Determination of the Sequences of Protein-Derived Peptides and Peptide Mixtures by Mass Spectrometry

By HOWARD R. MORRIS AND DUDLEY H. WILLIAMS

University of Cambridge Chemical Laboratory, Cambridge CB2 1EW, U.K.

AND RICHARD P. AMBLER

Department of Molecular Biology, University of Edinburgh, Edinburgh, U.K.

(Received 25 May 1971)

Micro-quantities of protein-derived peptides have been converted into *N*-acetylated permethyl derivatives, and their sequences determined by low-resolution mass spectrometry without prior knowledge of their amino acid compositions or lengths. A new strategy is suggested for the mass spectrometric sequencing of oligopeptides or proteins, involving gel filtration of protein hydrolysates and subsequent sequence analysis of peptide mixtures. Finally, results are given that demonstrate for the first time the use of mass spectrometry for the analysis of a protein-derived peptide mixture, again without prior knowledge of the protein or components within the mixture.

In recent years, considerable effort has been devoted to the study of the application of mass spectrometry to the amino acid sequence analysis of oligopeptides and achievements in the field have been reviewed (Shemyakin, 1968; Lederer, 1968; Shemyakin *et al.* 1970). The success of such studies depends critically on the volatility of the peptide, since the molecules are examined in the vapour phase within the mass spectrometer.

A significant advance as regards the size limit of peptides studiable by mass spectrometry was effected by the introduction of the technique of *N*-alkylation of the peptide bond (Das, Gero & Lederer, 1967). The inclusion of this permethylation step in derivative preparation has also led to a simplification in the spectra produced, since the resulting derivatives fragment almost entirely at the peptide bonds.

Considerable emphasis has also been placed (Biemann, Cone, Webster & Arsenault, 1966; Senn, Venkataraghavan & McLafferty, 1966) on the application of high-resolution mass spectrometry together with computer-aided data interpretation to the problem of sequence analysis of peptides.

Despite these advances in the preparation of derivatives and instrumentation, published work has dealt almost entirely with the examination of relatively large quantities of either synthetic peptides of known sequence or peptide antibiotics (e.g. Prox, Schmid & Ottenheim, 1969; Thomas, Lederer, Bodanszky, Izdebski & Muramatsu, 1968b). Two exceptions to this have been work on

the protein silk fibroin (Morris, Geddes & Graham, 1968; Geddes *et al.* 1969), and the peptide hormone feline gastrin (Agarwal, Kenner & Sheppard, 1969), and in both these cases the peptides examined were of partially or completely known sequence. In addition, a mixture of two octadecapeptides from normal immunoglobulin λ -chains has been examined by mass spectrometry (Franek, Keil, Thomas & Lederer, 1969) and it was possible to demonstrate the sequence of the first ten residues from the *N*-terminus, and that the heterogeneity (approx. 1:1 mixture of two components) was due only to the changed amino acid composition at the residue second from the *N*-terminus. Nevertheless, 5mg of the peptide was used to obtain this information.

In making the transition from the study of synthetic model peptides and peptides available in relatively large quantities, to peptides derived from the enzymic hydrolysis of proteins, a number of problems are apparent. In particular, the quantity of peptide available for study, and also the standard of purity of the peptide, are normally decreased. Both these factors are a result of the isolation procedures by which a single peptide is obtained from an enzymic hydrolysate. The above factors may affect the study in two ways. First, the intensity of the spectrum, which is dependent on sample size, may be decreased to a point at which interpretation of the full sequence of the peptide is not possible. Secondly, ions derived from the impurity present in the sample may 'mask' the

sequence ions of the peptide, and thus prevent interpretation. This was demonstrated in the aforementioned work on silk fibroin.

We have been engaged in studies designed to assess the applicability of the mass spectrometric method to the full sequential analysis of an 'unknown' protein. This work has involved the determination of the sequences of isolated peptides and peptide mixtures, all samples being derived from enzymic hydrolysis of proteins.

We believe mixture analysis to be the distinct and most important advantage that mass spectrometry affords over other methods of sequential analysis. Use of this strategy minimizes the need for laborious isolation and purification procedures, which are the rate-determining steps in classical techniques for the determination of sequences. In addition, the normal process of amino acid analysis before sequence determination has little or no meaning when working with peptide mixtures. However, this lack of knowledge of amino acid composition, or length of an individual component peptide in a mixture, places extra importance on the correct interpretation of the spectra of small amounts of unknown protein-derived peptides.

It is therefore essential, before undertaking mixture analysis, to ascertain the problems involved in examining 'unknown' protein-derived peptides. In the present paper we describe the results obtained and the technique used in the examination of a selection of seven protein-derived 'unknown' peptides. We suggest a strategy, involving mixture analysis, for the determination of the sequence of an 'unknown' protein, and finally demonstrate the sequence of a peptide mixture derived from an 'unknown' protein hydrolysate. 'Unknown' in the present context is used to signify peptides of which the amino acid compositions and sequences were not known to those carrying out the mass spectrometric determination. It is emphasized that the mass-spectrometrically established sequences are reported exactly as determined, even to the extent of any residual ambiguities of sequence. It was important for us to work with peptides of which the sequences had been determined by classical methods, thus allowing a cross-check of the results at this early stage of development of the technique.

