Yonath, A., & Wittmann, H. G. (1988) *Methods Enzymol.* (in press).
Zimmermann, R. A. (1980) in *Ribosomes* (Chambliss, G., Craven, G. R., Davies, J., Davis, K., Kahan, L., & Nomura, M., Eds.) pp 135-169, University Park Press, Baltimore, MD.
Zwieb, C., & Dahlberg, A. E. (1984) *Nucleic Acids Res. 12*, 4361-4375.

Accelerated Publications

Correction of the cDNA-Derived Protein Sequence of Prostatic Spermine Binding Protein: Pivotal Role of Tandem Mass Spectrometry in Sequence Analysis†

Robert J. Anderegg,‡,§ Steven A. Carr,*,‡ I. Yih Huang,‡ Richard A. Hiipakka,∥ Chawnshang Chang,∥ and Shutsung Liao∥

Department of Physical and Structural Chemistry, L-940, and Department of Analytical Chemistry, L-950, Smith Kline and French Laboratories, Box 1539, King of Prussia, Pennsylvania 19406, and Ben May Institute and Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, Illinois 60637

Received March 17, 1988; Revised Manuscript Received April 18, 1988

ABSTRACT: Spermine binding protein (SBP) is a rat ventral prostate protein that binds various polyamines, and the level of this protein and its mRNA is regulated by androgens. Previously, the cDNA for SBP was cloned and sequenced and an amino acid sequence deduced from the cDNA. Data from partial amino acid sequencing of the purified protein were consistent with the amino acid sequence deduced from the cDNA. However, the amino terminus of the protein was blocked, and therefore, direct protein sequence information confirming the cDNA reading frame of this region could not be obtained by Edman degradation. We have now employed an integrated approach using fast atom bombardment mass spectrometry, tandem mass spectrometry, and conventional sequencing methodologies to establish the amino-terminal sequence of the protein and to identify an amino acid sequence (35 residues) present in the purified protein but missing from the amino acid sequence deduced from cDNA clones for this protein. The missing piece of cDNA corresponds to an exon found in mouse genomic clones for a protein similar to rat SBP. Therefore, the cDNA clones for rat SBP may represent splicing variants that lack the sequence information of one exon. The blocked amino terminus of the protein was identified as 5-oxopyrrolidine-2-carboxylic acid. Mass spectrometry also provided evidence regarding glycosylation of the protein. The first of two potential glycosylation sites clearly carries carbohydrate; the second site is, at most, only partially glycosylated.

Mass spectrometry (MS)[1] is now a well-accepted technique in the overall structural analysis of peptides and proteins. The recent dramatic increase in the use of MS in such studies is largely due to the advent of fast atom bombardment (FAB) (Barber et al., 1982), an ionization technique that provides the molecular weights of as little as picomole amounts of peptides present in complex mixtures without the need for extensive chromatographic purification or chemical derivatization prior to analysis [for recent reviews, see Biemann and Martin (1987) and Carr et al., 1988)]. The FABMS peptide mapping procedure can be used to rapidly corroborate protein sequences obtained by DNA or cDNA sequencing. Peptide signals that do not map into the predicted sequence may indicate the presence of cloning errors, sequencing errors, or posttranslational modifications. In favorable cases the modification or error may be localized to a specific region of the otherwise correct protein sequence. Because sequence information is generally very limited or not obtainable by FABMS alone, the strategy described above is only applicable to proteins of known primary structure.

Tandem mass spectrometry employing two consecutive mass analyzers can provide structural information on individual peptides in complex mixtures, and therefore, this technique has great potential in the sequencing of peptides derived from proteins (Biemann & Martin, 1987; Hunt et al., 1986; Johnson & Biemann, 1987; Crabb et al., 1986; Biemann & Scoble, 1987; Carr et al., 1988). Protonated molecular ions $(M + H)^+$ of peptides in a mixture can be selectively fragmented, and under favorable circumstances, the sequences can be deduced

† This work was supported by grants from the National Institutes of Health to R.J.A. (GM 32602) and S.L. (AM 09461) and from Smith Kline and French Laboratories to S.A.C., R.J.A., and I.Y.H. These results were presented in preliminary form at the 35th Annual Conference on Mass Spectrometry and Allied Topics, May 24-29, 1987, Denver, CO (Carr, et al., 1987a), and at the 78th Annual Meeting of the American Society of Biological Chemists, June 7-11, 1987, Philadelphia, PA (Carr et al., 1987b).

* To whom correspondence should be addressed.
‡ Smith Kline and French Laboratories.
§ On sabbatical leave from the University of Maine, Orono, ME.
∥ The University of Chicago.

[1] Abbreviations: FABMS, fast atom bombardment mass spectrometry; MS, mass spectrometry; FAB, fast atom bombardment; SBP, spermine binding protein; HPLC, high-performance liquid chromatography; PITC, phenyl isothiocyanate; PTH, phenylthiohydantoin; TFA, trifluoroacetic acid; Pca, 5-oxopyrrolidine-2-carboxylic acid; Tris, tris(hydroxymethyl)aminomethane; EDTA, ethylenediaminetetraacetic acid; DTT, dithiothreitol; RP-HPLC, reversed-phase high-performance liquid chromatography.

0006-2960/88/0427-4214$01.50/0

from the resulting daughter ion spectra. The approach is ideal for use in corroborating DNA or cDNA predictions: the frame of translation is easily determined, and any errors in the DNA sequence can be located and corrected. Furthermore, the sequences of posttranslationally modified peptides can be assigned by tandem MS (Carr et al., 1988; Crabb et al., 1986; Biemann & Scoble, 1987). The interpretation of daughter ion spectra from completely unknown peptides (i.e., in the absence of a predicted or homologous sequence) is obviously more difficult and has only been demonstrated in a limited number of cases.

Herein we describe the results of our studies on the spermine binding protein (SBP) from rat ventral prostate. SBP is an androgen-regulated glycoprotein that binds spermine as well as other polyamines (Hiipakka et al., 1984). The cDNA for SBP has been cloned and sequenced (Chang et al., 1987). The amino acid sequence deduced from the cDNA is consistent with the data from partial amino acid sequencing of tryptic and cyanogen bromide peptides of SBP (Chang et al., 1987). However, a blocked amino terminus prevented sequencing this region and, so, prevented confirming the reading frame of SBP cDNA coding for the amino terminus. With a combination of FABMS, tandem MS, and automated Edman degradation, the amino-terminal region of SBP has now been sequenced and the N-terminal blocking group identified as 5-oxo-pyrrolidine-2-carboxylic acid (Pca). However, and even more significantly, we were able to show that, in the reverse transcription from mRNA to cDNA prior to sequencing, a portion of the message corresponding to 106 bases was omitted, creating a frame shift and a gap of some 35 amino acids, including the amino terminus, in the predicted protein sequence. Although tandem mass spectrometry was indispensable in the rapid solution of the problem, it also displayed some limitations that were best overcome by other methods. Some of these specific advantages and disadvantages are discussed in this paper. Mass spectrometric results also showed that one of the two potential glycosylation sites carries carbohydrate, in agreement with other studies (Chang et al., 1987), but that the other site is, at most, only partially glycosylated.

