near the C-terminus from nucleotide 3576 through 3588. The arrow represents the position where a cytosine (3584) has been inserted in the published sequences, yielding a frameshifted and prematurely terminated assignment of the open reading frame. The numbering scheme for the nucleotides corresponds to the open reading frame determined in this work.

lysine, an observation consistent with its identity as the C-terminal peptide. To prove that the masses of the ultimate and penultimate peptides were derived from the published DNA sequences, both were completely sequenced by traditional Edman degradation to confirm their identity. The penultimate 2.2-kDa LysC peptide begins with tyrosine 1189 (Figure 6), as deduced from the sequence of Old et al., but it runs through the site predicted by the stop codon in their sequence. The ultimate peptide runs to and includes aspartate 1227, the C-terminal residue consistent with the mass spectral results and predicted by the frameshifted DNA reading.

Reestablishment of the Reading Frame. Once the translated protein sequence at the C-terminus was well characterized, the inconsistency with the DNA sequence was immediately localized and identified as a frameshifted reading of the sequence. With the intent of distinguishing between a frameshifted reading of the mRNA and a misreading of the DNA sequence, both DNA strands were resequenced (Figure 7) using primers chosen to allow optimal nucleotide assignment through the region in question. Both strands gave the same reading and confirm that a cytosine (nucleotide 3584 in the coding sequence of Old et al.; see Figure 7) was incorrectly added to the sequence. Because of the large size of the plasmid and insert (9.0 kB for the plasmid p4B6.3), secondary sequences are common throughout the sequencing gels. A review of the basis gels for the first published sequence (Banerjee et al., 1989) showed that this was a likely cause of the incorrect sequence interpretation.

Discrepancies between Published DNA Sequences. Three discrepancies in the protein sequences deduced from the DNA sequences of Banerjee et al. (1989) and Old et al. (1990) occur in the C-terminal 65.7 kDa of methionine synthase characterized in this report. Analysis of these discrepancies illustrates the utility and limitations of sequence confirmation by electrospray mass spectrometric analysis. The first discrepancy occurs at residue 760 in the 28.0-kDa cobalamin-binding domain, but here the error in the domain mass determination (±20 Da) precluded distinguishing the 14-Da difference between the aspartate and the glutamate residue at this site deduced from the DNA sequences. However, if the domain were to be further proteolyzed with LysC, the smaller peptides derived from the domain could be determined with lower absolute uncertainty. For example, given an error of ~0.1% in the mass determination, a peptide of mass 3000 Da could now be measured to ± 3 Da. This process was demonstrated by constructing the LysC peptide map for the 37.2-kDa C-terminal domain. The two remaining sequence differences, occurring at amino acid residues 1037 and 1079 (Figure 8), were localized within individual LysC peptides, and the observed masses provided immediate support for one sequence over the other.