EXPERIMENTAL

A selection of peptides obtained by enzymic hydrolysis of *Staphylococcus aureus* penicillinase (Ambler & Meadway, 1969) has been studied. The peptides were available in quantities varying from 0.1 μ mol to 1.0 μ mol, and at a purity level acceptable for classical sequence studies (Ambler & Brown, 1967). Peptide C3a was obtained from a cytochrome (Ambler, 1963).

We have found, when dealing with very small quantities of peptide, that care must be taken in the choice of derivatives. The technique used in both the preparation of derivatives and mass spectrometric analysis is also important.

Preparation of derivatives. Of the selection of *N*-terminal blocking groups which is available, the *N*-acetyl derivative is to be preferred, both from the viewpoint of its volatility and ease of preparation.

Acetylation was performed by dissolving the dry peptide residue in an excess of the reagent acetic anhydride-methanol (1:4, v/v) (Thomas, Das, Gero & Lederer, 1968a). The reaction was allowed to proceed at room temperature for 3 h in a stoppered tube, and terminated by evaporation of reagents by means of a mechanical vacuum pump. This procedure results in the acetylation of basic free amino groups in the peptide; hydroxyl groups (e.g. of tyrosine, serine or threonine) are not acetylated.

We have chosen the Hakamori reaction (Hakamori, 1964) for permethylation, involving the generation of a methyl sulphanyl carbanion base by using dimethyl sulphoxide and sodium hydride. *N*-Permethylation of peptide derivatives by using this reaction has been described (Vilkas & Lederer, 1968; Thomas, 1968). Although this is less convenient than the Coggins & Benoiton (1968) technique, it has been shown that the latter can lead to extensive *C*-methylation of glycine and aspartic acid (Thomas, 1969).

We have found that to obtain good base generation, it is necessary to keep the reagent dry by periodic distillation from calcium hydride under reduced pressure.

The permethylation procedure used was as follows: a base was prepared by heating sodium hydride in dimethyl sulphoxide (50 mg/ml) at 90°C for 15 min, or until the evolution of H₂ ceased. An appropriate quantity of the base (slight excess) was added, when cool, to the acetylated peptide (0.1–1.0 μ mol) dissolved in 0.1 ml of dimethyl sulphoxide, followed by an excess of methyl iodide. The reaction was allowed to proceed for 1 h in a stoppered tube after which the permethylated product was isolated by dilution with water and extraction with chloroform (1 ml). The chloroform layer was washed twice with water, and evaporated *in vacuo*. In the above procedure it is important that the reactions are carried out in the same tube, thereby avoiding losses due to transfer. The method leads to methylation of amide nitrogen [$-\text{CONH}_2 \rightarrow -\text{CON}(\text{CH}_3)_2$; $-\text{CONH}- \rightarrow -\text{CON}(\text{CH}_3)-$], hydroxyl groups and carboxylic acid functions.

The above procedures may be applied successfully to all peptides except those containing the sulphur-amino acids, histidine and arginine. Modification reactions are available which, when used before the above procedure, render all peptides amenable to mass spectrometric analysis (see Vliegthart & Dorland, 1970; Lenard & Gallop, 1969; Lederer, 1968; Shemyakin *et al.* 1967; Thomas *et al.* 1968a). The problem of identifying the peptides that contain these amino acids is dealt with in the section on mixture analysis.

Mass spectrometric analysis. All mass spectra were recorded on an AEI MS902 double-focusing instrument, operating at an accelerator voltage of 8 kV and at an electron beam energy of 70 eV. The operating source pressure was typically in the range $(0.2\text{--}1.0) \times 10^{-6}$ mmHg.

The instrument may be operated in two basic modes, low and high resolution. The distinction is ill-defined, but essentially information as to the atomic composition of ions may be obtained by using the high-resolution mode. This setting also lends itself to computer interpretation and data reduction and hence automation, and its use in the determination of peptide sequences has been advocated and demonstrated often. Unfortunately, however, the tuning required to obtain high resolution also reduces the sensitivity of the instrument by a factor of about 10 to 100. Clearly therefore, it is not as yet the most desirable approach for use on small amounts of protein-derived peptides, since more sequence information may be obtained manually at low resolution. The resolution setting for all the peptides in the present paper was 1200, which implies that two adjacent peaks 1200 mass units and 1201 mass units will be resolved by 90% of their peak heights.

A complete transfer of sample to the instrument was effected as follows. The sample was dissolved in 20 μ l of chloroform, transferred by means of a fine capillary and slowly dried on to the quartz tip of the direct insertion probe. The process of dissolution of tube contents and transfer was repeated twice, to ensure complete removal of the sample. The probe was then introduced into the ion source of the mass spectrometer by use of the standard vacuum-lock system.

The separation of the peptide from impurities present in the sample was achieved by the use of a steady temperature gradient, which was effected by fractional movements of the probe tip into the ion source. Thus, as the probe tip is slowly moved towards the region of the electron beam, the temperature of the sample is steadily increased and the more volatile components evaporate first. The spectrum was continuously monitored on the oscilloscope until peptide volatilization occurred. At this point, the temperature gradient was rapidly increased, and the spectrum recorded.