Materials and Methods

SBP from rat ventral prostate was isolated as previously described (Hiipakka et al., 1984). The protein was reduced and carboxymethylated in 0.3 M Tris-HCl buffer, pH 8.1, containing 2 mM EDTA and 8 M urea with dithiothreitol in 50-fold molar excess over cysteinyl residues. The reaction was allowed to proceed at 50 °C for 4 h under N_2. After being cooled to room temperature, the reaction mixture was treated with a 100-fold molar excess of sodium iodoacetate at 20 °C for 20 min in the dark. Reagents were removed by dialysis against 50 mM NH_4HCO_3 at 4 °C.

Enzymatic Digests. Typically 1 mg of reduced and carboxymethylated protein was dissolved in 1 mL of 0.050 M ammonium bicarbonate buffer, pH 8.5. Trypsin (TPCK treated, Cooper Biomedical) was dissolved in the same buffer, and an aliquot was added to the protein solution to give an enzyme:substrate ratio of 1:100. The digest was incubated at 37 °C for 9–12 h. The digestion was stopped by addition of 2 drops of glacial acetic acid prior to lyophilization. The digest was redissolved in acetonitrile–water–trifluoroacetic acid (TFA) (200:800:1). Aliquots were removed for mass spectrometry or HPLC. A portion of the tryptic digest of SBP was treated with peptide:*N*-glycosidase F as previously described (Carr & Roberts, 1986) and then analyzed by FABMS. Chymotryptic (Cooper Biomedical) digests (3 h) were obtained in a similar fashion.

Peptides digested with pyroglutamate aminopeptidase (Boehringer Mannheim) were dissolved in 100 μL of 0.1 M Na_2HPO_4, 10 mM in EDTA, 5 mM in DTT, and 5% (v/v) in glycerol at pH 8. Enzyme (7 μL of a 1.0 μg/μL solution) was added to the peptide in buffer, and the reaction proceeded at 4 °C for 4 h under N_2. An additional 3 μL of enzyme solution was added and the reaction continued at 4 °C for an additional 12 h. The mixture was brought to room temperature, an additional 10 μL of enzyme solution added, and the reaction allowed to continue for 7 h. Peptides were purified by HPLC prior to analysis by FABMS.

Peptide mixtures were fractionated by HPLC as previously described (Hemling et al., 1988).

Manual and Automated Edman Degradation. Manual Edman degradation was performed on mixtures of peptides as previously described (Hemling et al., 1988). Automated Edman degradation was carried out with a Beckman System 890 M-2 sequencer with HPLC identification and quantitation of PTH-amino acids performed as previously described (Hawke et al., 1982).

FAB Mass Spectrometry and Tandem Mass Spectrometry. FAB mass spectra were obtained with a VG ZAB 1F-HF mass spectrometer equipped with a standard FAB ion source and an Ion Tech fast atom gun as previously described (Hemling et al., 1988). Approximately 1–2 μL of a peptide solution [0.05–0.5 nmol of peptide/μL dissolved in acetonitrile–water (2:1 v/v) and 0.1% TFA (v/v)] was used per analysis. The peptide solution was added to ca. 1 μL of thioglycerol (Sigma) on the FAB probe tip, and the solvents were evaporated under vacuum in the probe lock of the mass spectrometer. An additional 1 μL of thioglycerol was added prior to analysis by FABMS.

FAB tandem MS experiments were conducted on a VG ZAB-SE 4F (BE-EB) four-sector magnetic deflection mass spectrometer. Conditions for FAB were the same as described above. The resolution of MS-1 of the two coupled, double-focusing mass analyzers was adjusted for unit mass resolution of the selected parent, while the resolution of MS-2 was set between 800 and 1000. Daughter ions were produced by introducing He into the collision cell located between MS-1 and MS-2 at a pressure sufficient to reduce the selected parent ion beam by 75%. Either 10- or 5-keV collisions were employed by grounding or electrically floating the collision cell to 5 keV, respectively. Daughter ion spectra were obtained by a computer-generated linked scan (exponential down, 20–30 s/decade) of MS-2 such that the ratio of E/B was maintained constant. Six to eight scans were summed in the raw profile mode with a VG 11-250J data system to acquire and process all data. MS-2 was calibrated in the FAB mode with a mixture of lithium, sodium, and cesium iodides by using a second FAB ion source located between MS-1 and MS-2.

Results

The cDNA corresponding to the rat SBP has been sequenced (Chang et al., 1987) and a protein sequence predicted (Figure 1). The amino acid numbering is based on the cDNA sequence and the supposition that Met_1 is the start of initiation. The sequence reveals an extremely acidic carboxyl-terminal region extending from residue 182 to residue 307, undoubtedly accounting for the affinity of this protein for polyamines. Edman degradation previously confirmed the sequence of residues 83–190 (Chang et al., 1987). From these studies it was concluded that the amino terminus of the protein is blocked, and our mass spectrometric investigations were initiated to determine the nature of the blocking group and to establish the translational start site that could not be deter-

