Mass Spectrometric Studies of Peptides. III.[1] Automated Determination of Amino Acid Sequences

Martin Senn, R. Venkataraghavan, and F. W. McLafferty[2]

Contribution from the Department of Chemistry, Purdue University, Lafayette, Indiana 47907. *Received June 13, 1966*

Abstract: By utilizing the exact mass of the N-terminal group, the amino acid sequence can be determined from a variety of oligopeptides whose components include 14 of the known amino acids. Most derivatives giving suitable sample vapor pressure and thermal stability can be used if a mass measuring accuracy of ±2 mmu is available. All histidine-containing peptides run as the methyl esters appear to have a H atom replaced by CH_3. Automatic measurement of the spectral data plus computer calculation and interpretation of these data give promise that this method may find general use in protein research.

Key information leading to increased understanding of a number of important biological processes has been provided by the determination of the amino acid sequence in particular proteins. Present common practice for such determinations[3] involves enzymatic digestion of the protein followed by column and paper chromatographic separation and purification of the oligopeptide products. The sequence of amino acids in these peptides is then determined by a stepwise degradation procedure which is monitored by total hydrolysis with amino acid analysis by column chromatography, a time-consuming procedure requiring relatively large samples. An alternate source of structure information on particular oligopeptides which utilizes much smaller samples is indicated by recent publications on the mass spectra of linear and cyclic peptides and depsipeptides.[4] We have recently reported[5] a mass spectrometric method which appears to be much more generally applicable and unequivocal, and is amenable to automated calculations. This method is based on recognition of the fact that, barring rearrangements, *the structure of a linear molecule is determined unequivocally by using only the possible fragments which contain one end of the chain.* Thus in a hypothetical molecule A–B–C–D–E–F only six pieces of the molecule containing part A are possible without rearrangement, and determination of these will find the sequence of the parts unequivocally. Many more combinations of these parts are possible, but their identification is not necessary for the sequence determination.

A number of methods for marking the end of the peptide chain are possible;[4a,5] high-resolution mass spectrometry appears to provide a general method.[5] Application of the latter to the determination of amino acid sequences has also been reported recently in parallel independent work by Biemann, Cone, and Webster.[6] For almost any peptide derivative of sufficient volatility, the presence of the terminal functional group makes a unique mass contribution to the exact mass of a particular fragment ion which can be combined with the predictable mass contributions of known amino acids to identify the fragments containing the terminal group. This method demands mass measurements of high accuracy on several hundred peaks in the spectrum of an average oligopeptide; such determinations are now possible on several spectra per hour using a fully automatic comparator–microdensitometer and computer techniques.[5,7] In the variety of peptide structures and derivatives examined to date it appears to be possible to extend the use of a computer to the complete characterization of the amino acid sequence in oligopeptides.

To obtain sufficient volatility and stability, both of the terminal amino and carboxy groups are converted to suitable derivatives. Table I shows examples of the types of compounds from which satisfactory spectra are obtained. A number of other peptides examined gave results similar to these. For some other samples, suitable derivatives could not be prepared, and investigation of these is continuing. Note, however, that among

Table I

	Peptide	Temp[a]
I	N-Trideuterioacetyl-Ala-Leu-Ala-Val-Val-Val methyl ester	210
II	N-Trifluoroacetyl-Leu-Gly-Phe methyl ester	150
III	N-Acetyl-Pro-Gly-Phe-Gly methyl ester	190
IV	N-Trifluoroacetyl-His-Pro-Tyr methyl ester	200
V	N-Trifluoroacetyl-His-Met-(β-O-methyl-Asp) methyl ester	190
VI	N-Trifluoroacetyl-Pro-Phe-His-Leu methyl ester	215
VII	N-Trideuterioacetyl-Pro-Phe-His-Leu-Leu methyl ester	235
VIII	N-Carbobenzoxy-Val-(O-*t*-butyl-Glu) methyl ester	140
IX	N-Carbobenzoxy-Ile-(S-benzyl-Cys)-Ser methyl ester	190
X	N-Carbobenzoxy-Val-Gly-Ala-Leu-Ala methyl ester	200

[a] Ion source temperature (°C) at which the mass spectrum was obtained.

