PEPTIDE SEQUENCES BY MASS SPECTROMETRY

A thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Biochemistry at Massey University

STEPHEN BRIAN HENRY KENT

1970

"One significant result of this work is that it demonstrates that insulin, and probably other proteins, are homogeneous substances with unique structures and are not merely statistically random polymers of amino acids. This brings the science of proteins into the realms of classical organic chemistry and opens up the way to similar studies on the many other proteins that exist in nature and hence to a better understanding of the chemistry of life".

F. Sanger

1957

ABSTRACT.

The preparation and mass spectrometry of permethylated peptide derivatives was investigated. Procedures for the modification of free peptides prior to permethylation were examined. Acetylation with methanol+ acetic anhydride was found to result in partial esterification of the peptide. Specific cleavage of the C-terminal residue was also observed; a mechanism is proposed for this reaction. Esterification with HCl in methanol followed by acetylation of the peptide ester gave a mixture of products due to random methanolysis during the esterification. Methods of acetylating free peptides were examined, and it was found that the use of water + acetic anhydride at room temperature resulted in rapid quantitative acetylation, with no significant side reactions. Reaction of an ethereal solution of diazomethane with the acetyl-peptide gave quantitative esterification with negligible byproduct formation.

Use of dimethylsulfinyl sodium in dimethylsulfoxide, and methyl iodide for the permethylation of peptide derivatives was investigated. Suitable conditions were found for the preparation of the reagent and for its use in the permethylation reaction. Substitution at existing ester groups was found to occur during the permethylation, and the products were partially characterised. Use of the free acetyl-peptide rather than its methyl ester eliminated this side reaction. Introduction of more than the expected number of methyl groups was observed. This extra-methylation was found to occur mainly at specific residues, although some random methylation was observed. The conditions of permethylation were adjusted to minimise extra-methylation and limit it to specific sites in the molecule. Peptides containing aspartyl residues undergo chain cleavage; the products of this reaction were identified and a mechanism proposed for their formation. The permethylation reaction is discussed in relation to the formation of these artefacts; it is thought to

involve deprotonation of the peptide to form a multiple anion. Reaction conditions are suggested to eliminate these side reactions.

The mass spectrometry of permethylated peptide derivatives is discussed and the mass spectra of peptides of known sequence reported. The mass spectra show the sequence-determining fragments as the principal ions. This observation is rationalised in terms of the negative-inductive effect of the N-methyl groups. The simple procedure for interpreting the mass spectra of permethylated peptide derivatives is outlined, together with the use of minor fragmentation modes in identifying the molecular ion and sequencing peaks. Deuteriated methyl iodide, high resolution mass spectrometry and the detection of metastable transitions can all be used to confirm the deduced sequence.

The techniques developed were applied to a mixture of free peptides isolated from cheese; the three peptides present were sequenced. The results were confirmed by high resolution mass measurement and permethylation with deuteriated methyl iodide.

The present state of peptide sequence determination by mass spectometry is evaluated and possible future developments discussed.

ACKNOWLEDGMENTS.

It has been a privilege and a pleasure to work under my supervisor, Prof. R. Hodges, whose attitudes and knowledge of organic chemistry have been most instructive; I should like to thank him for his assistance with the mass spectrometry reported here, especially for performingall the high resolution mass measurements, and for his patient instruction in the operation of the MS902 and the full use of its capabilities.

Thanks also to the staff members of the Chem/Biochem Department for their aid in various aspects of this work; I have learned a great deal in the course of many discussions with helpful listeners. There are numerous others from Massey University, the D.S.I.R. and D.R.I. who have been involved at some point in this work and to whom I am grateful.

Finally, without my wife Joy and our daughter Julia, this work would probably never have been completed.

TABLE OF CONTENTS.

HISTORICAL DEVELOPMENT.

Protein Structure:

Historically, the determination of peptide sequencesfits into the classical chemical approach to the study of the covalent structure of proteins.