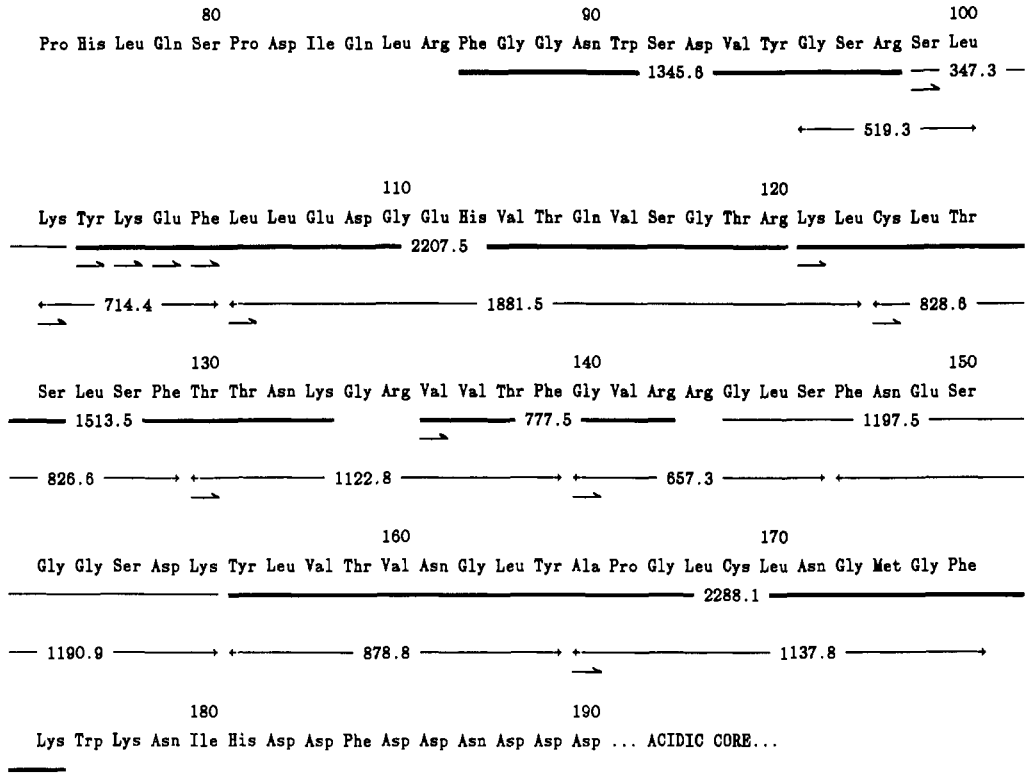

FIGURE 1: Predicted protein sequence of prostatic SBP from rat (Chang et al., 1987). Peptides identified by FABMS are identified as follows: chymotryptic peptides = ↔; tryptic peptides = —. In places where overlapping or redundant peptides were observed in the same digest, only the largest peptide is shown. Peptides appearing after treatment with peptide:*N*-glycosidase F are included (see text). Peptides sequenced by tandem MS are underlined with a heavy bar. Half arrows indicate amino acids removed by manual Edman degradation of the peptide mixtures and identified by FABMS. The amino acid numbering is based on the cDNA sequence and the supposition that the first methionine prior to the observed sequence is the site of initiation and is given the number 1. Edman degradation previously confirmed the sequence of residues 83–190 (Chang et al., 1987).

mined by Edman degradation. As our studies described below revealed, the predicted amino acid sequence prior to residue 82 (Figure 1) is entirely incorrect.

SBP isolated from rat ventral prostate was digested with trypsin, and the resulting peptide mixture, as well as fractions obtained after partial fractionation by RP-HPLC, was analyzed by FABMS to determine the molecular weights of the peptides (Morris et al., 1983; Biemann & Martin, 1987; Carr et al., 1988). Many of the expected peptides predicted from the cDNA sequence were observed (Figure 1), with the exception of (i) tryptic peptides encompassing the two N-linked glycosylation sites at residues 90 and 148, (ii) peptides derived from the acidic peptide "core", and (iii) the amino-terminal region. Lack of coverage in the FABMS mapping experiments of regions i and ii, noted above, was not surprising because, in FABMS, signals from very hydrophilic molecules such as glycopeptides (Carr et al., 1988; Carr & Roberts, 1986), or the hydrophilic, acidic peptides derived from the carboxyl-terminus of SBP, are generally suppressed in the presence of amphiphilic and hydrophobic peptides and are often not detected (Clench et al., 1985; Naylor et al., 1986). However, signals derived from the predicted amino-terminal region should have been observed, since this is the region of the protein containing the smallest number of charged residues. Furthermore, four peptide signals were observed in the FABMS experiments that were not predicted on the basis of the cDNA. The protonated molecular weights of these peptides were 345.2, 586.4, 1174.7, and 2562.2. Mixtures containing these peptides were subjected to one step of manual Edman degradation and reanalyzed by FABMS, which indicated that the 586 peptide begins with Gly and that the 1174 peptide begins with xLeu (xLeu is the designation we use for leucine or isoleucine, which is isobaric and cannot usually be differentiated in conventional MS). However, the tryptic fragment of mass 2562.2 did not change its mass upon Edman degradation, indicating that this peptide is blocked and, therefore, probably represents the N-terminus of the protein.

In order to determine the sequences of these undefined signals, which yield only molecular weight information in their conventional FAB mass spectra, HPLC fractions of the tryptic digest containing these peptides were subjected to tandem mass spectrometry by using a high-performance four-sector instrument (Biemann & Martin, 1987; Carr et al., 1988; Johnson & Biemann, 1987; Crabb et al., 1986; Biemann & Scoble, 1987). In this technique, a protonated molecular ion of an intact peptide is selected from a mixture of such signals by using the first of two coupled, double-focusing mass analyzers (i.e., MS-1 = B_1E_1, where B and E signify magnetic and electric sectors, respectively). The resolving power of MS-1 is generally set to select only the monoisotopic $(M + H)^+$. This results in simpler daughter ion mass spectra. The mass-selected peptide is transmitted into a collision cell located in the field-free region between MS-1 and MS-2. Collision with a neutral gas such as helium or argon at either 10 (collision cell at ground potential) or 5 keV (collision cell at 5 keV) energetically activates the parent ions and causes them to fragment. The fragment or daughter ions are detected by scanning the second, high-resolution mass spectrometer (MS-2). The daughter ion spectra so recorded carry sequence information from which the amino acid sequence of the parent can be frequently deduced. Peptides for which tandem MS data were obtained are underlined with heavy bars in Figures 1 and 4. In most cases (with a few exceptions noted below), these daughter ion spectra independently established the sequences of the peptides that gave $(M + H)^+$ signals that did not map into the predicted protein sequence. Tandem MS data were