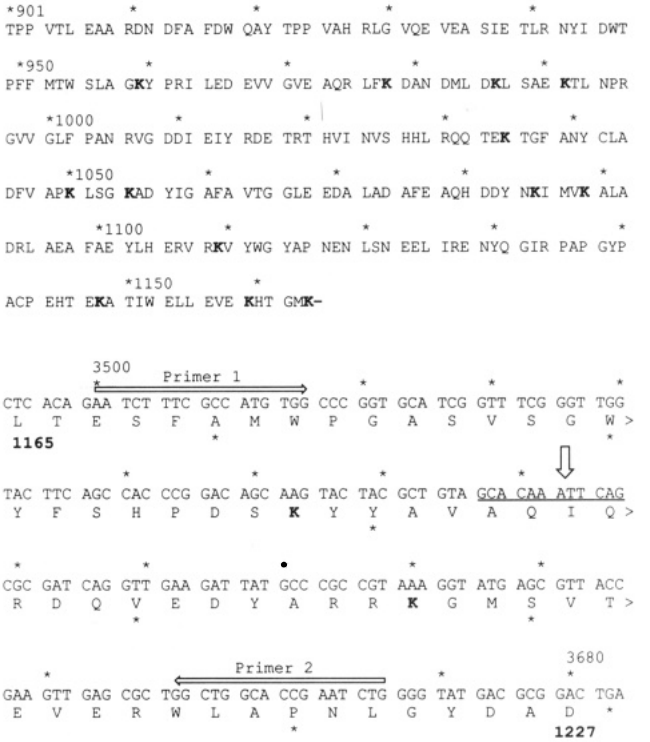

FIGURE 7: Corrected sequence at the C-terminus. The lysine residues are presented in bold, and they mark the C-termini of the LysC peptides. The vertical arrow indicates the position where a cytosine (3584) has been twice incorrectly inserted in the DNA sequence, and horizontal arrows indicate the locations of primers chosen to optimize sequence reading through the underlined region. The underlined region corresponds to that presented in Figure 6. The entire protein sequence from leucine 1165 through the C-terminus was confirmed by Edman degradation. Note that the ultimate peptide is the only one that does not end in lysine.

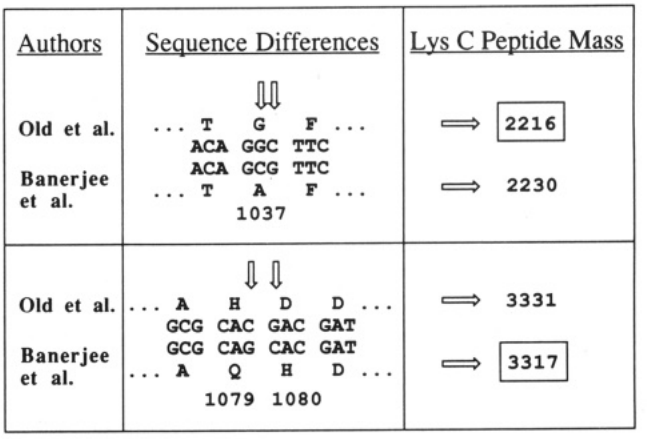

FIGURE 8: Prediction of DNA sequence from peptide mass. Vertical arrows indicate the two locations where a transposed GC pair occurs, and the boxes highlight masses of peptides isolated following LysC digestion. The first discrepancy occurs at position 1037, and the presence of glycine was confirmed by amino acid sequencing. The second affects the two amino acid residues following alanine 1078, and the peptide identity was confirmed by amino acid sequencing. This supports, but does not unambiguously establish, the assignment of Banerjee et al.

Both disagreements are the result of the transposition of a guanine and cytosine pair (GC vs CG) in the published DNA sequences. The first transposition occurs within a codon (Figure 8) and results in a single amino acid disagreement: Is glycine or alanine present at amino acid residue 1037? A LysC peptide with a mass consistent with glycine occupying position 1037 was isolated, and because this discrepancy was located near the N-terminus of the peptide, amino acid

sequencing confirmed both the identity of the peptide and the presence of glycine. At the second site of disagreement, the transposed pair of DNA bases altered the interpretation of two codons, but again an overall mass difference of 14 was predicted. A LysC peptide possessing a mass of 3317 Da was isolated (Figure 4), and its identity as the disputed peptide was confirmed by N-terminal sequencing of the first four residues. However, because the disagreement occurred late in this long peptide, the specific amino acids were not verified by Edman degradation. The evidence strongly supports, but does not confirm, the sequence of Banerjee et al. at this site.

The example of mapping a domain by mass presented here also illustrates a limitation of the technique. Agreement between a peptide mass predicted by sequence and a mass experimentally measured is strong but not conclusive evidence for peptide identification. N-terminal amino acid analysis can then confirm peptide identity, but does not rigorously validate either the protein or the DNA sequence. For example, lysine and glutamine possess the same mass, as do leucine and isoleucine, and these amino acids are translated through a set of codons that differ by single nucleotides. Depending on the size of the peptide and the error in the mass measurement, amino acids that differ by a single mass unit may be indistinguishable components of a peptide, although these problems could be addressed by high-resolution tandem mass spectrometry studies using high-energy, collision-induced dissociation (Biemann, 1990). So while the relatively low-resolution technique described here is a powerful tool to confirm the relationship between observed protein products and sequenced DNA, minor or offsetting mass differences may remain silent in this analysis, as will differences in codon triplets that predict the same amino acids. Direct interpretation of mass as proof of sequence should therefore be made with caution.