The polypeptide structure of proteins was first proposed independently by Emil Fischer and F. Hofmeister in 1902 (ref. 1). They postulated that α-amino acid residues were covalently linked by amide bonds through their α-amino and carboxyl groups, to form a linear molecule made up of a sequence of amino acyl residues. This remained unsupported by significant chemical evidence until the determination of the complete covalent structure of insulin by Sanger and his co-workers between 1945 and 1955 (ref. 2). The approach used in this classic study is essentially that used up to the present day. It involves the fragmentation (by at least two methods) of the large protein molecule and the separation and isolation of the peptides produced. The sequence of amino acyl residues in these is then determined and the overall sequence of the original protein molecule is deduced from this information.

Thus, the key to protein structure is the determination of the sequences of a large number of small peptides, usually available only in micro-amounts in their pure state. As with protein structure, basic strategy in determining peptide sequences has not changed since the early work in the field. The several methods originally used (ref. 3) are still those normally used today, though techniques have been refined.

Chemical Methods:

One general approach is to chemically modify a terminal residue, completely hydrolyse the peptide and identify the "labelled" amino acid. There are procedures for both N-terminal (ref. 4) and C-terminal (ref. 3) amino

acids. Similar information can be gained by using proteolytic enzymes such as the exopeptidases, carboxypeptidase A which liberates C-terminal amino acids and leucine aminopeptidase which is specific for N-terminal amino acids.

The most useful approach to sequencing peptides is due to Edman (ref. 7) and involves the removal, one at a time, of the N-terminal residues from the peptide and identification of the amino acid removed at each step. This is the only chemical method yet devised for such sequential degradation, although its usefulness has been extended by modifications since its original introduction in 1950. Briefly, the peptide is treated with phenyl isothiocyanate to give a phenylthiocarbamyl derivative at the N-terminal residue; treatment with anhydrous acid specifically cleaves this modified residue, finally giving a phenylthiohydantoin of the amino acid and the remaining peptide chain, which may then be subjected to another cycle of reactions. After each cycle, the amino acid split off is identified. Decreasing yields for each cycle generally limits to between six and ten the number of residues whose sequence can be determined in this way.

Edman (ref. 8) has described an automated modification of his procedure, for application to intact protein molecules. Sixty N-terminal residues were sequenced using only 0.25 μmole of myoglobin. Each cycle of reactions was performed with a yield of 98%. The "Edman sequenator", though commercially available, has not yet made a great impact on methods of sequence determination. It cannot be used for small peptides, as it depends for its successful operation on the insolubility of the protein in all solvents used.

Use of modifications of the Edman method is the best procedure available at present for determining peptide structure. Combined with the general approach to protein structure already outlined, it has enabled the determination of over four hundred protein sequences with a length of thirty or more residues, from a wide variety of sources (ref. 9). The complete structure of a protein molecule containing 1320 amino acid residues has been reported (ref. 69). Despite

this, the determination of the amino acid sequence of a protein cannot be considered a routine matter and the methods described have definite and sometimes serious limitations, outlined in the following section.

Mass Spectrometry and Peptide Sequencing:

Although the principles of mass spectrometry were demonstrated in the first years of this century, mass spectrometers adequate for organic chemical use have been available only since the early 1950's. In fact, it in only in the last ten years that organic chemists have made extensive use of this technique.

The advantages of mass spectrometry over more classical techniques are usually given as the requirement of only a small sample size (for even a complete unknown) and the provision of a large amount of information about the molecule in a short time. Samples in the order of 10 to 100 μg are routinely used and a few hours' work provides data such as molecular weight, elemental composition and structural information.

However, potential advantages of mass spectrometry in the determination of peptide sequences do not lie in small sample size or in shorter analytical time, despite assertions to the contrary (e.g. ref. 59). Mass spectrometry can, at best, only match the 20 nanomoles to 1 micromole routinely needed for the successful sequencing of penta- or hexapeptides by modifications of the Edman technique (ref. 10), while concurrent handling of many peptides considerably reduces the "per peptide" analytical time by normal methods, so that mass spectrometry is not considered to offer significant advantages in this area.

