Application of Mass Spectrometry to Protein Chemistry. I. Method for Amino-Terminal Sequence Analysis of Proteins*

William R. Gray and Ursino E. del Valle

ABSTRACT: A method is described for analyzing the amino-terminal sequences of proteins by mass spectrometry. It is applicable to as little as 1 mg of protein, and does not depend on the presence of a free α-amino group. The protein is acetylated, so as to block amino groups, and digested with a protease; the resulting mixture of peptides is permethylated. The amino-terminal peptide is then extracted into $CHCl_3$, and its mass spectrum is examined at a resolving power of 1000–2000. Simple fragmentation patterns are observed that are easily interpreted in terms of amino acid sequence.

Mass spectrometry has found increasing use as a tool for analyzing the sequence of amino acids in peptides (Biemann *et al.*, 1966; Shemyakin *et al.*, 1966; Thomas *et al.*, 1968). Apart from a few instances where it has been applied to peptides isolated from proteins (Geddes *et al.*, 1969) the method has been used principally upon synthetic peptides of known structure or on peptide antibiotics. We are attempting to use mass spectrometry as a routine method for protein structural studies, and have devised a procedure for analyzing the amino-terminal sequences of peptides and proteins.

The spectra are simplest to interpret when the protein has a free α-amino group, but the method can be applied to proteins whose amino terminals are blocked (*e.g.*, by *in vivo* acetylation or by pyrrolidonecarboxylic acid).

Principle of the Method

The protein is first acetylated, using an equimolar mixture of acetic anhydride and perdeuterioacetic anhydride (van Heijenoort *et al.*, 1967), thus blocking and labeling the amino terminus, if it is not naturally blocked. Digestion with a protease of broad specificity (such as subtilisin or pronase) then produces a mixture of peptides, of which the amino-terminal peptide is the only one not possessing a free amino group. The mixture of peptides is then permethylated using methyl iodide and methylsulfinyl carbanion (Hakomori, 1964; Thomas, 1968). This converts the amino-terminal peptide into a neutral, very hydrophobic derivative which readily extracts into chloroform. All other peptides are converted into quaternary ammonium salts which partition largely in the aqueous phase. Mass spectra of the crude extraction products are examined at a resolving power of 1000–2000, using a direct insertion probe at a temperature of 150–300°.

One of three basic spectral types can normally be expected. (1) Proteins having a free amino terminus show twin peaks separated by three mass units for all fragments containing the terminal group. These correspond to $CH_3 \cdot CO-$ and $CD_3 \cdot CO-$, and make interpretation of the spectra very simple. (2) Proteins containing a naturally acetylated amino terminus give singlet peaks at the principal cleavage points. (3) Proteins containing pyrrolidonecarboxylic acid as their amino terminus also give singlet peaks, including a very intense peak at 98 (C_5H_8NO).

N,O-Permethylation of peptides, introduced by Lederer's group (Das *et al.*, 1967, 1968; Thomas *et al.*, 1968), has the principal advantages of increasing the volatility of the peptides, and of simplifying the fragmentation pattern. The principal cations formed, apart from fragments of amino acids, are usually the sequence-determining acyl peptide fragments. This makes interpretation of the spectra fairly straightforward, even at low resolution, and even when no isotopic marker can be put onto the α-amino group.

Example of Method

Acetylation and Digestion. Protein (4 mg, human κ-type Bence-Jones protein Mil (Dreyer *et al.*, 1967)) was dissolved in 250 μl of aqueous pyridine (10%, v/v), and a further 200 μl of pyridine was added. To this was added 60 μl of an equimolar mixture of acetic anhydride and perdeuterioacetic anhydride, with rapid stirring. After 5 min a further 60 μl of the same mixture was added, and the reaction mixture was allowed to stand for 15 min at room temperature. Acetone (2.5 ml) was then added, causing the protein to precipitate as white floccules; these were allowed to settle for 30 min, and were then spun into a pellet, the supernatant being discarded. The pellet was suspended in 200 μl of 0.1 M ammonium bicarbonate buffer, pH 8.5, and 0.3 mg of subtilisin was added. Digestion was allowed to proceed for 5 hr at 37°, after which time the solution was frozen and lyophilized in a desiccator containing P_2O_5 and NaOH pellets.