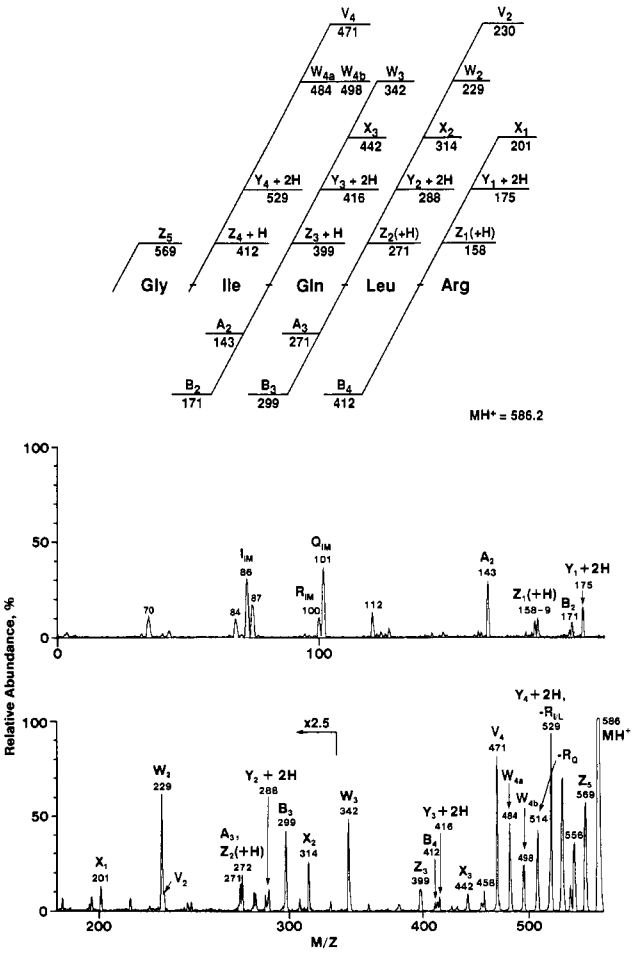

FIGURE 2: Tandem mass spectrum of the tryptic peptide of $(M + H)^+ = 586.4$. FABMS in MS-1 of the tandem mass spectrometer was used to identify this parent ion (present in an HPLC fraction containing four other peptide signals) and to select it for collision with He at 10 keV (pressure sufficient to reduce parent to 25% of original intensity) and mass analysis in MS-2. See Figure 3 for an explanation of the notation used to identify the fragment ions.

also obtained on many peptides that fit the deduced sequence, and in each case, the spectra confirmed the peptide sequence predicted from the cDNA. The specifics of these analyses are described below.

The daughter ion spectrum of the SBP tryptic peptide of $(M + H)^+ = 586.4$ is shown in Figure 2. The sequence Gly-Ile-Gln-Leu-Arg can be readily discerned (as described below), which represents the first four residues of the previously determined protein sequence of SBP (Chang et al., 1987), residues 83–86, but shows a glycine in position 82 rather than aspartic acid as predicted from the cDNA. In the tandem mass spectra of peptides, sequence-defining "backbone" fragments are formed by cleavage of the amide backbone, with or without H rearrangement, resulting in ion series derived from the C-terminus (X_n, Y_n, and Z_n; Figures 2 and 3) and complementary series of fragments (A_n and B_n) from the N-terminus. Fragment ions arising by loss of side chains from backbone fragments (labeled W_n and V_n; Figures 2 and 3) are also commonly observed (Biemann & Scoble, 1987; Stults & Watson, 1987). The V_n fragment ions come from the Y_n + 2H by loss of RH, where R is the entire side chain of the nascent amino-terminal residue. The W_n fragment ions originate from the Z_n + H by β–γ cleavage of the side chain of the nascent amino-terminal residue. Multiple W_n ions are obtained when R_n is branched at the β-position as with Thr, Val, and Ile (Stults & Watson, 1987). (In some cases this secondary peak multiplicity may confuse interpretation; see below.) The W ions for Leu and Ile (Figure 3) are particularly useful, since they allow these two isobaric residues to be distinguished. For example, position 2 of this peptide is defined as Ile by the presence of the peaks at m/z 484 and 498 (W_{4a} and W_{4b}, respectively) and absence of a peak at m/z 470, the calculated mass for the W_4 fragment if Leu were at this position. Similarly, residue 4 is defined as Leu by the presence of the W_2 at m/z 229 and absence of peaks at m/z 243 and 257, the calculated masses for the W_{2a} and W_{2b} fragments, respectively, if this position were Ile. Other fragment ions that are commonly observed in the tandem mass spectra of peptides include loss of amino acid side chains from the $(M + H)^+$ (e.g., $-R_Q$, m/z 514; Figure 2) and immonium ions $(H_2N{=}CHR)^+$ (e.g., Q_{im}, m/z 101; Figure 2), where R is the amino acid side chain.

Tandem mass spectra were also obtained on three other unmapped tryptic peptides with $(M + H)^+ = 345.2$, 1174.7, and 2562.2 present in HPLC subfractions of the digest. The sequence Gly-Ile-Arg was established from the daughter ion spectrum of the $(M + H)^+ = 345$ peptide. Unfortunately, the FABMS signal for the N-terminally blocked peptide of $(M + H)^+ = 2562.2$ was weak, and no useful sequence data were obtained from its daughter ion spectrum.[2] The daughter ion spectrum of the tryptic peptide $(M + H)^+ = 1174.7$ did not give an unambiguous answer because the sequence ions from the middle region of the peptide were very weak. Nevertheless, the partial sequence xLeu-Phe-Leu-Thr-?-?-?-Leu-Ile-Lys could be established. Two possible sequences for residues 5–7, -Ile-Ile-Thr- or -Val-Ile-Asp-, could be supported by relatively weak peaks in the middle region of the daughter ion spectrum. The latter sequence was better supported in that seven of ten key sequence defining peaks (e.g., Y_n, Z_n, V_n, where n = 4, 5, and W_n, where n = 4–6) were present as opposed to only five out of eleven for the former sequence.

It was important to establish the sequence of the $(M + H)^+ = 1174.7$ peptide, since the partial sequence defined by the tandem MS data did not fit the sequence predicted from the cDNA, nor did it fit the sequences derived from translation of the cDNA in either of the two other possible reading frames. Therefore, the peptide was purified to homogeneity and subjected to automated Edman sequencing. A sequence Ile-Phe-Leu-Thr-Val-Ile-Asp-Leu-Ile-Lys was revealed, which fit the sequence best supported by the MS/MS data. No portion of this sequence can be found in any reading frame of the rat SBP cDNA in the region prior to the known protein terminus, indicating a serious problem in the predicted sequence for the N-terminal region of the protein.

On the basis of the FABMS, MS/MS, and Edman data a tentative alignment of the N-terminal tryptic peptides could be made: X-(MH = 2562.2)-?-(MH = 1174.7)Ile-Phe-Leu-Thr-Val-Ile-Asp-Leu-Ile-Lys-?-(MH = 586.4)Asp-Ile-Glu-Leu-Arg(86)..., where X = a blocking group and ? indicates the two possible locations of the tripeptide Gly-Ile-Arg (the numbering refers to that of the original sequence data, Figure 1). In an effort to obtain a smaller peptide containing the blocked N-terminus, as well as peptides that would finalize the ordering of the tryptic fragments, the SBP was digested with chymotrypsin for a brief time (3 h). Once again, a large portion (>85%) of the known protein sequence was represented by FABMS of the chymotryptic digest (Figure 1). Two peptides, $(M + H)^+ = 671.3$ and 1326.5, demonstrate that

[2] In general, good tandem mass spectra are more difficult to obtain for peptides with masses greater than 2500 Da on currently available instrumentation. However, useful tandem mass spectra have been obtained on peptides up to 4000 Da (Carr et al., 1987c).