There are, however, several limitations of classical sequencing methods. One restriction is peptide size: normally only six to ten residues can be determined (ref. 1, p. 144). A much more serious limitation is the lack of universal applicability of the Edman method. It cannot be used on peptides with blocked N-terminal residues, such as acetyl- or formylpeptides or those

peptides with pyrrolidone carboxylic acid formed from an N-terminal glutaminyl residue during isolation.

Neither is the Edman method suited to the detection of new or unusual amino acid residues, making one large area of peptide chemistry particularly unsuited to classical sequencing methods. That is, the determination of the structures of naturally occurring small peptides, such as some antibiotics and fungal metabolites. These compounds frequently contain unusual or entirely new constituents and even common residues are often linked by other than the normal peptide bonds.

In these areas mass spectrometry potentially has something to offer the protein chemist. It is uniquely applicable in the verification of the sequence and purity of synthetic peptides where methods involving degradations must be avoided if unusual artefacts are to be found with any degree of reliability (refs. 34, 43).

Modification of Peptides:

Before peptides can be examined by mass spectrometry, the interrelated problems of their low volatility and lack of thermal stability must be overcome. Although free dipeptides (ref. 11) and tripeptides (ref. 12) give mass spectra containing molecular ions, extensive artefact formation, such as cyclisation to dioxopiperazines (ref. 11) occurs, often dominating the spectra.

The low volatility of free peptides has been ascribed to their zwitterion character (ref. 13) and to hydrogen-bonding involving the amide groups (ref. 14). Early methods of increasing volatility relied on the elimination of the zwitterion nature of the molecule by acylation of the N-terminal residue (ref. 15), usually followed by esterification of the terminal carboxyl group (ref. 13); or on reduction of the peptide to give the polyamino alcohol (ref. 16) or polyamine (ref. 17). More recently, methylation of the amide nitrogens of acyl peptides was introduced to eliminate intermolecular hydrogen-

-bonding (ref. 18).

Early Studies:

The development of mass spectrometry of peptides has occurred in several distinct steps. Starting in 1958, pioneering studies were carried out (see Table I). As organic mass spectrometry was in its infancy, these involved little more than cursory examinations of a few derivatives of simple synthetic peptides to see if mass spectra could be obtained and what type of information they could be expected to give.

The studies of Biemann and his co-workers (refs. 16, 17) on the mass spectrometry and gas liquid chromatography (GLC) of reduced peptides appeared very promising. The very volatile derivatives fragmented in such a way as to give information on the sequence of the original peptide. However, with increasing use of the direct insertion probe, there was no longer a need for derivatives of such volatility. Modification of the peptide by reduction was probably considered too drastic a procedure and this approach has not been pursued.

In the work of Heyns and Grutzmacher (ref. 15) concepts were introduced that had a great influence on later developments. They realised that the acyl group, originally used to eliminate the zwitterion nature of the peptide, acted as a marker for ions containing the N-terminal residue of the peptide. Secondly, they noted that characteristic ions are formed due to cleavage on both sides of the carbonyl group of the peptide linkage and that identification of these ions in the mass spectrum of a N-trifluoroacetyl--peptide methyl ester was sufficient to establish the amino acid sequence. Work in the field has since relied heavily on these two observations.

At about the same time, Weygand and Prox examined the combined GLC-mass spectrometry of mixtures of N-trifluoroacetyl-peptide methyl esters (ref. 67). The peptides were derived from oligopeptides by partial acid hydrolysis. These two workers also stressed that the sequence of a peptide

TABLE I.

EARLY MASS SPECTROMETRY OF PEPTIDES.