Permethylation. Methylsulfinyl carbanion solution was prepared as follows. Sodium hydride dispersion (100 mg, 57% by weight, dispersed in oil) was rinsed with three 1-ml portions of dry ether. The sodium hydride was then dried in a gentle stream of nitrogen, and 1 ml of Me_2SO was added;

* From the Department of Biology, California Institute of Technology, Pasadena, and the Space Sciences Division, Jet-Propulsion Laboratory, Pasadena. *Received December 29, 1969.* This work was supported by U. S. Public Health Service Grant GM-06965, and the President's Fund, California Institute of Technology. Inquiries should be sent to W. R. Gray. Address after July 1970: Biology Department, University of Utah, Salt Lake City, Utah.

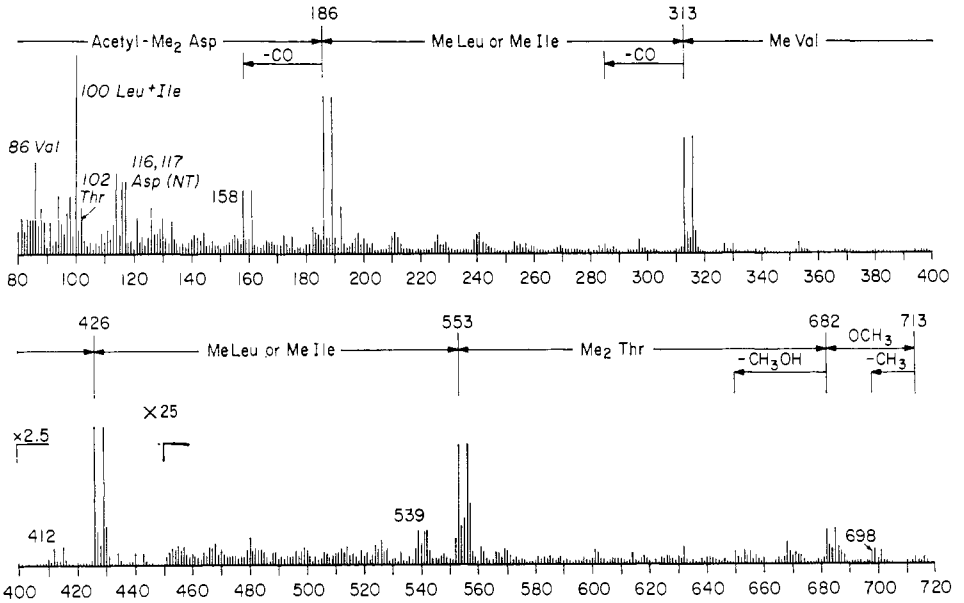

FIGURE 1: Mass spectrum of terminal peptide derivative from HBJ (3 Mil), omitting peaks below *m/e* 80. See text for experimental conditions.

a nitrogen atmosphere was maintained through this step until the extraction procedure. Traces of water in the Me_2SO gave rise to a brisk effervescence, which soon subsided. The mixture was then heated at 85° for 10 min, after which time hydrogen evolution had ceased, and the solution was a pale yellow color.

Dry Me_2SO (0.2 ml) was added to the peptide mixture, and the tube was flushed with nitrogen. To this was added 0.2 ml of the methylsulfinyl carbanion solution, and the mixture was allowed to stand for 30 min at room temperature. Three portions of 100 μl of methyl iodide were added, the first addition being accompanied by vigorous evolution of heat. Reaction was allowed to proceed for 1 hr at room temperature.

Chloroform (0.5 ml) was then added, and the mixture was extracted six times with 2 ml of water to remove Me_2SO, sodium iodide, and as much as possible of the quaternary ammonium iodide derivatives of the peptides. The chloroform layer was dried in a stream of nitrogen, leaving a film of yellow oil.

Mass Spectrometry. The oil was redissolved in 50 μl of chloroform, transferred to a direct insertion probe, and the chloroform was removed in a stream of nitrogen. Spectra were recorded on an AEI Model MS9 at several temperatures up to 200°, using an ionizing voltage of 70 eV. The resolving power used was approximately 1500 (10% valley).

Results

Figure 1 shows the spectrum recorded with a probe temperature of 170°. The sequence of the first five residues (Asp-Ile-Val-Leu-Thr-) is defined by the principal CH_3/CD_3 doublets, although it is not possible to decide clearly whether residues two and four are leucine or isoleucine. The major sequence-determining doublets are given (in order) in Table I where all amino acids listed are present as their *N*-methyl derivatives, and aspartic acid and threonine are also O methylated in their side chains. At low temperatures (80–100°) the doublet at 158/161 was much higher than that at 186/189, and probably was due to *N*-acetylserine, rather than to *N*-acetylaspartic acid; the only other doublets not attributable to the main sequence also correspond to acetyl serine fragments (126/129 and 130/133). Apart from this, changing the temperature varied the intensity of the spectrum, but had little effect upon its principal features.