RESULTS

In the following discussion, plausible structures for certain structurally useful ions are indicated where the constituent atoms are simply related to the original amino acid.

The structures of the amino acids derived from the peptides after acetylation and permethylation, are given in Table 1, together with relevant mass numbers and associated ions. The sulphur-containing amino acids have been omitted from the table, since they have not been dealt with in this study, and require treatment before permethylation (Thomas *et al.* 1968a).

Peptide C69. This peptide gave rise to a spectrum (Fig. 1 shows the region above m/e 100) which in its simplicity is typical of those obtained in the present study and which could be readily interpreted as follows. The abundant ions at m/e 170, 142 and 100 indicate *N*-terminal leucine. (No distinction can yet be made between leucine and isoleucine mass spectrometrically.) We have noted also that m/e 171 is always present in the spectra of lysine-

containing peptides, and at this early stage of the interpretation we may expect a lysine residue in the sequence. Similarly, the presence of m/e 98 in the spectrum is indicative of a glutamic acid or glutamine residue in the peptide; this ion is usually particularly abundant if the glutamic acid or glutamine residue is at the *N*-terminus. The ion is derived from a small amount of cyclization at the glutamic acid residue, accompanied by chain cleavage, and loss of carbon monoxide to give m/e 98 (see structure I). Although these ions provide good confirmatory evidence of the presence of a certain amino acid in the sequence, only the sequence ions can be fully relied on to give the exact amino acid mass differences present and hence the sequence. The exception to this is the presence of m/e 121 and m/e 192 in a peptide spectrum, as here. These, if present, are always very abundant ions, and considerable experience of tyrosine-containing peptides has shown that m/e 121 (see structure II) is indicative of the tyrosine side chain, but moreover m/e 192 fixes the tyrosine in the *C*-terminal position (Morris, 1970). The m/e 192 ion (structure III) is not given by tyrosine residues elsewhere in the chain, as it contains the *C*-terminal blocking group.

The next most abundant ion in the spectrum is at m/e 327, and corresponds to a Leu-Glu- sequence. The abundant ion after m/e 327 is at m/e 497, a difference of 170 mass units, and corresponds to a glutamine unit. The sequence so far is thus Leu-Glu-Gln. This peptide clearly demonstrates the ease of differentiation between an acid and its amide, which if, as here, present in the same peptide, can cause difficulty by classical techniques. The abundant ion after m/e 497 is at m/e 610, the difference corresponding to a valine unit, giving the sequence: Leu-Glu-Gln-Val. After this, a sharp drop in intensity occurs, the next sequence ion being 97 mass units higher at m/e 707. The mass difference corresponds to proline, and this intensity loss is common after proline residues. The next group of ions in the spectrum lie around m/e 900. One difficulty in examining peptides of unknown amino acid composition and length, is deciding

Table 1. *Integral mass numbers corresponding to derivatives of the amino acids*

Amino acid	Structure of derivative	<i>A</i> <i>N</i> -Terminal mass	<i>B</i> Mass of residue	<i>C</i> <i>C</i> -Terminal mass	Associated ions* (commonly observed)
Glycine	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CH}_2-\text{CO}- \end{array}$	114	71	102	
Alanine	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CH}-\text{CO}- \\ \\ \text{CH}_3 \end{array}$	128	85	116	<u>Loss of CO from <i>A</i> or <i>B</i></u>
Valine	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CH}-\text{CO}- \\ \\ \text{CH}(\text{CH}_3)_2 \end{array}$	156	113	144	<u>Loss of CO from <i>A</i> or <i>B</i></u>
Leucine†	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CH}-\text{CO}- \\ \\ \text{CH}_2-\text{CH}(\text{CH}_3)_2 \end{array}$	170	127	158	<u>Loss of CO from <i>A</i> or <i>B</i> plus further loss of keten from <i>A</i> (<i>m/e</i>100)</u>
Serine	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CH}-\text{CO}- \\ \\ \text{CH}_2-\text{O}-\text{CH}_3 \end{array}$	158	115	146	<u>Loss of CH₃OH from <i>A</i> or <i>B</i> (increases with increasing temperature)</u>
Threonine	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CH}-\text{CO}- \\ \\ \text{CH}_3-\text{CH}-\text{O}-\text{CH}_3 \end{array}$	172	129	160	<u>Loss of CH₃OH from <i>A</i> or <i>B</i> (increases with increasing temperature); occasional loss of side chain minus H (58 mass units) from <i>B</i> or <i>C</i></u>
Aspartic acid	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CH}-\text{CO}- \\ \\ \text{CH}_2-\text{CO}-\text{O}-\text{CH}_3 \end{array}$	186	143	174	Loss of CO from <i>A</i> or <i>B</i>
Glutamic acid	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CH}-\text{CO}- \\ \\ \text{CH}_2-\text{CH}_2-\text{CO}-\text{O}-\text{CH}_3 \end{array}$	200	157	188	Loss of CO from <i>A</i> or <i>B</i>
Asparagine	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CHCO}- \\ \\ \text{CH}_2-\text{CO}-\text{N}(\text{CH}_3)_2 \end{array}$	199	156	187	Loss of CO from <i>A</i> or <i>B</i>
Glutamine	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CH}-\text{CO}- \\ \\ \text{CH}_2-\text{CH}_2-\text{CO}-\text{N}(\text{CH}_3)_2 \end{array}$	213	170	201	Loss of CO from <i>A</i> or <i>B</i>
Phenylalanine	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{N}-\text{CH}-\text{CO}- \\ \\ \text{CH}_2-\text{C}_6\text{H}_5 \end{array}$	204	161	192	<u>Loss of CO from <i>A</i>; <i>m/e</i>91, 162</u>

