# POLYPEPTIDE SEQUENCING: USE OF DIPEPTIDYLAMINOPEPTIDASE I AND GAS CHROMATOGRAPHY/MASS SPECTROMETRY

R. M. Caprioli, W. E. Seifert Jr., and D. E. Sutherland
Department of Chemistry, Purdue University
West Lafayette, Indiana 47907

Received September 7, 1973

## Summary

A new method of sequencing polypeptides has been investigated in which the polypeptide is enzymatically hydrolyzed to dipeptides by the action of dipeptidylaminopeptidase I and the products identified by gas chromatography/mass spectrometry. The technique involves two hydrolyses, the original polypeptide and the des N-terminal amino acid polypeptide, providing overlapping dipeptide sequences. The technique is illustrated by the determination of the sequence of porcine insulin A chain, a polypeptide containing 21 amino acids. Methods and procedures using standard dipeptides are also described and the advantages and present limitations of the technique discussed.

One of the most promising new techniques for the determination of the sequence of amino acids in polypeptides involves enzymatic hydrolysis of the polypeptide to dipeptides by dipeptidylaminopeptidase I (DAP I)* (1) followed by analysis of the products by gas chromatography/mass spectrometry (GC/MS). A typical analysis would therefore first involve hydrolysis of the polypeptide, acylation and esterification of the dipeptides, and separation and identification of the derivatized dipeptides. Since this gives only the sequence of the individual dipeptides and not the order of the dipeptides, a second sample of the polypeptide whose N-terminal amino acid has been removed by the Edman technique would be hydrolyzed and analyzed as above. The sequence of the original polypeptide could then be deduced from this overlapping set of data.

Preliminary work by Ovchinnikov and Kiryushkin (2) has demonstrated the potentials of this approach by partially sequencing a peptide containing 14 amino acids. Others have utilized alternative analytical techniques for the separation and identification of the dipeptides, such as paper chromatography and column chromatography followed by end-group analysis (3,4). It is ap-

*The abbreviations used in the text are; DAP I for dipeptidylaminopeptidase I, GC/MS for gas chromatography/mass spectrometry, PFP for pentafluoropropionyl, cmc for carboxymethylcysteine, and mce for 2-mercaptoethanol.



*Copyright © 1973 by Academic Press, Inc.*
*All rights of reproduction in any form reserved.*

parent, however, that the use of this general technique for the routine sequencing of polypeptides with speed, accuracy, and high sensitivity rests upon the development of dipeptide analysis by GC/MS.

In this paper, we will describe a system for the analysis of dipeptides by GC/MS and the application of the DAP I-GC/MS technique for the sequence determination of carboxymethylated porcine insulin A chain. The advantages and disadvantages of the technique will be discussed, including the problems remaining and the direction in which we believe their solution lies.

## Materials and Methods

DAP I was isolated from fresh bovine spleen by the method of McDonald et al. (1). Edman degradations were performed according to the procedure of Blombäck and coworkers (5). Peptide digestion was carried out at pH 5.0 and 37° for four hours according to Callahan, et al. (3). The digest was then passed through an Amicon CF50A Centriflo membrane ultrafilter to remove the enzyme and the filtrate was lyophilized.

Pentafluoropropionyl (PFP) dipeptide methyl esters were prepared in the following manner: the lyophilized dipeptide samples were dissolved or suspended in 1 ml. methanol (spectral grade) and, after cooling in a dry ice - acetone slush bath, 0.4 ml. of thionyl chloride were slowly added. After several minutes, the mixture was heated at 45° for 30 min. followed by removal of the reagents in vacuo. The residue was dissolved in about 200 µℓ PFP anhydride and allowed to react at room temperature for 15 min. The excess reagent was removed in vacuo and the PFP-dipeptide methyl esters dissolved in 200 µℓ dry dioxane for injection into the gas chromatograph. Arginine containing dipeptides were first derivatized with 2,4-pentanedione (6) to yield the dimethylpyrimidylornithine analog and then derivatized as above.