A number of strong individual peaks are also present, especially below *m/e* 200. Several of these are attributable to the amino-terminal peptide, but some are also found with several other proteins and probably come from particularly hydrophobic peptides which were not completely separated by the extraction. Strong peaks at *m/e* 121 (C_8H_9O) and 192 ($C_{11}H_{12}O_3$) are found in spectra of tyrosine-containing peptides under these conditions of derivatization. Two other peaks have been encountered at *m/e* 112 and 144 with several proteins, though they are relatively minor in the spectrum shown in Figure 1. They have also been found with peptides containing amino-terminal serine, and may represent *N*-formyl-*N*-methyldehydroalanyl ($C_5H_6NO_2$) and *N*-formyl-

TABLE I: Mass Spectral Data for CH_3 and CD_3 Doublets.

me		
CH_3	CD_3	Amino Acid Residue
158	161	Acetyl-Asp-, minus CO
186	189	Acetyl-Asp-
313	316	Acetyl-Asp-Ile-
426	429	Acetyl-Asp-Ile-Val-
553	556	Acetyl-Asp-Ile-Val-Leu-
650	653	Acetyl-Asp-Ile-Val-Leu-Thr-, minus CH_3OH
682	685	Acetyl-Asp-Ile-Val-Leu-Thr-
713	716	Acetyl-Asp-Ile-Val-Leu-Thr-OCH_3

N,O-dimethylseryl ($C_6H_{10}NO_3$), respectively. Another commonly encountered peak at *m/e* 144 comes from the N-methylated side chain of tryptophan ($C_{10}H_{10}N$). The resolution used in this work was not sufficient to establish elemental compositions of the major peaks found at these masses, so the actual source of them remains unknown.

Other important peaks which are attributable to the amino-terminal peptide include the ketiminium fragments ($CH_3N^+H{=}CHR$) typical of the individual amino acids. These are at *m/e* 86 (valine, R = C_3H_7), 100 (leucine and isoleucine, R = C_4H_9), 102 (threonine, R = $-CH(CH_3)OCH_3$), and 116 (aspartic acid, R = $-CH_2COOCH_3$). The latter is accompanied by an equivalent peak at 117 ($CH_3N^+D{=}CH\text{-}CH_2COOCH_3$), since the nitrogen derives its proton equally from CH_3CO- and CD_3CO-. None of these peaks gives rise to confusion with sequence-determining peaks.

Two potentially confusing peaks occur at *m/e* 98 and 114. The former mass is diagnostic for peptides having pyrrolidonecarboxylic acid as their terminal group, the latter for acetylglycine. At *m/e* 98 the peak is actually a doublet, the component of lower mass being indistinguishable at this resolution from the peak due to *N*-methylpyrrolidonyl (C_5H_8NO). Several possible sources of such an ion will be discussed later; the source of the second component is not known. Different proteins gave spectra containing the two components in widely varying ratios and amounts. The peak at *m/e* 114 is readily resolved from that due to *N*-acetyl-*N*-methylglycyl ($C_5H_8NO_2$), its mass being approximately 0.08 amu higher.

CH_3/CD_3 doublets also occur as satellites at +14 or −14 mass units from the main sequence-determining doublets. These are probably due to two main causes. (1) Methylation of the β-carbon of the aspartyl side chain in 5–10% yield, giving rise to minor doublets at *m/e* 172/175, 200/203, 327/330, 440/443, and 567/570. (2) Incomplete methylation at each of the peptide nitrogens following the sterically hindered isoleucine and valine residues, giving rise to satellite doublets at *m/e* 412/415 (approximately 10% of that at 426/429) and 539/542 (approximately 20% of that at 553/556).

Discussion

This approach uses the permethylation system of Thomas *et al.* (1968) for a combination of reasons. Primarily, it enables us to convert the amino-terminal peptide into a form which can be separated very readily from the derivatives of all other peptides. The selectivity of this is evident from Figure 1, where the spectrum is overwhelmingly that of the amino-terminal peptide, although it represented only about 2% of the molecule. Thomas *et al.* (1968) noted also the increased volatility of the permethylated peptides, and the promotion of a simple cleavage pattern. This also is obvious in Figure 1, where almost all of the CH_3/CD_3 doublets (fragments which contain the amino terminal) are attributable to cleavage at peptide bonds. It is likely that part of the selectivity of the method is due to the high volatility of the amino-terminal peptide derivative compared with that of any other peptide derivative which might remain in the chloroform phase during extraction.