(1) Paper II: S. Takeuchi, M. Senn, R. W. Curtis, and F. W. McLafferty, submitted to *Phytochemistry*.

(2) To whom request for reprints should be sent.

(3) For example, see C. H. Li, W. K. Liu, and J. S. Dixon, *J. Am. Chem. Soc.*, **88**, 2050 (1966).

(4) (a) E. Bricas, J. van Heijenoort, M. Barber, W. A. Wolstenholme, B. C. Das, and E. Lederer, *Biochemistry*, **4**, 2254 (1965); (b) K. Heyns and H. F. Grützmacher, *Ann. Chem.*, **669**, 189 (1963); (c) B. J. Millard, *Tetrahedron Letters*, 3041 (1965); (d) F. Weygand, A. Prox, H. H. Fessel, and K. K. Sun, *Z. Naturforsch.*, **20b**, 1169 (1965); (e) N. S. Wulfson, V. A. Puchkov, B. V. Rozinov, A. M. Zyakin, M. M. Shemyakin, Yu. A. Ovchinnikov, A. A. Kiryuskin, and V. T. Ivanov, *Tetrahedron Letters*, 2793 (1965).

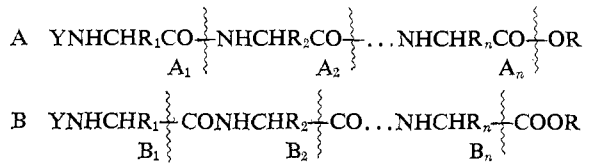
(5) M. Senn and F. W. McLafferty, *Biochem. Biophys. Res. Commun.*, **23**, 381 (1966).

(6) K. Biemann, C. Cone, and B. R. Webster, *J. Am. Chem. Soc.*, **88**, 2597 (1966).

(7) F. W. McLafferty, *Science*, **151**, 641 (1966).

the components of the peptides of Table I are *14* of the natural amino acids.

The fragmentation patterns of a variety of peptides show[4] that cleavage of the chain involves two main pathways.



These cleavages may also be accompanied by the rearrangement gain or loss of a hydrogen atom. The relative probabilities of these reactions depend on the particular peptide and derivative. "Amine fragments" (those containing the N-terminal group Y) are generally much more abundant than "ester fragments" (those containing the –COOR group). In suitable derivatives of peptides containing only Gly, Ala, Val, Leu, Pro, or Phe, almost all fragment ions corresponding to scheme A and scheme B could be detected; in I–III, only the $(M - COOR)^+$ peak (B_6) in I is not found. For the histidine-containing peptides IV–VII several fragments from scheme A or B, or both, are missing.

The complementary ion fragments containing the terminal COOR group and part of the peptide chain decrease in abundance very rapidly with increase of the number of amino acids in the particular ion fragment. Ester peaks corresponding to $H_2N^+{=}CR_nCOOR$ and $H_3N^+CHR_nCOOR$ (containing only the terminal amino acid) are usually detectable.

These pathways account for only a small fraction of the detectable ions in the observed spectra, but fortunately the exact masses of these ions usually are unique. Ambiguities can arise in the case of valine or leucine owing to the rearrangement loss of side-chain C_3H_6 or C_4H_8, respectively, yielding a peak which is identical in mass with the corresponding glycine-containing fragment. In some cases, such as in the spectrum of II, such rearrangement peaks are prominent; but in all spectra examined such peaks are of lower intensity than the corresponding peaks due to nonrearranged ions. The occurrence of such a rearrangement is an obvious conclusion when both glycine and valine (or leucine) are indicated as the next amino acid in the sequence.