Gas chromatographic analysis were performed using a Hewlett-Packard 5700A instrument with a 2 mm.(ID) x 2 m glass column packed with 1% Dexsil 300 on Gas Chrom Q (80-100 mesh). On-column injection was used with the injector temperature at 250°. GC/MS analysis was performed under the same conditions

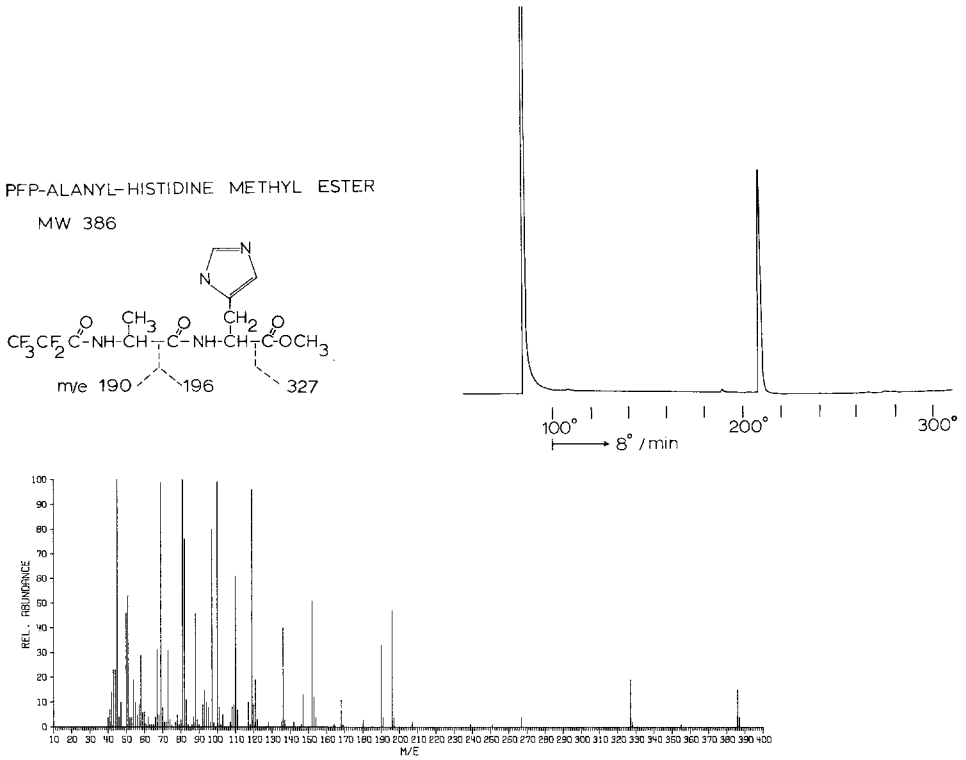


Figure 1 The gas chromatogram and the mass spectrum of PFP-alanylhistidine methyl ester.

using a Varian-Mat CH7 instrument. The mass spectra of the individual dipeptide standards were obtained using the direct insertion probe of a 21-110B double-focusing mass spectrometer.

## Results and Discussion

The PFP-dipeptide methyl esters give homogeneous peaks by gas chromatography, with a range of temperatures from 130° to 300°. This provides a means for gross separation of the derivatized dipeptides in the mixture so that identification can be made by mass spectrometry. The mass spectra of these compounds are easily interpretable, generally giving intense molecular ions and other ions at (M-15), (M-31) and (M-59) due to the loss of a methyl, methoxyl and carbomethoxy group, respectively. A major fragmentation of the molecular ion occurs between the alpha carbon of the N-terminal amino acid and the carbonyl carbon of the peptide bond with charge retention on the