Table 1—continued

Amino acid	Structure of derivative	<i>A</i> <i>N</i> -Terminal mass	<i>B</i> Mass of residue	<i>C</i> <i>C</i> -Terminal mass	Associated ions* (commonly observed)
Tyrosine		234	191	222	<u>Loss of CO from <i>A</i>;</u> <u><i>m/e</i> 121, 161, 192</u>
Tryptophan		257	214	245	<u>Loss of CO from <i>A</i>;</u> <u><i>m/e</i> 144, 215</u>
Lysine		241	198	229	<u>Loss of CO from <i>A</i>;</u> <u><i>m/e</i> 171</u>
Histidine†		208	165	196	<u><i>m/e</i> 95;</u> <u>satellites of <i>B</i> at ± 14 mass units</u>
Ornithine§		227	184	215	Loss of CO from <i>A</i>
Proline		140	97	128	

* Associated ions that are normally abundant indicated by underlining.

† Not differentiated from isoleucine.

‡ Permethylated of this residue was not always successful; for modification see Vliegthart & Dorland (1970).

§ Result of hydrazinolysis of arginine.

whether the ions at highest mass in the spectrum indicate the termination of the molecule (molecular ion) or whether sequence ions exist beyond this, but cannot be differentiated from the noise and background level. Normally, a molecular ion region is indicated by strong M^+ and $M^+ + 1$ ions (the latter largely an isotope peak under normal circumstances), and losses of 15 mass units (CH_3^-) and 31 mass units (OCH_3^-) from the molecular ion. In this case therefore we may be confident that m/e 898, 914, 929 and 930 indicate *C*-terminal tyro-

sine, giving the full sequence as Leu-Glu-Gln-Val Pro-Tyr.

Peptide C47a. After removal of impurities as described earlier, a good spectrum of this peptide was produced and interpreted as follows. Abundant ions at m/e 156 and 128, together with a 'metastable peak' corresponding to the transition $156 \rightarrow 128$ indicate *N*-terminal valine. A 'metastable peak' corresponding to the transition $m_1 \rightarrow m_2$ is found in the mass spectrum, at a position m^* such that $m^* = m_2^2/m_1$, when the reaction $m_1 \rightarrow m_2$ occurs in

Fig. 1. Mass spectrum (above m/e 100) of the modified peptide C69. The peptide has the sequence Leu-Glu-Gln-Val-Pro-Tyr. To simplify the figure, the normal amino acid abbreviations are used to represent the modified amino acids.

the field-free region preceding the magnetic analyser (Beynon, Saunders & Williams, 1968).

A strong sequence ion at m/e 227, 71 mass units above m/e 156, corresponds to a Val-Gly- sequence. An even stronger sequence ion at m/e 425 shows the sequence of the first three amino acids to be Val-Gly-Lys- (m/e 171 is present as expected). After this, m/e 568 and m/e 695 indicate unambiguously the sequence Val-Gly-Lys-Asp-Leu-. Beyond this two peaks are present that do not correspond to amino acid mass differences from m/e 695 or to mass differences which may be anticipated for those amino acids undergoing additional facile losses from their side chains (e.g. loss of methanol from a serine or threonine residue, see below). However, higher in the spectrum, an obvious molecular ion region is present, as shown by m/e 982, 951 (–31 mass units) and 967 (–15 mass units). It would appear therefore that the peptide bond connecting the penultimate and C-terminal amino acids has not broken on ionization, as is normally the case. Fortunately this is a rare occurrence. It is possible to calculate a limited number of possibilities for the sequence by taking the mass difference from m/e 695 to 982, and testing for every combination of two amino acids.

The following combinations are possible:

Peptide C42. A number of intense spectral scans were obtained with this peptide, and were interpreted as follows: m/e 158 indicates N-terminal serine, and this is verified by the expected loss of methanol to give m/e 126. An intense peak at m/e 324 (356–32) shows that the second residue is lysine. The ion expected at m/e 171 is also present. A mass difference of 157 from m/e 356 to 513 places glutamic acid in the third position giving the sequence so far as Ser-Lys-Glu-. The next most intense ion in the spectrum is m/e 669 and corresponds to a Ser-Lys-Glu-Asn- unit. Beyond this, m/e 867 shows the next residue to be lysine and a further ion at m/e 1065 establishes the sequence of the peptide as Ser-Lys-Glu-Asn-Lys-Lys.

Since m/e 1065 is not a molecular ion, this is not in fact the full sequence of the peptide. The presence of m/e 171 and 98 are again noted, as should be the unambiguous assignment of asparagine in position four.