A surprising number of derivatives appear to be of a suitably unique mass if the mass measurement accuracy is better than 3 millimass units (mmu). Although N-terminal derivatives containing relatively low proportions of hydrogen have the greastest advantage in terms of mass defects, the exact mass contributions of even CD_3CO and, to a lesser extent, CH_3CO are sufficiently distinct in most cases. Such acetyl derivatives have the advantage of simplicity of preparation and high volatility, and give suitable fragmentation patterns for many types of peptides. However, more investigation is necessary to find optimum derivatives for particular amino acids from the point of view of both thermal stability and fragmentation patterns. For example, of the carbobenzoxy derivatives examined, the dipeptide VIII gives the peaks predicted by both schemes A and B in low abundance, while the tripeptide IX only yields the amine fragments containing the N-terminal amino acid (A_1 and B_1), although a molecular ion is produced. Pentapeptide X gives no sequence peaks corresponding to scheme A and B and no molecular ion, although other peaks in the spectrum are consistent with its structure.

Although satisfactory spectra can sometimes be obtained from peptides containing the free carboxyl group, it is more satisfactory in general to convert this to the methyl ester. The trideuteriomethyl ester derivative usually increases the ease of distinguishing the masses of the $H_2N^+{=}CR_nCOOR$ and H_3N^+-CHR_nCOOR peaks.

General Method of Sequence Determination. Amino acid sequence is elucidated utilizing a computer program of the following general format. The initial step is the identification of all peaks corresponding in mass to peaks possible from fragmentation schemes A and B. The spectrum is first checked for a peak corresponding to the sum of the exact masses of the N-derivative moiety and the glycine unit $NHCH_2CO$, for example, 46.03722 (CD_3CO) + 57.02146 = 103.05868 (see Table II). The search is repeated using combinations of the

Table II

N derivatives	Mass		
CH_3CO	43.01839		
CD_3CO	46.03722		
CF_3CO	96.99012		
Amino acid	Mass	Amino acid	Mass
Gly	57.02146	Asp-O-CH_3	129.04259
Ala	71.03711	Met	131.04048
Ser	87.03203	His	137.05891
Pro	97.05276	Glu-O-CH_3	143.05824
Val	99.06841	Phe	147.06841
Thr	101.04768	Tyr	163.06332
Leu	113.08406	Try	186.07931

N derivative with each of the other possible amino acids. Fragment ions corresponding to scheme B are checked by subtracting the exact mass of CO (27.99491 amu) from each of these combinations. Correspondence within experimental error of a fragment from either scheme indicates the N-terminal amino acid. If more than one amino acid is indicated as the next unit (ordinarily due to an experimental artifact or measuring error), an attempt is made to determine the next amino acid in the sequence for both of the possibilities; this almost always resolves the ambiguity. To identify the next amino acid unit of the chain, this process is repeated, with the addition of the mass (value from Table II) of the newly identified amino acid unit to the sum of exact masses described above. This process is repeated until the search for an additional chain member finds no suitable match for either A or B. A check is now made for the molecular ion by adding the exact mass of the ester group OR to the mass of the identified chain, A_n. A fit then establishes the molecular size, and the identification is complete.

For some larger peptides neither a peak corresponding to the next amino acid by paths A or B nor one corresponding to the addition of OR is found. In these cases a check for a molecular ion at higher mass is made using the sum of A_n + OR + each of the amino acid fragment masses (Table II). A similar sequence determination can be attempted from the C-terminal

```
1 AMINO ACID SEQUENCE ANALYSIS

-SAMPLE NUMBER 66-2

 N-TRIDEUTERDACETYL METHYL ESTER

- THE FOLLOWING ARE THE SEQUENCE PEAKS FOUND IN THIS PEPTIDE

  SEQUENCE IDENTITY        FOUND          CALC          ERROR

     A-1 ALA          89.079700     89.079410      -0.29
     B-1 ALA         117.073170    117.074321       1.15

     A-2 LEU         202.164000    202.163470      -0.53
     B-2 LEU         230.158401    230.158380      -0.02

     A-3 ALA         273.201201    273.200500      -0.70
     B-3 ALA         301.196100    301.195491      -0.61

     A-4 VAL         372.267500    372.268990       1.49
     B-4 VAL         400.264310    400.263900      -0.41

     A-5 VAL         471.336600    471.337400       0.80
     B-5 VAL         499.333301    499.332310      -0.99

     B-6 VAL         598.402000    598.400721      -1.28

     MOL-ION         629.421000    629.419110      -1.89

- AMINO ACID SEQUENCE IN THIS PEPTIDE IS

        *(D)AC-ALA *LEU *ALA *VAL *VAL *VAL*
```