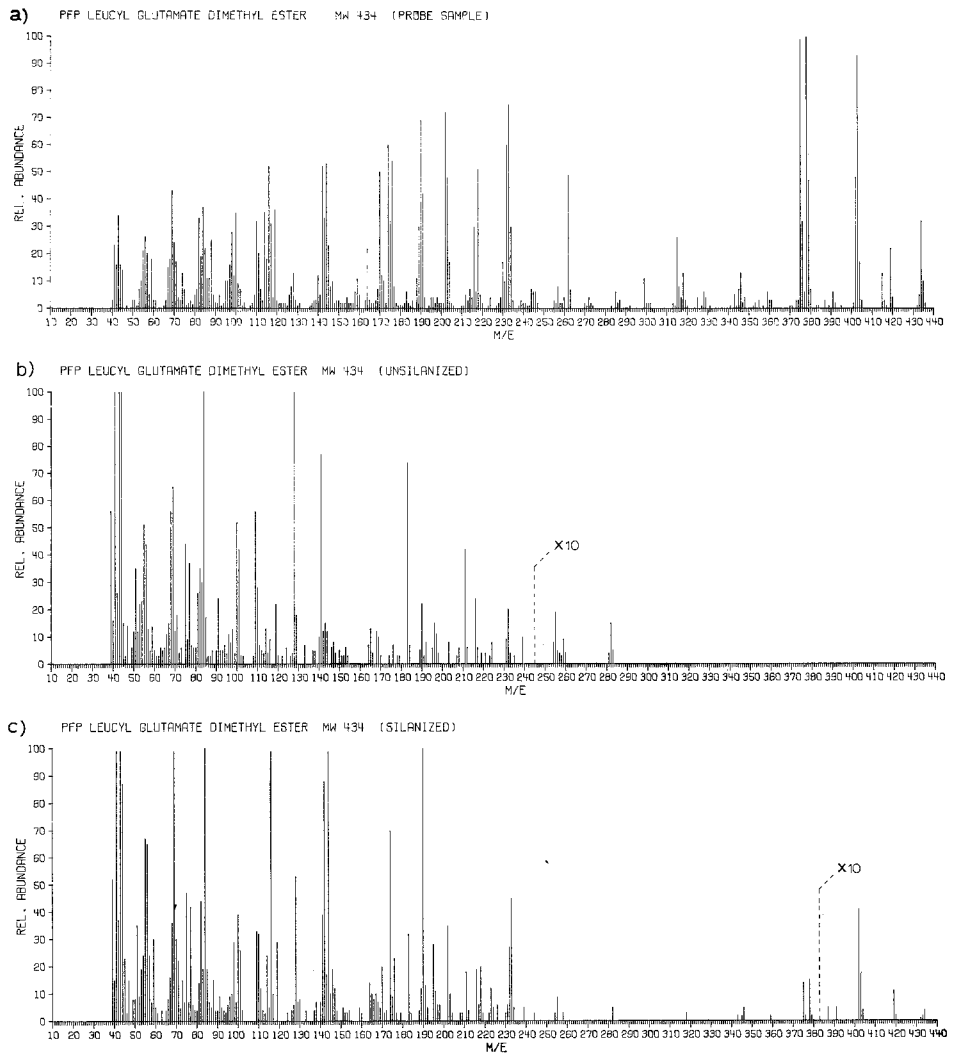


Figure 2 The mass spectrum of PFP-leucylglutamic acid dimethyl ester obtained (a) from the direct insertion probe, (b) by GC/MS using a glass-frit separator, and (c) by GC/MS using a silanized glass-frit separator.

acylated fragment. This ion together with the molecular ion is usually sufficient to identify the dipeptide. Many other ions are also formed which are derived from both ends of the molecule, giving diagnostic ions with which to sequence the dipeptide. For example, Figure 1 shows the gas chromatogram and the mass spectrum of PFP-alanylhistidine methyl ester. The molecular ion appears at m/e 386 with an ion due to loss of the carbomethoxy group at m/e

327. Two major sequence ions can be seen, one at m/e 190 due to the N-terminal fragment and the other at m/e 196 due to the C-terminal fragment. Either of these fragment ions together with the molecular ion or the two fragment ions alone are sufficient to identify the dipeptide. The gas chromatographic properties and mass spectra of a variety of dipeptides have been studied and will be published elsewhere.

One of the difficulties encountered when using combined GC/MS was decomposition of some of the dipeptides in the glass-frit separator. This is illustrated in Figure 2 for PFP-leucylglutamic acid dimethyl ester. It can be seen that spectrum (b) is completely different from spectrum (a) and is basically the spectrum of decomposition products, while spectrum (c) is essentially the same as spectrum (a) with the exception of less intense high mass ions. This shows that although decomposition takes place in the separator, it can largely be eliminated by silanization of the separator. In the work which follows, the GC/MS system was silanized to minimize decomposition. In the few cases where high mass ions were still not seen, sequence ions derived from both ends of the molecules were used to identify the dipeptides. In the figures shown later, these dipeptides are marked with an asterisk.

In order to illustrate the sequencing technique, two mixtures of dipeptides were prepared to simulate the two DAP I digests of a tryptic peptide (one for the original polypeptide and the other for the des-amino acid polypeptide) having the sequence,

tyr-leu-gly-ala-thr-gly-ile-ala-trp-phe-leu-gly-ala-met-glu-ala-arg

In the total ion monitor recordings shown in Figure 3, each peak was identified from its mass spectrum. The two sets of data were submitted to computer analysis to reconstruct the "original polypeptide".