DISCUSSION

Our goal in this work was to provide a practical example of the application of electrospray mass spectrometry to proteins that are too large to yield useful information by direct mass analysis. In general, we found that cycles of proteolysis followed by characterization of the resulting fragment masses afforded increasingly well resolved information concerning sequence, structure, and function. We studied cobalamin-dependent methionine synthase for *E. coli*, a large monomeric enzyme of 136.1 kDa, and our initial approach was to divide the enzyme by limited proteolysis with trypsin into a set of fragments amenable to electrospray mass analysis. Following the first round of proteolytic division, we were able to purify and characterize a 28.0-kDa domain that retained the noncovalently bound cobalamin prosthetic group and a 37.2-kDa domain that included the carboxyl terminus. Because we have been able to assign specific binding and catalytic functions to these fragments, this work provided part of the structural framework for our working model of how these fragments contribute to the overall turnover scheme of methionine synthase (Drummond et al., 1993). We propose that this approach should be an excellent complement to traditional methods used to characterize complex protein structures.

Perhaps the most striking use of electrospray mass spectrometry reported here is the observation of a complex between a small domain and the noncovalently bound cobalamin prosthetic group. When the domain was introduced into the mass spectrometer in a neutral, aqueous solution that allows the native structure to be preserved, we were able to observe primarily the holodomain. This result is analogous to the demonstration of bound complexes of myoglobin with non-covalently bound heme, although we were unable to completely resolve the prosthetic group from the protein by acidic denaturation. Even so, the ability to characterize noncovalently bound complexes by electrospray mass spectrometry holds the promise that coenzyme or substrate binding regions of other proteins may be located within a larger structure, assuming that smaller fragments or domains retain the determinants required for tight binding of these molecules. It should be emphasized that the mass analysis reported here was performed on ~125 pmol of the domain, although we initially isolated the domains in nanomolar quantities.

A second important feature inherent to the determination of a domain mass is the ability to assign the boundaries of the domain within the primary amino acid sequence. In combination with N-terminal sequence analysis and a DNA sequence, the corresponding C-terminal residue can be deduced, as shown in Figures 1 and 5. When a fragment containing the C-terminus from the parent protein can be purified, mass determination can be used to verify the identity of the deduced translational stop site. During the division of a protein into domains by limited proteolysis, it is always possible that small peptides will be released from regions between domains or from the C-terminus of the domain. When this happens, the observed mass should still be consistent with the deduced protein sequence, and the fragment isolated should end in a residue consistent with the sequence specificity of the protease.

The strategy of proteolyzing a large structure into smaller units to isolate individual regions of interest is not limited to the intact protein. This was illustrated by the lysine-specific proteolysis of the C-terminal domain into a set of peptides that were again separated and characterized by the combination of electrospray mass analysis and Edman degradation. The peptide mass values generated are likely to be unique, and a map that orders these peptides along the amino acid sequence can be constructed from this information alone. The peptide map can contribute more than the confirmation of primary sequence predictions; in this case, it allowed us to establish the correct reading frame and translational stop site for methionine synthase. More generally, peptide mapping by mass should be a useful tool to allow rapid scanning of long polypeptides for modification, derived either from chemical alterations or from mutagenesis at the DNA level. For example, the peptide map of the 37.2-kDa domain has enabled us to detect and define covalent modifications of this domain that are associated with the inactivation of methionine synthase by nitrous oxide. These results will be presented in a paper that is currently in preparation.

ACKNOWLEDGMENT

We would like to credit the personnel in the Biomedical Research Core Facilities at the University of Michigan, under the direction of Dr. Philip Andrews, for their expertise in protein and peptide analysis. Ms. Natalie Dales performed the peptide mass spectrometric analyses, and Ms. Sari Vlahakis carried out the N-terminal sequencing of peptides. We are grateful to Mr. Brian Ernsting and Ms. Sha Huang for the DNA sequencing reported in this paper.

REFERENCES

Allen, M. H., & Vestal, M. L. (1992) *J. Am. Soc. Mass Spectrom. 3*, 18–26.

Andrews, P. C., Allen, M. H., Vestal, M. L., & Nelson, R. W. (1992) in *Techniques in Protein Chemistry III* (Angeletti, R. H., Ed.) pp 515–523, Academic Press, San Diego.

Baca, M., & Kent, S. B. H. (1992) *J. Am. Chem. Soc. 114*, 3992–3993.

Banerjee, R. V., Johnston, N. L., Sobeski, J. K., Datta, P., & Matthews, R. G. (1989) *J. Biol. Chem. 264* (23), 13888–13895.