Peptide C53b. A satisfactory spectrum was obtained from this peptide, although because of its low volatilization temperature some difficulty was encountered in separating it from the impurity in the sample.

A peak at m/e 114 indicates *N*-terminal glycine. The next abundant ion beyond this is at m/e 312, and corresponds to a Gly-Lys- unit. The placing of the next residue is not immediately obvious, since although a strong ion is present at m/e 441, corresponding to a Gly-Lys-Thr- unit, the expected loss of methanol from threonine is not obvious at m/e 409. However, on closer inspection of the spectrum, an abundant 'metastable peak' is present at m/e 379.2 corresponding to this loss ($441 \rightarrow 409$), thus confirming threonine in position three. A clear molecular ion region (M^+ at m/e 599) indicates that the full sequence of the peptide is Gly-Lys-Thr-Leu.

A further feature of this spectrum is the intense ion at m/e 541 arising presumably via the loss of the threonine side chain from the molecular ion, with accompanying hydrogen transfer to the peptide backbone.

Peptide C612a. A fairly good spectrum of this peptide was obtained at a source temperature of 260°C.

Sequence ions at m/e 200, 327, 484, 611 and 767 unambiguously indicate the sequence Glu-Leu-Glu-Leu-Asn. The next most abundant ion is at m/e 958 corresponding to a tyrosine unit (not *C*-terminal), and a number of much weaker peaks are observed higher in the spectrum. Although the tyrosine unit at m/e 958 is not *C*-terminal, the presence of m/e 192 in the spectrum would indicate the additional presence of a *C*-terminal tyrosine (see III). Thus, it can be clearly stated that the spectrum shows the sequence Glu-Leu-Glu-Leu-Asn-Tyr- and that this first tyrosine is not the *C*-terminus, which is due to an additional tyrosine unit further along the chain.

Peptide C610. A good spectrum was obtained from this peptide, clearly indicating the full sequence of the peptide to be Glu-Gln-Val-Pro-Tyr- via the ions at m/e 200, 370, 483, 580 and 802 (M^+).

Peptide C57c. A good spectrum, almost entirely free from contamination, was obtained from this

peptide. The sequence is clearly indicated as Ser-Pro-Leu-Leu-Glu-Lys-Tyr. The peaks at m/e 171 and m/e 121 and 192 confirm the presence of lysine and tyrosine. Other points to note are m/e 98 from the glutamic acid residue, and loss of methanol from serine to give m/e 126, m/e 223 and m/e 360.

Peptide C3a. The separation of this peptide from the impurity present in the sample was made somewhat more difficult than usual because of its low volatilization temperature. However, a good spectrum was obtained, and the sequence determined as follows. The most abundant ion in the spectrum is m/e 144 which does not correspond to an *N*-terminal amino acid, but to the side chain fragment of tryptophan. The complete absence of m/e 130 shows that the indole nitrogen is completely methylated by the technique employed. The next abundant ion in the spectrum is m/e 215 which is the tryptophan analogue to the *C*-terminal tyrosine fragment shown in structure III, and hence is indicative of *C*-terminal tryptophan. The *N*-terminus of the peptide is located at m/e 128, and the sequence ion at m/e 326 corresponds to an Ala-Lys- sequence. After m/e 326 the next most abundant ion in the spectrum is the molecular ion at m/e 571. The ions normally indicative of a molecular ion region at M^+-15 and M^+-31 , although present, are much decreased in intensity. This is no doubt due to the excellent positive charge-stabilization properties of the indole nucleus in this small peptide. Two smaller ions are present in the spectrum at m/e 581 and m/e 587 which may be artifacts of the *N*-acetylation and permethylation of tryptophan or may be due to the impurity in the sample. They do not, however, hinder spectral interpretation.

The classically determined (Gray, 1963) sequences of the above peptides, made available only after the mass spectrometric analyses, are given in Table 2.

It is evident therefore that working on 0.1–1.0 μ mol quantities of completely 'unknown' peptides, we were able to ascertain unambiguously 42 of the 46 residues in the above peptides (bearing

Table 2. *Classically determined sequences of the peptides*

Peptide	Sequence	Residue No. (Ambler & Meadway, 1969)
C69	Leu-Glu-Gln-Val-Pro-Tyr	55–60
C47a	Val-Gly-Lys-Asp-Ile-Thr-Leu	85–91
C42	Ser-Lys-Glu-Asn-Lys-Lys-Phe	172–178
C53b	Gly-Lys-Thr-Leu	159–162
C612a	Glu-Ile-Glu-Leu-Asn-Tyr-Tyr	138–144
C610	Glu-Gln-Val-Pro-Tyr	56–60
C57c	Ser-Pro-Ile-Leu-Glu-Lys-Tyr	78–84
C3a	Ala-Lys-Trp	

in mind the lack of differentiation between leucine and isoleucine); these results are sufficiently encouraging to allow an investigation of mixtures containing an unspecified number of 'unknown' component peptides.

Because of the limitations of the mass spectrometric technique for the determination of peptide sequences, a strategy must be defined which will enable the maximum amount of information to be obtained from an unknown protein in as short a time as possible.