Porcine S-carboxymethylated insulin A chain (Schwarz/Mann), a polypeptide containing 21 amino acids of known sequence, was used as a test sample. The total ion monitor recording of the derivatized products of the DAP I hydrolysis of this polypeptide is given in Figure 4. The first peak, labeled (PFP-

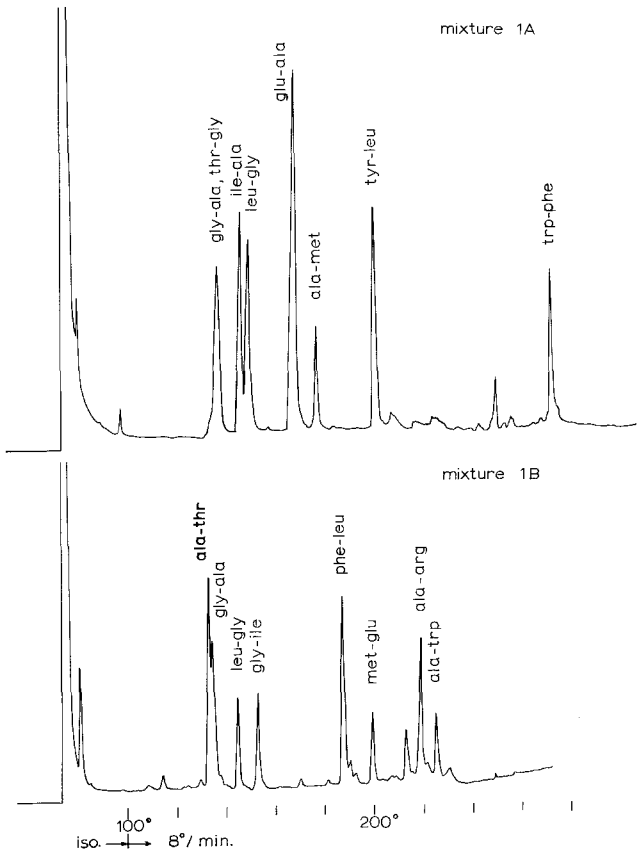


Figure 3 The total ion monitor recordings of dipeptide mixtures simulating the products of digestion of a tryptic peptide, where mixture 1A are those of the original peptide and mixture 1B those of the des N-terminal amino acid peptide.

$mce)_2$, is due to the diacylated dimer of 2-mercaptoethanol, a buffer constituent for the enzymatic hydrolysis. A peak was observed and identified for each of the 10 possible dipeptides of the carboxymethylated insulin A chain. Asparagine and glutamine residues were analyzed as their respective acids due to hydrolysis during derivatization. The total ion monitor recording of the products of hydrolysis of the des-glycine insulin A chain obtained from one round of Edman degradation is also given in Figure 4. Each of the 10 possible dipeptides was again identified by its mass spectrum. The total ion recording for the des-gly chain is somewhat more complicated because the N-

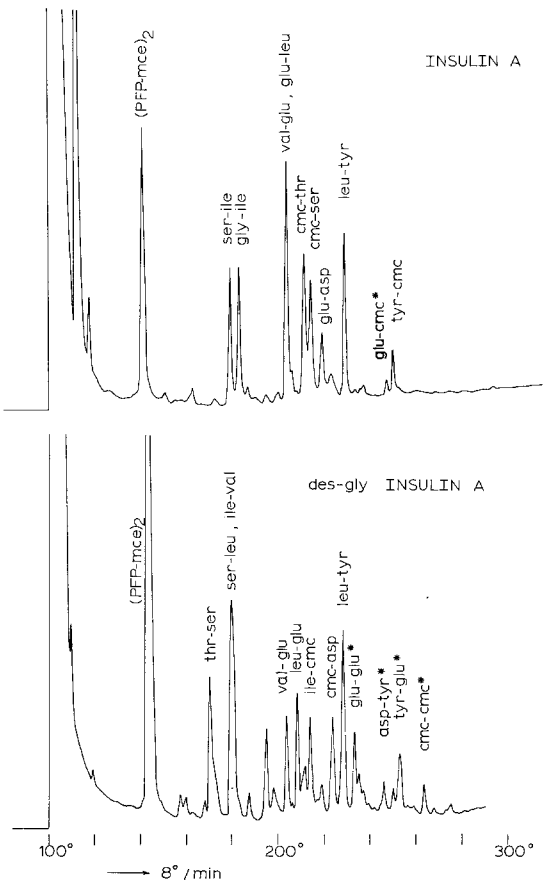


Figure 4 The total ion monitor recordings of the digests of carboxymethylated insulin A chain and des-gly insulin A chain. For those dipeptides marked with an asterisk, partial decomposition was observed.

terminal glycine was not quantitatively removed in the Edman reaction, a well known difficulty with glycine residues. However, the dipeptides derived from the 10-15% of the insulin A chain remaining did not cause difficulty in the identification of the des-gly dipeptides. The N-terminal glycine removed as the PTH-derivative was also identified by its mass spectrum.