Deduction of the amino acid sequence from the spectrum is completely straightforward in the example shown, except that there is no direct means of distinguishing between leucine and isoleucine residues. Some indirect evidence is available, however, in the form of incomplete methylation of the peptide nitrogen following the sterically hindered isoleucine residue. It remains to be seen whether this will be a reliable diagnostic aid. In some cases the possibility arises that satellite peaks are not due to incomplete methylation, but to the peptide's containing partial replacement of an amino acid by a lower homolog. Similarly, aspartic acid is somewhat prone to C methylation in the side chain, and could be mistaken for a glutamic acid residue. These ambiguities can be avoided by using CD_3I as the methylating agent (Das *et al.*, 1968), and we plan to use it routinely.

The appearance of amino-terminal serine in the spectrum is probably genuine. This protein contains 30 serine residues, which could become partially labeled by at least three processes: (1) partial hydrolysis of labile peptidyl serine bonds; (2) N–O peptidyl transfer followed by acetylation of the serine amino group, and subsequent hydrolysis of the peptidyl ester bonds; (3) acetylation of serine hydroxyls with subsequent O–N acyl shift following digestion at the peptidyl serine bonds. It was not possible to trace a sequence out from the acetyl-serine, suggesting the presence of many sequences in low yields. The same phenomenon has been seen with other serine-rich proteins, but not with cytochrome *c* which contains no serine.

Several other proteins have been investigated by this procedure, and satisfactory spectra were obtained from cytochrome *c* (1 mg), insulin (1 mg), ribonuclease (2 mg), rennin (2 mg), and lysozyme (2 mg). With cytochrome *c* no significant pattern of doublets was observed, but the sequence acetyl-Gly-Asp-Val-Glu- was indicated by a series of major peaks at *m/e* 114, 257, 370, and 527, showing that the protein has a naturally acetylated amino terminus. The peak at *m/e* 114 had two main components, differing in mass by approximately 0.08 amu, the component of lower mass being due to *N*-acetyl-*N*-methylglycyl ($C_5H_8NO_2$). Insulin, as expected, gave two sequences, both of which were interpretable because of the differential volatility of the peptides from the two chains. Four residues of the A chain were traceable (Gly-Ile/Leu-Val-Glu-), while two residues of the B chain (Phe-Val-) were clear, and a third (Asn) could be tentatively assigned. The sequences found for the other three proteins were all in agreement with published sequences found by conventional methods: ribonuclease (Lys-Glu-Thr-Ala-), rennin (Gly-Glu-Val-Ala-), and lysozyme (Lys-Val-Phe-). In none of these examples was a peak observed for the intact peptide containing the ester group at its C terminus, though a weak parent peak was observed in the spectrum shown in Figure 1. This in no way hampers the structural analysis, because the deductions are made by working from lower masses toward the higher masses, and not *vice versa* as is usual with most classes of compounds.

A human λ-type Bence–Jones protein (HBJ8 (Hood *et al.*, 1966)) containing a pyrrolidonecarboxylic acid group was also examined (1.4 mg). The first two residues of the sequence (Glp-Ser-) were unambiguously assigned, but several possible branches were available beyond this point, including the correct sequence (Glp-Ser-Ala-Leu). We are trying to gain wider experience with pyrrolidonecarboxyl proteins, to give a broader basis for interpretation. It has already been mentioned that the peak at *m/e* 98 (C_5H_8NO) which is characteristic of pyrrolidonecarboxylic acid was obtained with

some of the other proteins also, although no sequence of amino acids could be traced out in these cases. This peak could arise by several processes. (1) Partial cyclization of amino-terminal glutamic acid and glutamine, as these are released during proteolysis; this is the source most likely to give rise to ambiguities, because some peptides cyclize quite rapidly and efficiently. (2) Partial cleavage and cyclization at internal residues of glutamic acid and glutamine during permethylation; this was reported to occur with the older methods using silver oxide, but is not significant when the methylsulfinyl carbanion method is used (Thomas *et al.*, 1968). (3) Elimination of methanol or dimethylamine from the ketiminium fragments of glutamic acid or glutamine, respectively; this would occur whether the residues were terminal or internal. (4) An ion of the same composition could also arise by elimination of CO and CH_3OH from *N*-acetyl-*N*,*O*-dimethylseryl ion. This should give a corresponding CD_3 peak at 101 also; only minor amounts of such a doublet have been observed.