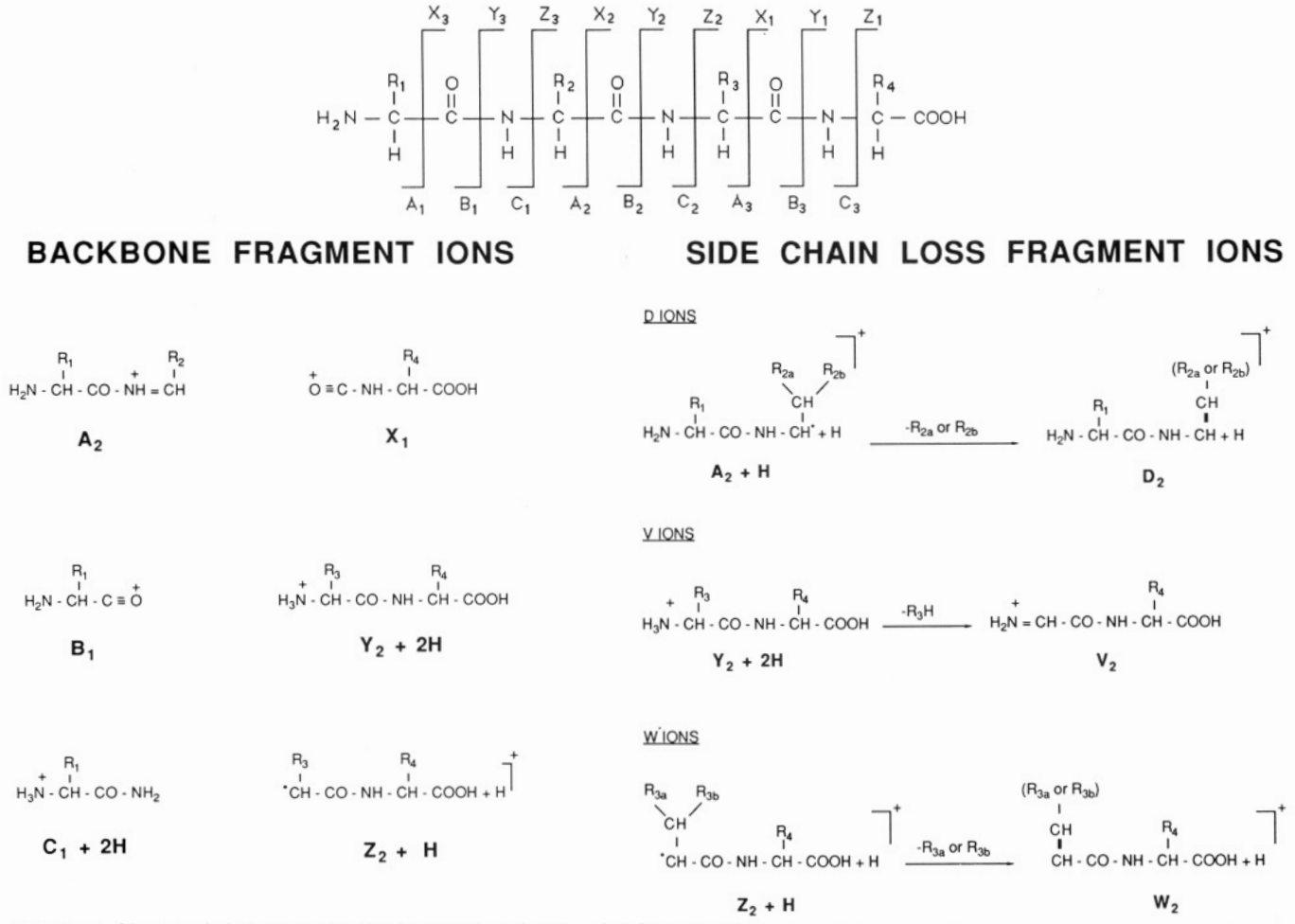

FIGURE 3: Characteristic backbone and side-chain-loss daughter ion fragments formed by high-energy (i.e., kilovolt) collisional activation of peptides and observed in the resulting MS/MS spectra (see text for discussion).

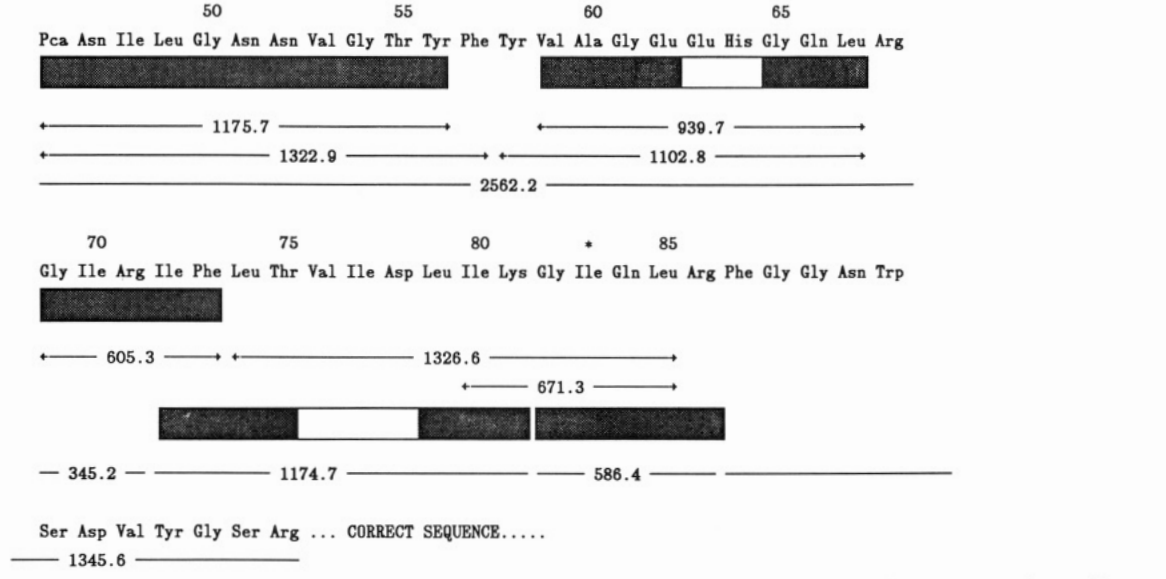

FIGURE 4: N-Terminal sequence of prostatic SBP established by FABMS and tandem MS: chymotryptic peptides = ↔; tryptic peptides = —. Peptides sequenced by tandem MS are indicated by hashed boxes. Uncertainties in the sequences of the peptides of $(M + H)^+ = 1174$ and 939 (nonhashed regions of boxes) were resolved by Edman degradation of the isolated peptides (see text). The previously determined Edman sequence began at residue 83 (see legend to Figure 1) and is indicated by an asterisk.

the tryptic fragment of $(M + H)^+ = 1174.7$ immediately precedes the tryptic $(M + H)^+ = 586.4$ peptide and, therefore, that the tripeptide Gly-Ile-Arg connects the blocked N-terminus with the remainder of the sequence (Figure 4).