Anderson:	1958	ref. 13	an N-trifluoroacetyl-tripeptide methyl ester.
Biemann:	1959	ref. 16	polyaminoalcohols and polyamines from di-,
	1960	ref. 17	tri- and tetrapeptides; combined GLC-MS.
Stenhagen:	1961	ref. 19	N-trifluoroacetyl-peptide ethyl esters, several di- and one tripeptide.
Manusadzhyan:	1964	ref. 20	N-acetyl-peptide ethyl esters and polyamino-alcohols from di- and tripeptides.
Heyns and Grutzmacher:	1963	ref. 15 ref. 21	N-acetyl-peptides, N-trifluoroacetyl-peptides, up to a pentapeptide.
Weygand and Prox:	1963	ref. 67	N-trifluoroacetyl-peptide methyl esters, up to a pentapeptide; combined GLC-MS.

was defined by the ions formed due to cleavage on both sides of the carbonyl groups of the peptide bonds.

Naturally Occurring Peptides:

The next development was the use of mass spectrometry to examine a number of naturally occurring peptide derivatives, needing little or no prior modification. For most of these compounds, normal chemical studies had led to partial or complete structures (see Table II). The series of papers by Lederer's group at the Institute of Natural Product Chemistry in France, working in conjunction with A.E.I. Laboratories in England, represented a fundamental breakthrough as the first use of mass spectrometry in determining the sequence of oligopeptides of unknown structure. A particularly exciting discovery was that "Fortuitine", a peptidolipid of molecular weight 1359 containing nine amino acids, gave a mass spectrum in which the sequencing peaks were the principal ions (ref. 27).

In the naturally occurring long-chain N-acyl-oligopeptide methyl esters examined by Lederer's group, the principal mode of fragmentation observed was rupture of the peptide bond. It was soon shown that there was no basic difference in the fragmentation modes of short and long-chain N-acyl derivatives of synthetic peptide esters. However, with the long-chain N-acyl group, fragment-ions containing the intact N-terminal residue and part of the peptide chain were shifted to higher mass and were thus more easily recognised (ref. 36).

Acyl Marking Groups:

The determination of the structure of fortuitine and subsequent study of acyl-peptide derivatives led to a rash of papers stressing the importance of identifying by mass spectrometry those fragments containing the intact N-terminal residue. The climate of thought at that time is exemplified by McLafferty's generalisation that "the structure of a linear molecule is determined unequivocally by using only the possible fragments which contain one end of

TABLE II.

MASS SPECTROMETRY OF NATURAL PEPTIDES

Shannon:	1964a,b	ref. 22	"sporidesmolides", cyclodepsipeptides.
	1964c	ref. 23	"angolide", a cyclodepsipeptide.
Pais:	1964	ref. 24	a partially hydrolysed tripeptide alkaloid.
Lederer, et al.	1964 1965a	ref. 25 ref. 26	"peptidolipin NA_1", a cyclic hepta-peptidolipid.
	1965b	ref. 27	"fortuitine", a nonapeptidolipid.
	1965c	ref. 28	a pentapeptidolipid.
	1966a	ref. 29	"peptidolipin NA_2".
	1966b	ref. 31	"mycoside C_b", a tripeptidoglycolipid.
Vining:	1966	ref. 30	"isariin", a cyclic pentapeptidolipid.
Kiryushkin:	1966	ref. 32	"isariin".
	1967	ref. 33	"staphlomycin S", a cyclodepsi-hexapeptide.
Biemann:	1966	ref. 34	"isariin".
Kingston:	1966	ref. 35	"ostreogrycin A", a cyclodepsipeptide.

the chain" (ref. 38).

Special methods of marking the N-terminal residue were proposed for easy recognition of sequencing peaks in low resolution mass spectra, usually involving N-acyl substituents of characteristic composition. Other authors proposed recognising sequence peaks by routine high resolution measurement of the entire spectrum, followed by computerised interpretation (see Table III). Peptides used as examples in computer-aided interpretation of mass spectra were generally synthetically derived, though two groups (refs. 34,41) used derivatives of "isariin", a naturally occurring cyclic pentapeptidolipid. Their results confirmed the structure proposed by Wolstenholme and Vining (ref. 30) rather than that suggested by Kiryushkin (ref. 32), both of which had been inferred from mass spectrometry.