Twelve of the twenty common protein amino acids have been encountered in application of the method so far, plus pyrrolidonecarboxylic acid (glycine, alanine, valine, leucine, isoleucine, phenylalanine, lysine, aspartic acid, glutamic acid, asparagine, serine, and threonine). Our experience with small peptides, plus the work of Thomas *et al.* (1968), leads us to feel confident that the method will also work without complications for proline, tryptophan, tyrosine, and glutamine. This leaves only cystine, methionine, arginine, and histidine posing special problems. We are attempting to extend the method to these four amino acids also. It should be borne in mind that some modification procedures which are suitable for pure peptides of known composition, are quite unsuited for proteins. An example is the conversion of arginine into ornithine by hydrazinolysis (Thomas, 1968). Extensive hydrolysis and hydrazinolysis of many peptide bonds occurs even under mild conditions, and gives rise to a confusing array of terminal groups when applied to proteins.

A comparison between this procedure and the phenyl isothiocyanate degradation (Edman, 1956) is pertinent. The latter is completely dependent upon the presence of a free α-amino group, while the mass spectrometric method is applicable to blocked proteins also. However, with rigorously purified reagents, and precise attention to detail, the stepwise degradation of a protein can be carried through many cycles, especially with the aid of an automatic "sequenator" (Edman and Begg, 1967). By its nature the mass spectrometric method is unlikely to give long sequences, partly due to lack of volatility of the peptides, and partly due to the characteristics of the proteases used for digestion. It is interesting to note that the characteristic fragmentation of peptides in the mass spectrometer is due in part to a stepwise degradation from the carboxyl terminus (Shemyakin *et al.*, 1966).

It is not yet clear what the practical sensitivity of the method will be. This will probably depend less upon the absolute sensitivity of the mass spectrometer (which is very high) than upon the ability to maintain a low background against which to identify the sample. We have used samples in the range 50–200 nmoles, and feel that a considerable reduction in those amounts should be possible. It thus promises to be a valuable tool for characterizing small amounts of protein, especially since a single procedure is applicable to proteins containing either free or blocked amino termini. Since a short sequence (3–5 amino acids) is defined, rather than a single end group, one may use the method to investigate possible genetic relationships among proteins, even when very small amounts are available. With simple mixtures of proteins the method may reveal the terminal sequences of more than one component, as was the case with the two chains of insulin; with more complex mixtures it will be advantageous to use a combination gas chromatograph–mass spectrometer to separate and analyze each component. Such a combination could prove to be a powerful tool for structural analysis of proteins if used in conjunction with new strategies for avoiding the complicated peptide fractionations used at present (Gray, 1968).

Acknowledgment

We wish to thank Dr. H. G. Boettger of the Space Sciences Division, Jet Propulsion Laboratory, for making his facilities and experience available to us.

References

Biemann, K., Cone, C., Webster, B. R., and Arsenault, G. P. (1966), *J. Amer. Chem. Soc. 88*, 5598.

Das, B. C., Gero, S. D., and Lederer, E. (1967), *Biochem. Biophys. Res. Commun. 29*, 211.

Das, B. C., Gero, S. D., and Lederer, E. (1968), *Nature 217*, 547.

Dreyer, W. J., Gray, W. R., and Hood, L. E. (1967), *Cold Spring Harbor Symp. Quant. Biol. 32*, 353.

Edman, P. (1956), *Acta Chem. Scand. 10*, 761.

Edman, P., and Begg, G. (1967), *Eur. J. Biochem. 1*, 80.

Geddes, A. J., Graham, G. N., Morris, H. R., Lucas, F., Barber, M., and Wolstenholme, W. A. (1969), *Biochem. J. 114*, 695.

Gray, W. R. (1968), *Nature 220*, 1300.

Hakomori, S. I. (1964), *J. Biochem. (Tokyo) 55*, 20.

Hood, L. E., Gray, W. R., and Dreyer, W. J. (1966), *J. Mol. Biol. 22*, 179.

Shemyakin, M. M., Ovchinnikov, Yu. A., Kiryushkin, A. A., Vinogradova, E. I., Miroshnikov, A. I., Alakhov, Yu. B., Lipkin, V. M., Shvetsov, Yu. B., Wulfson, N. S., Rosinov, B. V., Bochkarev, V. N., and Burikov, V. M. (1966), *Nature 211*, 361.

Thomas, D. W. (1968), *Biochem. Biophys. Res. Commun. 33*, 483.

Thomas, D. W., Das, B. C., Gero, S. D., and Lederer, E. (1968), *Biochem. Biophys. Res. Commun. 32*, 199.

van Heijenoort, J., Bricas, E., Das, B. C., Lederer, E., and Wolstenholme, W. A. (1967), *Tetrahedron 23*, 3403.