From these two overlapping sets of data, the possible sequences for insulin A-chain were deduced. If the amides of the acidic amino acids could be distinguished, then the overlapping sets of dipeptides would yield two possible sequences for insulin A-chain. However, due to our present inability

to make this distinction, the number of possible sequences is increased to twelve. Nevertheless, we were able to reduce these twelve to the single correct sequence by providing additional information through a time-course study. Thus, when insulin A-chain was hydrolyzed, samples were taken with time and analyzed by GC/MS. From this, certain correlations were made; leu-tyr appeared before glu-asp, cmc-thr before tyr-cmc, cmc-thr before cmc-ser, ser-ile before glu-leu, and val-glu before cmc-thr. Of these correlations, only three are necessary to specify the correct sequence: val-glu before cmc-thr, leu-tyr before glu-asp, and ser-ile before glu-leu. Thus, the sequence of S-carboxymethylated porcine insulin A-chain is,

gly-ile-val-glx-glx-cmc-cmc-thr-ser-ile-cmc-
ser-leu-tyr-glx-leu-glx-asx-tyr-cmc-asx

The results presented above show that the DAP I-GC/MS technique is an extremely promising method for the determination of the amino acid sequence of polypeptides in terms of speed, sensitivity, and accuracy. However, before becoming a viable alternative to classical sequencing methods, several remaining problems must be solved. The most serious of these concerns the specificity of DAP I toward proline containing peptides, i.e., the enzyme cannot hydrolyze peptide bonds involving proline. We are currently investigating two approaches to the solution of this problem, the first involving the use of a second enzyme, DAP IV (7), which has been shown to preferentially split proline bonds to form dipeptides. Further work is necessary, however, to better characterize the specificity of this enzyme. The second approach basically involves the use of two rounds of Edman degradation. A second specificity problem of DAP I is its inability to hydrolyze a dipeptide from a polypeptide when either lysine or arginine is in the N-terminal position. This is a less serious problem than that for proline, since tryptic peptides could be used as substrates. Nevertheless, we are investigating chemical derivatization methods to mask these residues in the polypeptide and the use of other enzymes (DAP II and III) (7) to eliminate this problem. Other procedures for

the formation of methyl esters of dipeptides are being studied to prevent hydrolysis of asparagine and glutamine to aspartic and glutamic acids, respectively. With the solution of these problems we believe the DAP I-GC/MS technique will constitute a primary method for sequencing polypeptides.

Acknowledgements: The authors wish to thank Mr. Carter Cook of the Mass Spectrometry Center, Univ. Illinois, for his aid and the use of the GC/MS. W.E.S., Jr., is a David Ross Fellow and thanks the Purdue Research Foundation for support.

References:

1. McDonald, J.K., Callahan, P.X., and Ellis, S. (1972) in Methods of Enzymology, XXV, (C.H.W. Hirs and S.N. Timasheff, eds.), pp. 272-281, Academic Press, New York.
2. Ovchinnikov, Y.A., and Kiryushkin, A.A. (1972) FEBS (Fed. Eur. Bio. Soc.) Lett. 21, 300-302.
3. Callahan, P.X., McDonald, J.K., and Ellis, S. (1972) in Methods of Enzymology, XXV, (C.H.W. Hirs and S.N. Timasheff, eds.), pp. 282-298, Academic Press, New York.
4. Lindley, H. (1972) Biochem. J. 126, 683-688.
5. Blombäck, B., Blombäck, M., Edman, P., and Hessel, B. (1966) Biochim. Biophys. Acta 115, 371-396.
6. Morris, H.R., Dickinson, R.J., and Williams, D.H. (1973) Biochem. Biophys. Res. Commun. 51, 247-255.
7. McDonald, J.K., Callahan, P.X., Ellis, S., Ellis, S., and Smith, R.E. (1971) in Tissue Proteinases (A.J. Barrett and J.T. Dingle, (eds.) pp. 69-107, North-Holland Publishing Co., Amsterdam.