Although very promising, the computer-aided interpretation of the mass spectra of peptide derivatives has fallen into disuse because of the lack of universal and reliable methods of preparing peptides for mass spectrometry.

N-methylated Peptides:

At the beginning of 1967, mass spectrometry was normally limited to small peptides of six or seven residues made up of simple, non-polar amino acids; even then it was difficult to interpret the spectra in terms of sequence. The single exception to this generalization was the aptly named "Fortuitine".

This was a peptidolipid isolated from Mycobacterium fortuitum and was known to be an N-acyl-oligopeptide methyl ester. The preliminary structure (I) had been proposed:** (over page).

** (All abbreviations for structural formulae are in accord with the IUPAC-IUB rules: see Biochem. 5, 2485 (1966)).

$CH_3(CH_2)_nCO$-Val-Val-Val-Thr(Ac)-Thr-Ala-Pro-OMe

(I) n=18, 20.

Mass spectrometry of fortuitine gave a mass spectrum containing two molecular ions at m/e1331 and m/e1359, due to the homologous fatty acyl substituents. The spectrum consisted predominantly of the sequencing peaks corresponding to cleavage of the peptide bonds and allowed the complete structure to be determined as:

$CH_3(CH_2)_nCO$-Val(Me)-Leu-Val-Val(Me)-Leu-Thr(Ac)-Thr(Ac)-Ala-Pro-OMe

(II) n=18, 20.

In retrospect, the exceptional volatility of this compound can be attributed to the presence of three amino acid residues with substituted amide nitrogens: the two N-methylleucines and the proline.

In the light of this structural feature of fortuitine and the fact that derivatives of several synthetic hepta- and octapeptides were not volatile enough to give mass spectra, J. Van Heijenoort, working with Lederer's group, proposed that the factor limiting the volatility of peptide derivatives was intermolecular hydrogen-bonding through the -CO-NH- groups (ref. 14). It was found that acyl peptides could be quantitatively converted to N-methyl amides, -CO-NMe-, thus eliminating the possibility of such interaction (ref. 44). The expected increase in volatility of these N-permethylated derivatives resulted in useful mass spectra from much larger peptides (refs. 45 and 48). But now an unexpected problem became

TABLE III.

IDENTIFICATION OF SEQUENCING PEAKS.

Lederer, et al:	1965	ref. 36	long-chain acyl groups; mixed homologous acyl groups.
	1967	ref. 14	mixed CH_3CO/CD_3CO groups; mixed $CH_3(CH_2)_8CO/CD_3(CH_2)_8CO$ groups.
Prox:	1966	ref. 37	aryl groups containing a hetero-atom with fixed isotopic ratio.
McLafferty:	1966a	ref. 38	separate spectra of CH_3CO, CD_3CO derivatives; use of CF_2BrCO; high resolution measurement and computer interpretation.
	1966b	ref. 39	high resolution measurements and computer interpretation.
Biemann:	1966a	ref. 40	high resolution measurements and computer interpretation.
	1966b	ref. 34	high resolution measurements and computer interpretation; acyl group containing a halogen atom; aryl group.
Barber:	1966	ref. 41	high resolution measurements and computer interpretation.
	1967	ref. 42	automated metastable detection, high resolution measurements and computer interpretation.

apparent: successive sequencing peaks decreased in intensity with increasing mass and ions beyond a certain point (ten to twelve residues) were too weak to be observed (ref. 45).

This search for a means of increasing the volatility of peptide derivatives was to lead to an unexpected bonus: the mass spectra of permethylated peptide derivatives consist almost exclusively of sequence peaks resulting from cleavage of the peptide bond (ref. 45). It is this simplification, even more than the increase in volatility, that makes N-permethylation such an important technique in the sequence determination of peptides by mass spectrometry.