Biemann, K. (1990) *Methods Enzymol. 193*, 455–479.

Chait, B. T., & Kent, S. B. H. (1992) *Science 257*, 1885–1894.

Chowdhury, S. K., Katta, V., & Chait, B. T. (1990) *Biochem. Biophys. Res. Commun. 167* (2), 686–692.

Covey, T. R., Huang, H. C., & Henion, J. D. (1991) *Anal. Chem. 63*, 1196–1200.

Craigen, W. J., Cook, R. G., Tate, W. P., & Caskey, T. C. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 3616–3620.

Drummond, J. T., Huang, S., Blumenthal, R. M., & Matthews, R. G. (1993) *Biochemistry* (following paper in this issue).

Edmonds, C. G., Loo, J. A., Ogorzalek Loo, R. R., & Smith, R. D. (1991) in *Techniques in Protein Chemistry II* (Villafranca, J. J., Ed.) pp 487–495, Academic Press, San Diego.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989) *Science 246*, 64–71.

Flower, A. M., & McHenry, C. S. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 3713–3717.

Ganem, B., Li, Y.-T., & Henion, J. D. (1991a) *J. Am. Chem. Soc. 113*, 6294–6296.

Ganem, B., Li, Y.-T., & Henion, J. D. (1991b) *J. Am. Chem. Soc. 113*, 7818–7819.

Ganguly, A. K., Pramanik, B. N., Tsarbopoulos, A., Covey, T. R., Huang, E., & Fuhrman, S. A. (1992) *J. Am. Chem. Soc. 114*, 6559–6560.

Hogenkamp, H. P. C. (1982) in B_{12} (Dolphin, D., Ed.) Vol. 1, pp 293–323, Wiley-Interscience, New York.

Inglis, A. S. (1991) *Anal. Biochem. 195*, 183–196.

Katta, V., & Chait, B. T. (1991) *J. Am. Chem. Soc. 113*, 8534–8535.

Ling, V., Guzetta, A. W., Canova-Davis, E., Stults, J. T., Hancock, W. S., Covey, T. R., & Shushan, B. I. (1991) *Anal. Chem. 63* (24), 2909–2915.

Luschinsky, C. L., Drummond, J. T., Matthews, R. G., & Ludwig, M. L. (1992) *J. Mol. Biol. 225*, 557–560.

Michel, H., Griffin, P. R., Shabanowitz, J., Hunt, D. F., & Bennett, J. (1991) *J. Biol. Chem. 266* (26), 17584–17591.

Ogorzalek Loo, R. R., Goodlett, D. R., Smith, R. D., & Loo, J. A. (1993) *J. Am. Chem. Soc.* (in press).

Old, I. G., Margarita, D., Glass, R. E., & Saint Girons, I. (1990) *Gene 87*, 15–21.

Pratt, J. M. (1982) in B_{12} (Dolphin, D., Ed.) Vol. 1, pp 326–392, Wiley-Interscience, New York.

Prome, D., Blouquit, Y., Ponthus, C., Prome, J.-C., & Rosa, J. (1991) *J. Biol. Chem. 266* (20), 13050–13054.

Smith, R. D., Loo, J. A., Goodlett, D. R., Edmonds, C. G., Barinaga, C. J., & Udseth, H. R. (1990) *Anal. Chem. 62* (9), 882–889.

Stark, G. R. (1965) *Biochemistry 4*, 1030–1036.

Stark, G. R., Stein, W. H., & Moore, S. (1960) *J. Biol. Chem. 235* (11), 3177–3181.

Taylor, R. T. (1970) *Arch. Biochem. Biophys. 137*, 529–546.

Taylor, R. T., & Weissbach, H. (1967) *Arch. Biochem. Biophys. 119*, 572–579.

Tsuchihashi, Z., & Kornberg, A. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 2516–2520.

Van Dorsselaer, A., Bitsch, F., Green, B., Jarvis, S., Lepage, P., Bischoff, R., Kolbe, H. V. J., & Roitsch, C. (1990) *Biomed. Environ. Mass Spectrom. 19* (11), 692–704.

Wood, W. B., Wilson, J. H., Benbow, R. M., & Hood, L. E. (1974) in *Biochemistry, A Problems Approach*, W. A. Benjamin Inc., Menlo Park, CA.