Two chymotryptic peptides, $(M + H)^+ = 1175.7$ and 1322.9, did not change mass upon manual Edman degradation, as determined by FABMS. These are undoubtedly derived from the blocked N-terminus. The daughter ion spectrum of $(M + H)^+ = 1175.7$ shown in Figure 5 is very simple, consisting principally of B and Y series fragment ions. From it the sequence Pca-Asn-xLeu-xLeu-Gly-Asn-Asn-Val-Gly-Thr-Tyr can be readily deduced, where Pca is 5-oxo-

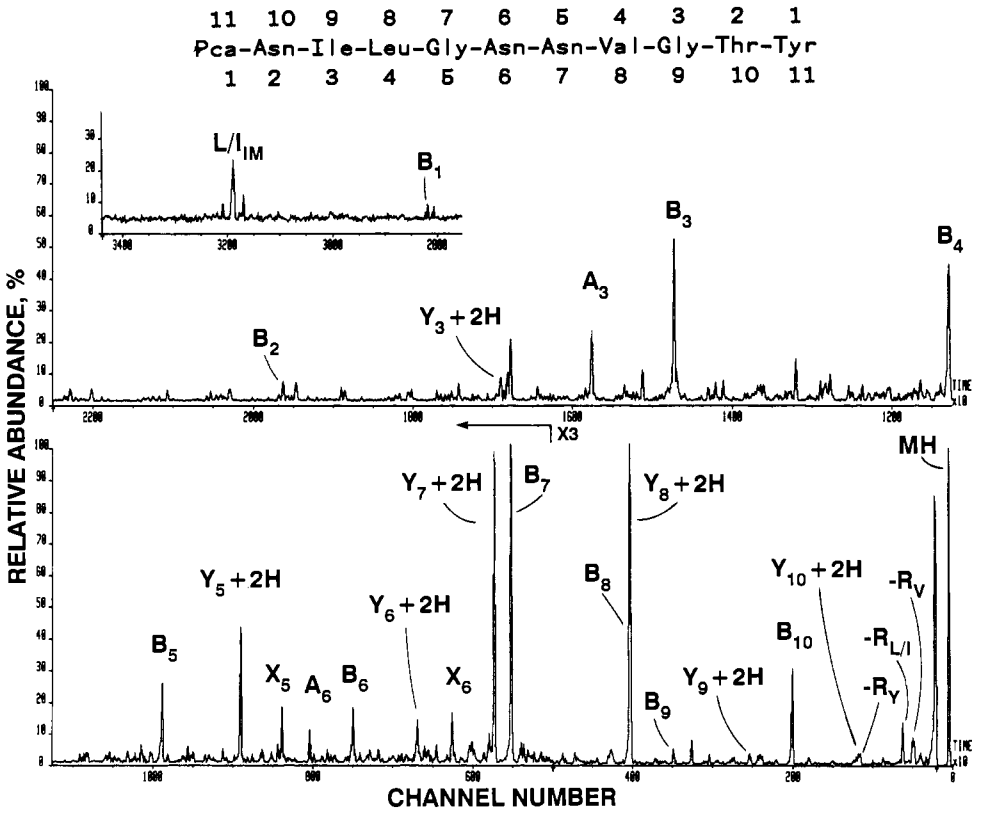

FIGURE 5: FAB tandem mass spectrum of the blocked N-terminal chymotryptic peptide of SBP. See legend to Figure 2 for experimental details and Figure 3 for an explanation of the notation used. Numbering above the printed sequence refers to C-terminal ions (X, Y series); that below the printed sequence refers to N-terminal ions (A, B series).

pyrrolidine-2-carboxylic acid (also known as pyrrolidonecarboxylic acid or pyroglutamic acid). The $(M + H)^+$ = 1322.9 peptide is presumed to have the same sequence, but to be longer by a phenylalanine (147 daltons) at the C-terminus. Two other chymotryptic peptides, $(M + H)^+$ = 939.7 and 605.3, were also analyzed by tandem MS. The sequence of the 605.3 peptide was found to be Gly-Ile-Arg-Ile-Phe, which represents residues 69–73, the overlap between the tryptic fragment $(M + H)^+$ = 1174.7 and the peptide immediately preceding it (Figure 4). The daughter ion spectrum of the $(M + H)^+$ = 939.7 peptide did not allow a complete sequence to be deduced, but a partial sequence Val-Ala-Gly-Glu-?-?-Gly-Gln-xLeu was postulated. Further purification of this peptide by HPLC and subsequent automated Edman degradation produced a sequence Val-Ala-Gly-Glu-Glu-His-Gly-Gln-Leu.

As further confirmation of the identity of the blocking group, the $(M + H)^+$ = 1175.7 chymotryptic peptide was treated with pyroglutamate aminopeptidase, an enzyme known to remove this blocking group from proteins (Podell et al., 1978). After HPLC cleanup of the digest, FABMS showed the expected ion at $(M + H)^+$ = 1064.9, corresponding to a peptide shortened by 111 daltons, the mass of the Pca group.

The final alignment of the peptides and the placement of the remaining residues were possible on the basis of mass spectral data (Figure 4). For example, a chymotryptic peptide of $(M + H)^+$ = 1102, after one step of manual Edman degradation, shifted mass to $(M + H)^+$ = 939. That shift, 163 daltons, is the in-chain mass of a tyrosine and placed Tyr-58 immediately before the chymotryptic $(M + H)^+$ = 939 peptide. Similarly, knowing that the N-terminal tryptic peptide was of $(M + H)^+$ = 2562.2 allowed the assignment of Arg-68, since the difference between 2562 and the sum of the known chymotryptic peptides [$(M + H)^+$ = 1175, 1322, 1102, 939] was 156 amu, the in-chain mass of Arg. The revised complete sequence of rat SBP is shown in Figure 6. The mature protein sequence defined in this work begins at residue 18, a Gln that is found cyclized to Pca. The sequence of a mouse prostatic SBP deduced from its cDNA (Mills et al., 1987) is shown for comparison.