We propose the following procedure: the unknown oligopeptide or protein is first cleaved specifically by a proteolytic enzyme e.g. chymotrypsin. Since the length limit presently useful for mass spectrometry is about ten residues, the enzymic hydrolysate must be fractionated according to mass. This is conveniently and cleanly done by gel filtration through Sephadex G-25. In other experiments we have used a column (65 cm $\times$ 1.75 cm diam.) of Sephadex G-15 resin, and eluted the peptides with 0.5 M-acetic acid (Morris, 1970). Calibration of the column by standard runs allows location of fractions containing material of molecular weight <1000. To ascertain which fractions contain peptides that are not

amenable to mass spectrometry by direct *N*-acetylation and permethylation we suggest electrophoretic analysis of a sample from each tube. Peptides containing the amino acids that are difficult to analyse may be detected by staining (*S*-containing amino acids, histidine, arginine) or by radioactive labelling (cysteine). Tubes containing these peptides may then be treated separately as referred to above. The remaining fractions may be pooled or split into suitable subfractions, and the mixtures examined as described below. Fractions of higher molecular weight must also be pooled or split into suitable subfractions and re-digested, preferably with another specific proteolytic enzyme. The mixture is refractionated on Sephadex, and the process repeated until mixtures of suitably sized components are formed. The whole procedure is then repeated, starting with a different enzymic hydrolysis, in an effort to create overlaps in sequence. An example of the use of part of this strategy, namely the use of gel filtration and mixture analysis is, given below.

We are grateful for the co-operation of Dr B. S. Hartley of the Medical Research Council Unit, Cambridge, in providing two adjacent fractions

Fig. 2 (a)

Fig. 2. Mass spectrum (above m/e 100) of the modified peptide mixture fraction 'B'. (a) Source temperature 200°C; (b) source temperature 220°C. Dotted lines show mass peaks of the sequence Ala-Asn-Asn-Lys. Solid lines show mass peaks of the sequence Val-Gly-Leu-Ala-Pro-Val-Ala. To simplify the figure, the normal amino acid abbreviations are used to represent the modified amino acids.

from the lower-molecular-weight region of a Sephadex fractionation of a tryptic digest of a protein. Neither the protein nor any information concerning the fractions was known to us at the time of examination. The results are presented below.

After *N*-acetylation and permethylation of the mixture fractions, the technique used for analysis was similar to that described earlier for the separation of impurity and peptide. The object is to enhance volatility differences of components in the mixture by the use of a temperature gradient (Morris, 1970; McLafferty, Venkataraghavan & Irving, 1970). In this way it is often possible to obtain scans of single components in the mixture.

Fraction 'A'. Examination of fraction 'A' indicated only one major component, the spectrum of which is easily interpreted as being caused by the sequence Val-Gly-Leu-Ala-Pro-Val-Ala. Since it was known that the peptides were products of tryptic digestion, the above complete sequence must correspond to the *C*-terminus of the protein.

Fraction 'B'. This fraction yielded spectra that changed continuously as the temperature was increased from 150°C to 250°C, resulting in what appeared to be a single component peptide at 260°C.

Examination of the earlier spectra showed them to be due to a mixture of two components (the material appearing at 260°C was not evident at that stage). Interpretation is effected by comparing consecutive scans (for representative consecutive scans see Figs. 2a and 2b), and looking for changes in peak intensity, as one component becomes either more or less intense relative to the other. Figs. 2(a) and 2(b) show that the component that is more abundant in the scan at a source temperature of 200°C (Fig. 2a) has the same sequence as that found for fraction 'A' (and must arise from an overlap of fractions 'A' and 'B'); the second component, which gives rise to ions of increased abundance at 220°C (Fig. 2b), is a tetrapeptide of sequence Ala-Asn-Asn-Lys- (see the dotted lines in the Figures).

Examination of the scan taken at 260°C (Fig.

Fig. 3. Partial mass spectrum (above m/e 100) of the modified peptide mixture fraction 'B'. All parts of Fig. 3 (a, b and c) are normalized relative to m/e 213 as base peak (arbitrarily taken as 100% relative abundance). To simplify the figure, the normal amino acid abbreviations are used to represent the modified amino acids. (a) Showing the normal sequence ions at a source temperature of 260°C. (b) Showing the pyrrolidonecarboxylic acid unit sequence ions at a source temperature of 260°C. (c) Showing the 'in-chain tyrosine' sequence ions at a source temperature of 260°C. The origin of these sequence ions is indicated below the spectrum.

3a-c) shows it to be a noteworthy spectrum from the point of view of interpretation. The spectrum is more complex than would be expected for a single component, and there are a number of reasons contributing to this. On interpretation it can be seen that the spectrum contains in fact three components, but all coming from the same peptide, and all combining to verify the sequence deduced. To clarify the discussion, the sequence ions corresponding to these three components are reproduced separately in Figs. 3(a), 3(b) and 3(c). Nearly all abundant ions in the spectrum are accounted for by the sequence ions, but, in addition, other peaks that are as large in a given region as those used in the sequence determination are indicated, and

distributed (arbitrarily) between the three parts of Fig. 3.