Several groups have adopted this procedure and unknown peptides, both naturally occurring and from proteins, have been successfully sequenced (see Table IV). Because of the simplified spectra of the N-methyl derivatives and the fact that the amino acid composition of the peptide was usually known, high resolution measurements were not taken. Lederer (ref. 43) has used mass spectrometry of permethylated derivatives to locate and identify artefacts in synthetic peptides; the nature of these artefacts, some residual N-methoxycarbonyl-peptide and a threonine present as a benzoate, was such that they would probably not have been found by normal degradative sequencing techniques.

Other Approaches:

As well as the developments outlined above, several other approaches to the mass spectrometry of peptides have been tried, with little success so far. Two research groups in particular should be mentioned, because of their continued work in the field.

A series of German papers on the mass spectrometry of N-trifluoroacetyl-peptide methyl esters have been published, based on the original work of Stenhagen (1961, ref. 19). GLC was used in conjunction with mass spectrometry to separate and identify small peptides (di-, tri- and tetra-peptides) from partial acid hydrolysates of oligopeptides, and the sequence of the original peptide was then

TABLE IV.

PERMETHYLATED UNKNOWN PEPTIDES.

Lederer's group:	1968a	ref. 43	pentadecapeptides; gramicidns A and B.
	1968b	ref. 47	tetrapeptidoglycolipid; mycoside C_{b1}.
	1968c	ref. 48	tetradecapeptide (10)*; stendomycic acid.
	1969a	ref. 46	heptapeptidolipid; esperinic acid.
	1969b	ref. 49	octadecapeptide (10); from pig immunoglobulin γ-chain.
	1969c	ref. 50	heptapeptide (6); from the zymogen of phospholipase A.
			docosapeptide** (6); from pig immunoglobulin γ-chain.
Geddes, et al.	1969	ref. 51	octapeptide; from silk fibroin.
Agarwal, et al.	1969	ref. 52	tetra- and octapeptides (7); from feline gastrin.
Hiramoto, et al.	1970	ref. 68	cyclodepsinonapeptidolipid; viscosin.

* the number in brackets indicates the number of residues successfully sequenced, if less than the total.

** docosapeptide, i.e. twenty-two amino acids.

reconstructed (refs. 53,67). The structure of one naturally occurring cyclodecapeptide has been deduced in this fashion (refs. 53,54,55). This method has also been used to check "failure sequences" in oligopeptides synthesised by the Merrifield solid phase technique (ref. 53), but such a degradative approach in the search for artefacts in synthetic peptides is questionable.

Shemyakin and his co-workers in the U.S.S.R. have also done a considerable amount of work in this field. A series of papers on the mass spectral fragmentations of synthetic compounds pioneered the study of depsipeptides by mass spectrometry (refs. 61,62).

In 1965 Shemyakin apparently independently realised the significance of mass spectral fragmentation at the peptide bond (Heyns and Grutzmacher, 1963. refs. 15,21) in the sequence determination of oligopeptides (ref. 63). This concept was developed in a number of papers in 1965 and 1966. The Russian workers then embarked on a systematic study of all the characteristic details introduced by each individual amino acid into the general pattern of the mass spectrum of acyl-peptide esters (ref. 56). Although a vast amount of work on synthetic peptides was done (ref. 59) and much useful chemistry of peptide derivatives was elucidated (refs. 57,58,64) there has been a singular lack of application to unknown peptides. Only two structural determinations have been published by the group: that of isariin (ref. 32), which has since been shown to be erroneous, and that of staphlomycin S (ref. 33). In a recent review article (1970, ref. 60), mention is made of the successful sequencing of three peptides from the partial hydrolysate of a protein; detailed results have not been published.

The aim of this group has been to develop mass spectrometry as a tool for the routine sequencing of small peptides from proteins. Reasons for doing this have been given: "mass spectrometry, with its great saving in time and in requirement of material used, is a major advance in the determination of the amino acid sequence of proteins" (Shemyakin, 1966 ref. 59). Any advantage of mass

spectrometry in protein analysis in either of these respects is dubious. In trying to compete on equal terms, rather than attack complementary problems not susceptible to normal chemical methods, Shemyakin and his co-workers have a difficult task. Only time will tell whether it proves to be impossible.