The presence of Asn-linked carbohydrate on the rat protein was also established by MS using the carbohydrate mapping technique (Carr & Roberts, 1986). With this approach, the FABMS-derived peptide map of the tryptic digest of SBP was compared with that obtained after digestion with peptide:*N*-glycosidase F. This enzyme cleaves the β-aspartylglycosylamine linkage and converts the attachment-site Asn to Asp, which weighs 1 dalton more (Carr & Roberts, 1986; Tarentino et al., 1986). After treatment with the glycosidase a new signal at m/z 1345.6 was observed by FABMS of the tryptic digest. This ion corresponds to the tryptic peptide residues 87–98 in which carbohydrate has been released from Asn_{90} and this residue converted to Asp. We saw no evidence of a new ion at m/z 1198.5, which would correspond to the tryptic peptide residues 144–155. Rather, an ion was observed at m/z 1197.5, in which the Asn_{148} bears no carbohydrate. In the chymotryptic digest (no carbohydrate removed), a weak signal was observed at m/z 1190.9, corresponding to residues 146–156 with no carbohydrate attached. The weakness of the signal may indicate partial glycosylation, but we have no direct evidence for it.

Discussion

There is a high degree of homology (66% identical amino acids) between our corrected sequence of SBP and the sequence predicted for SBP from the mouse (Mills et al., 1987). Our sequence prior to Ile_{83} is in no way related to the published cDNA sequence (Chang et al., 1987) unless a block of 106 bases is inserted into the cDNA in the region immediately preceding that which codes for Ile_{83}. With the 106-base ad-

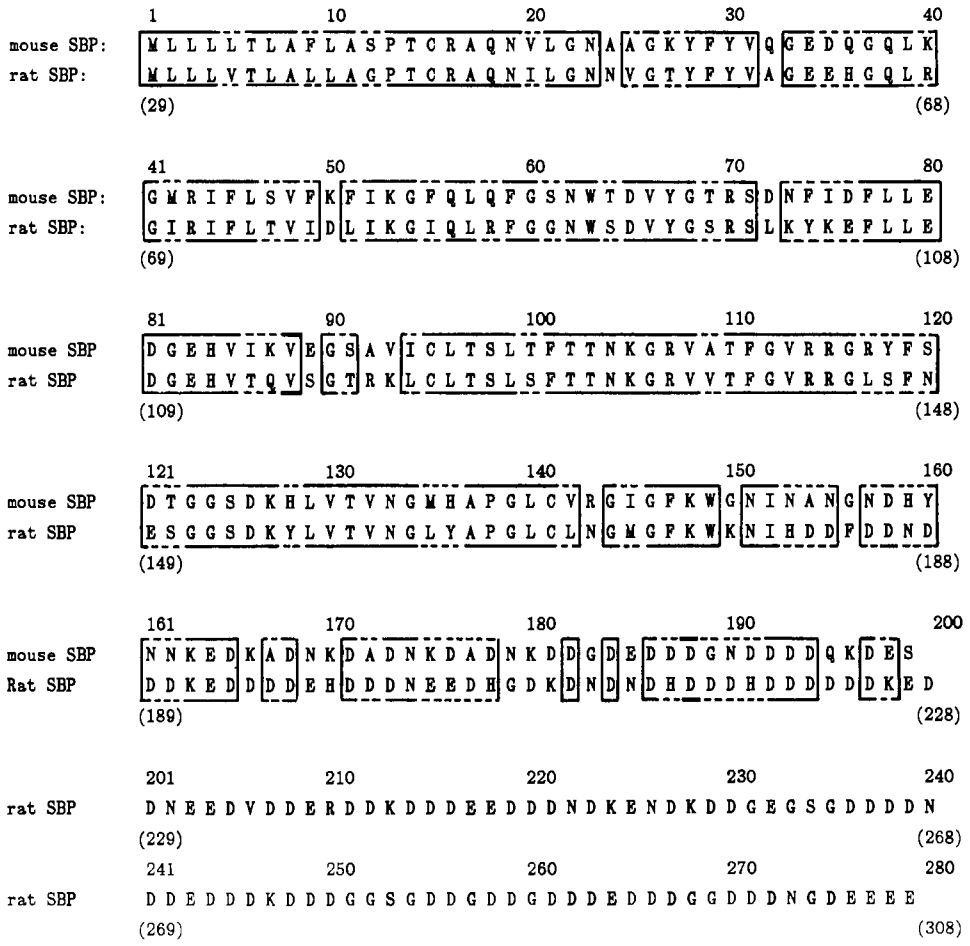

FIGURE 6: Complete corrected amino acid sequence of prostatic SBP from rat aligned with the predicted amino acid sequence of the same protein from mouse (Mills et al., 1987). The mature protein sequence has been renumbered to begin at residue 1, a Gln that is found cyclized to 5-oxopyrrolidine-2-carboxylic acid (Pca). Numbers in parentheses indicate the previous numbering scheme.

dition, which also causes a shift in the reading frame prior to Ile_{83}, the homology with the mouse sequence is maintained for 17 amino acids to the putative initiator Met and for an additional 87 bases into the 5′ noncoding region. Figure 6 compares the two predicted protein sequences.

In this paper we have described the synergistic use of MS, tandem MS, and conventional chemical sequencing strategies to determine the sequence of the first 36 residues of SBP, to confirm the sequence of ca. 90% of the previously sequenced portion of the molecule, and to establish the sites of attachment of Asn-linked carbohydrate; these procedures constitute an effective overall strategy for correcting deduced sequences of proteins and for defining the sites of attachment and chemical identity of posttranslational modifications in proteins. The instrumental, chemical, and biochemical approaches were used so as to take advantage of their unique strengths. Specifically, peptide mapping by FABMS was used to rapidly check the overall correctness of the deduced sequence and to define regions of the protein (and, therefore, of the cDNA) either where the sequence was in error or where a posttranslational modification may be present. Peptides observed in the FABMS data that did not fit the predicted sequence were then analyzed without extensive cleanup by tandem MS, and sequence information was obtained. Only in instances where ambiguities arose or only partial sequence data obtained was the peptide then isolated and purified for Edman degradation.

FABMS analysis of the mixture of tryptic peptides indicated that the protein sequence derived from the cDNA was clearly in error. The sequencing work described above, and comparison with mouse SBP cDNA (Mills et al., 1987), demonstrated that the discrepancy was due to a missing 106 bases in the rat cDNA. More than 20 different clones for rat SBP from two different cDNA libraries prepared from ventral prostatic mRNA have been analyzed. None of these clones contained the missing 106 bases of DNA when analyzed by sequencing or by hybridization to an oligomer specific for this region. However, SBP purified from the rat ventral prostate contained the amino acid sequence encoded by this missing sequence. Therefore, mRNA containing this information must be present in the cells, even though the cDNA corresponding to this mRNA was not found in our screening of our cDNA libraries. A number of possible explanations exist. It is possible that the SBP we have purified is encoded by a minor mRNA, perhaps a splicing variation at the time of mRNA processing. This seems unlikely, however, because of the high level of this SBP in the rat ventral prostate. It is also possible that the mRNA containing the missing sequence is not readily cloned because of difficulties in the synthesis of its cDNA or effects of the cDNA on the growth of host bacteria or on propagation of the vector. Finally, it is possible that some secondary structure in the mRNA prevents a part of the message from being reverse transcribed back into cDNA.