First, a search is made for the *N*-terminal ion. The only possibility found is m/e 213 corresponding to *N*-terminal glutamine (Fig. 3a). The peak at m/e 185 corresponds to a loss of carbon monoxide from this fragment. A number of abundant ions exist beyond m/e 213, but not until m/e 404 do we reach a significant mass difference, the whole unit being Gln-Tyr. Similarly the next ion corresponding to an amino acid mass difference is at m/e 595, which indicates a Gln-Tyr-Tyr- sequence. Beyond this, difficulties arise because of the number of equi-intense peaks in the region m/e 600-850. The small peak at m/e 724 (Fig. 3a) could be caused by an

addition of a threonine unit. However, the larger peak 32 mass units lower could confirm this or alternatively point to the next residue being proline (Thr. -32). This ambiguity is resolved in favour of a threonine unit by the presence of an abundant 'metastable ion' at m/e 662 for the reaction m/e 724 $\rightarrow$ 692 (loss of methanol).

The sequence thus far is Gln-Tyr-Tyr-Thr-. The next most intense peak following m/e 724 is m/e 805, 81 mass units higher. This does not correspond to an amino acid, but is caused by the loss of 32 mass units from the true sequence ion at m/e 837 (Fig. 3a). It is peculiar that this loss seems more pronounced than at the threonine sequence ion itself. The sequence so far is Gln-Tyr-Tyr-Thr-Val.

The higher mass region is again complicated, but a prominent ion is present at m/e 998, indicating that the next residue is phenylalanine. The spectrum intensity beyond this point diminishes rapidly. The sequence found so far is Gln-Tyr-Tyr-Thr-Val-Phe.

The other two components in the spectrum are as follows: first, the presence of a very intense m/e 98 peak (Fig. 3b) indicates a possible *N*-terminal pyrrolidonecarboxylic acid unit (see Fig. 3a), and this is in fact present. Glutamine is known to be easily cyclized in an *N*-terminal position, a cyclization of approx. 40% of the sample having taken place here, probably during the process of preparing the derivative. The sequence of the peptide can then be followed starting from the methylated pyrrolicarboxylic acid unit via m/e 317, 508, 637, 718 (-32) to the sequence ion of highest mass at m/e 911 (Fig. 3b).

Secondly the presence of a mid-chain tyrosine unit gives rise to fragments formed by tyrosine N-C bond breakage, with hydrogen transfer and charge retention on the *C*-terminal fragment (Morris, 1970). This fragment can then give rise to normal peptide-bond cleavage. Ions arising from this process can be seen at m/e 161, m/e 352 and m/e 481 (Fig. 3c). Since there are two in-chain tyrosine units, the additional possibility of a threonine sequence ion at m/e 290, immediately after m/e 161, exists and this is in fact observed (Fig. 3c).

In summary, the results indicate the presence of three components in the two fractions, having sequences:

Val-Gly-Leu-Ala-Pro-Val-Ala
Ala-Asn-Asn-Lys
Gln-Tyr-Tyr-Thr-Val-Phe

These sequences have been confirmed, and the third component contained a further two amino acids not shown on the spectrum. It must be remembered here that we have made no attempt in this study to modify any arginine residues present before permethylation. The inclusion of

such a step in the procedure might well have resulted in the full sequence of the third component being determined.

The sequence confirmation was made by Dr B. S. Hartley, who prepared the peptide mixtures as follows.

A sample (400mg) of pepsin (Worthington Biochemical Corporation; batch PM 702) was denatured by the method of Tang & Hartley (1970), dialysed against water at 2°C, freeze-dried, and dissolved in 20ml of 0.1M-NH₄HCO₃. After adding 4mg of trypsin, the mixture was incubated for 18h at 20°C and applied to a Sephadex G-25 column (90cm $\times$ 2.5cm diameter) eluted with 0.01M-NH₃. As expected, most of the protein emerged at the exclusion volume and only two peaks of small peptides were detected, which were pooled into two fractions, A and B in order of their emergence from the Sephadex column.

High-voltage paper electrophoresis at pH 6.5 and pH 2.1 showed that fraction A contained one main component, and fraction B contained three peptides, the mobilities of which were consistent with their presumed sequences (Tang & Hartley, 1970):

Val-Gly-Leu-Ala-Pro-Val-Ala
Ala-Asn-Asn-Lys
Gln-Tyr-Tyr-Thr-Val-Phe-Asp-Arg

The two fractions were freeze-dried and supplied for mass spectrometric examination as unknown mixtures.

DISCUSSION

The importance of mixture analysis cannot be overemphasized. It is not necessary to find the full sequence of every component present, since overlaps will be created by the use of different proteolytic enzymes. At the present time, certain amino acids require special modification in addition to the normal methods. An amino acid analysis of the total protein, before sequence studies, will indicate the extent of this problem. All amino acids except cysteine, methionine, histidine and arginine can be dealt with as outlined in the present paper, and it may be that at the present time a combination of classical and mass spectrometric techniques would be needed for a complete sequence determination of a protein.