To our knowledge, the present paper is only the second well-documented case of a serious discrepancy between mRNA and cDNA. Wong et al. (1987) recently reported on a deletion of 142 base pairs in a cDNA clone for the male-specific isozyme (C-P-$450_{16\alpha}$) of testosterone 16α-hydroxylase in livers of 129/J mice. Interestingly, the missing 142 bases in this study as well as the missing 106 bases of DNA from the rat SBP cDNA appear to correspond to whole or nearly whole

exons of their corresponding genes (Wong et al., 1987; Mills et al., 1987). It is difficult to know how often and under what circumstances this occurs, but these findings reemphasize the need to check the protein sequences derived from cDNA predictions carefully with data from the protein itself. Unquestionably, the fastest and most efficient means of checking is by use of peptide mapping by FABMS (Biemann & Martin, 1987; Carr et al., 1988). With nanomolar amounts of protein and little or no peptide separation, the analysis of molecular weights of peptides generated by a tryptic digest can be performed in less than 1 h by FABMS. Information is obtained from all regions of the protein simultaneously, and C-terminal as well as N-terminal peptides are represented.

The use of tandem mass spectrometry for protein sequencing has not yet been fully evaluated, although it shows great promise (Biemann & Martin, 1987; Carr et al., 1988; Hunt et al., 1986; Crabb et al., 1986; Biemann & Scoble, 1987). The key to its utility will be if conditions can be found under which a predictable and complete set of sequence ions can be generated in the daughter ion spectrum of unknown peptides. We have not yet defined such conditions. Some peptides, e.g., the blocked N-terminus of SBP, behave extremely well and their sequences can be readily deduced. Others, such as the $(M + H)^+ = 1174.7$ peptide, give only partial sequences. In this case the presence of a contiguous sequence of amino acids each with side chains branched at the β-position resulted in a multitude of C-terminal-related secondary sequence ions. This factor, together with the relatively low abundance of some backbone fragments and the absence of others, impaired our ability to unambiguously define residues 5–7 of this peptide. An accurate amino acid composition can often suffice to complete the sequence, but in mixtures of unknowns, such compositional data are unavailable. If peptides are purified, compositions can be obtained, but one sacrifices one of the great advantages of tandem MS, namely, the ability to use MS-1 as the separating device. Computer programs have recently been described by Scoble et al. (1987) to assist in the interpretation of daughter ion mass spectra of peptides. These programs will undoubtedly be of great benefit, both in generating and impartially considering all plausible structures and in assigning some relative probability estimate to structures in which the mass spectra are not completely unambiguous, as was the case with our $(M + H)^+ = 1174$ and 939 peptides. As more experience is gained with the technique or if suitable derivatization reactions are developed to direct mass spectrometric fragmentation to desired bonds, this unpredictability may be overcome.

ACKNOWLEDGMENTS

The technical assistance of Lynette Miles, Anthony Jurewicz, Bart Frederick, Franco Duarte, and Don Winter is gratefully acknowledged.

REFERENCES

Barber, M., Bordoli, R. S., Elliot, G. J., Sedgwick, R. D., & Tyler, A. N. (1982) *Anal. Chem. 54*, 645A–657A.

Biemann, K., & Martin, S. A. (1987) *Mass Spectrom. Rev. 6*, 1–76.

Biemann, K., & Scoble, H. A. (1987) *Science (Washington, D.C.) 237*, 992–998.

Carr, S. A., & Roberts, G. D. (1986) *Anal. Biochem. 157*, 396–406.

Carr, S. A., Anderegg, R. J., Hemling, M. E., & Roberts, G. D. (1987a) 35th Annual Conference on Mass Spectrometry and Allied Topics, Denver, CO, Collected Abstracts, pp 542–543.

Carr, S. A., Roberts, G. D., Hemling, M. E., & Anderegg, R. J. (1987b) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 46*, 541.

Carr, S. A., Green, B. N., Hemling, M. E., Roberts, G. D., Anderegg, R. J., & Vickers, R. (1987c) 35th Annual Conference on Mass Spectrometry and Allied Topics, Denver, CO, Collected Abstracts, pp 830–831.

Carr, S. A., Hemling, M. E., & Roberts, G. D. (1988) in *Macromolecular Sequencing and Synthesis: Selected Methods and Applications* (Schlessinger, D. H., Ed.) pp 83–99, Alan R. Liss, New York.

Chang, C., Saltzman, A. G., Hiipakka, R. A., Huang, I.-Y., & Liao, S. (1987) *J. Biol. Chem. 262*, 2826–2831.

Clench, M. R., Garner, G. V., Gordon, D. B., & Barber, M. (1985) *Biomed. Mass Spectrom. 12*, 355–357.

Crabb, J. W., Armes, L. G., Carr, S. A., Johnson, C. M., Roberts, G. D., Bordoli, R. S., & McKeehan, W. L. (1986) *Biochemistry 25*, 4988–4993.

Hawke, D., Yuan, P. M., & Shively, J. E. (1982) *Anal. Biochem. 120*, 302–311.

Hemling, M. E., Carr, S. A., Capiau, C., & Petre, J. (1988) *Biochemistry 27*, 699–705.

Hiipakka, R. A., Chen, C., Schilling, K., Oberhauser, A., Saltzman, A. G., & Liao, S. (1984) *Biochem. J. 218*, 563–571.

Hunt, D. F., Yates, J. R., III, Shabanowitz, J., Winston, S., & Hauer, C. R. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 6233–6237.

Johnson, R. S., & Biemann, K. (1987) *Biochemistry 26*, 1209–1214.

Mills, J. S., Needham, M., & Parker, M. G. (1987) *Nucleic Acids Res. 15*, 7709–7724.

Morris, H. R., Panico, M., & Taylor, G. W. (1983) *Biochem. Biophys. Res. Commun. 117*, 299–305.

Naylor, S., Findeis, A. F., Gibson, B. W., & Williams, D. H. (1986) *J. Am. Chem. Soc. 108*, 6359–6363.

Podell, D. N., & Abraham, G. N. (1978) *Biochem. Biophys. Res. Commun. 81*, 176–185.

Scoble, H. A., Biller, J. E., & Biemann, K. (1987) *Fresenius' Z. Anal. Chem. 327*, 239–245.

Stults, J. T., & Watson, J. T. (1987) *Biomed. Environ. Mass spectrom. 14*, 583–586.

Tarentino, A. L., Gomez, C. M., & Plummer, T. H., Jr. (1985) *Biochemistry 24*, 4665–4671.

Wong, G., Kawajiri, K., & Negishi, M. (1987) *Biochemistry 26*, 8683–8690.