The results given in the present paper must be judged in the light of two considerations in particular. First, incorrect sequence assignments were not made, and cannot be made by using the mass spectrometric method provided that due care is taken in interpretation. The case of threonine, mentioned twice here, is a good example of the type of fragmentation which must be documented, and

expected, when interpreting spectra. Secondly, and most important, the time needed for peptide modification, instrumentation and interpretation is at least a factor of 10 shorter than would be required by classical techniques for sequence determination. Use of the technique proposed in this paper, namely enzymic digestion, gel filtration then mixture analysis, eliminates the need for complicated isolation procedures and numerous amino acid analyses, and offers the prospect of a revolution in protein sequence determination.

Finally, we wish to stress that the above results have been obtained by using only low-resolution mass spectrometry. We suggest that the use of computer-aided high-resolution mass spectrometry as a solution to the problem of determination of sequences cannot be recommended as a general method at present for the following reasons.

1. The use of the high-resolution mode considerably decreases the sensitivity of this instrument.

2. The cost of a computer-mass spectrometer unit would be prohibitive to most laboratories engaged in sequence determination.

3. The reliability of the total system is decreased by the introduction of a further complex electronic unit.

4. Most important, our experience has indicated that the few ambiguities that may arise in mixture analysis may be resolved by isotopic labelling and/or individual manual high-resolution measurements (Morris, 1970), without resort to expensive sophisticated techniques.

We thank the Science Research Council for financial support of the work carried out in Cambridge.

REFERENCES

- Agarwal, K. L., Kenner, G. W. & Sheppard, R. C. (1969). *J. Am. chem. Soc.* **91**, 3096.
- Ambler, R. P. (1963). *Biochem. J.* **89**, 349.
- Ambler, R. P. & Brown, L. H. (1967). *Biochem. J.* **104**, 784.
- Ambler, R. P. & Meadway, R. J. (1969). *Nature, Lond.*, **222**, 24.
- Beynon, J. H., Saunders, R. A. & Williams, A. E. (1968). *The Mass Spectra of Organic Molecules*, p. 44. London: Elsevier Publishing Co. Ltd.
- Biemann, K., Cone, C., Webster, B. R. & Arsenault, G. P. (1966). *J. Am. chem. Soc.* **88**, 5598.
- Coggins, J. & Benoiton, L. (1968). *Abstr. Pap. Am. chem. Soc. 156th Meet., Atlantic City*, Biol. 18.
- Das, B. C., Gero, S. D. & Lederer, E. (1967). *Biochem. biophys. Res. Commun.* **29**, 211.
- Franek, F., Keil, B., Thomas, D. W. & Lederer, E. (1969). *FEBS Lett.* **2**, 309.
- Geddes, A. J., Graham, G. N., Morris, H. R., Lucas, F., Barber, M. & Wolstenholme, W. A. (1969). *Biochem. J.* **114**, 695.
- Gray, W. R. (1963). In *Methods in Enzymology*, vol. 11, p. 469. Ed. by Hirs, C. H. W. New York and London: Academic Press.
- Hakamori, S. I. (1964). *J. Biochem., Tokyo*, **55**, 205.
- Lederer, E. (1968). *Pure appl. Chem.* **17**, 489.
- Lenard, J. & Gallop, P. M. (1969). *Analyt. Biochem.* **29**, 203-209.
- McLafferty, F. W., Venkataraghavan, R. & Irving, P. (1970). *Biochem. biophys. Res. Commun.* **39**, 274.
- Morris, H. R. (1970). Ph.D. Thesis; Leeds University.
- Morris, H. R., Geddes, A. J. & Graham, G. N. (1968). *Biochem. J.* **111**, 38P.
- Prox, A., Schmid, J. & Ottoheym, H. (1969). *Justus Liebigs Annln Chem.* **722**, 179.
- Senn, M., Venkataraghavan, R. & McLafferty, F. W. (1966). *J. Am. chem. Soc.* **88**, 5593.
- Shemyakin, M. M. (1968). *Pure appl. Chem.* **17**, 313.
- Shemyakin, M. M., Ovchinnikov, Yu. A., Vinogradova, E. I., Feigina, M. Yu., Kirushkin, A. A., Aldanova, N. A., Alakhov, Yu. B., Lipkin, V. M. & Rosinov, B. V. (1967). *Experientia*, **23**, 430.
- Shemyakin, M. M., Ovchinnikov, Yu. A., Vinogradova, E. I., Kirushkin, A. A., Feigina, M. Yu., Aldanova, N. A., Alakhov, Yu. B., Lipkin, V. M., Morishnikov, A. I., Rosinov, B. V. & Kazaryan, S. A. (1970). *FEBS Lett.* **7**, 8.
- Tang, J. & Hartley, B. S. (1970). *Biochem. J.* **118**, 611.
- Thomas, D. W. (1968). *Biochem. biophys. Res. Commun.* **33**, 483.
- Thomas, D. W. (1969). *FEBS Lett.* **5**, 53.
- Thomas, D. W., Das, B. C., Gero, S. D. & Lederer, E. (1968a). *Biochem. biophys. Res. Commun.* **33**, 519.
- Thomas, D. W., Lederer, E., Bodanszky, M., Izdebski, J. & Muramatsu, I. (1968b). *Nature, Lond.*, **220**, 580.
- Vilkas, E. & Lederer, E. (1968). *Tetrahedron Lett.* p. 3089.
- Vliegenthart, J. F. G. & Dorland, L. (1970). *Biochem. J.* **117**, 31P.