Figure 1. Output from computer interpretation of the mass spectrum of CD_3CO-Ala-Leu-Ala-Val-Val-Val-OCH_3 (I).

end using the ester fragments, making possible the structure elucidation of peptides which give no molecular ion. Some typical fragment peaks, such as $(M - H_2O)^+$, can also serve to indicate the molecular ion when it is absent. For example, if an amino acid which typically loses H_2O has been identified in the earlier part of the computer program, the peak corresponding to $(A_n + OCH_3 - H_2O)$ should be the $(M - H_2O)$ ion.

Certain peaks from other fragmentation pathways are also useful to corroborate the sequence found. The use of "dipeptide fragments" containing two amino acid units will be illustrated in the spectrum of I.

Examples

CD_3CO-Ala-Leu-Ala-Val-Val-Val-OCH_3 (I). The amino acid sequence as determined by the computer program is shown in Figure 1; the bar graph of this spectrum at unit mass resolution is shown in Figure 2. Peaks due to $(M - 15)^+$ and $(M - 42)^+$ (rearrangement loss of the valine side chain) are noteworthy. Ions corresponding to the loss of the leucine chain are at very low intensity on the photoplate, and there is no B_6 fragment, $(M - COOCH_3)^+$. Some peaks from pathways A and B are accompanied by peaks containing one more or one less hydrogen atom, such as fragment B_2. Masses 170, 184, and 198 have elemental compositions corresponding to the "dipeptide fragments" Ala-Val, Ala-Leu, and Val-Val, respectively; no peaks are found corresponding in mass to Ala-Ala or Leu-Val. The base peak in the spectrum at mass 72 corresponds to the amine fragment of valine, $C_4H_{10}N$; the corresponding leucine fragment is at

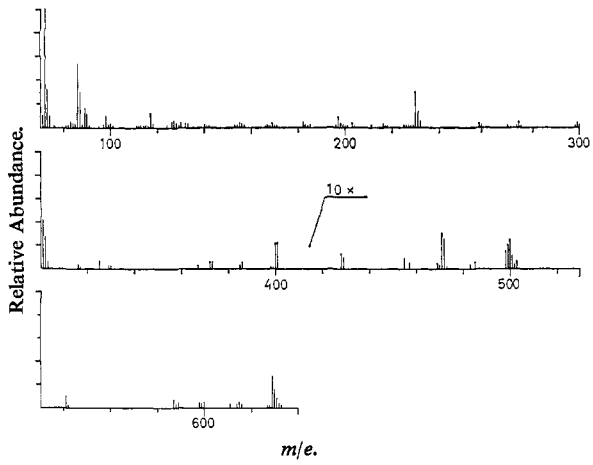


Figure 2. CD_3CO-Ala-Leu-Ala-Val-Val-Val-OMe.

m/e 86 ($C_5H_{12}N$). The two ester peaks also indicate that valine is the C-terminal amino acid.

CF_3CO-Leu-Gly-Phe-OCH_3 (II). See Table III. The spectrum shows abundant A_1 and B_1 fragment peaks which could indicate a terminal CF_3CO-Gly.

Table III. N-Trifluoroacetyl-Leu-Gly-Phe-OCH_3

Scheme A Measured masses	Calcd sequence	Scheme B Calcd masses	Scheme B Measured masses
	96.99012		
	113.08406 (Leu)		
210.0726 [1.6][a]	210.07418	182.07927	182.0781 [1.1]
	57.02146 (Gly)		
267.0956 [0.0]	267.09564	239.10073	239.0991 [1.6]
	147.06841 (Phe)		
414.1647 [0.7]	414.16405	386.16914	386.1682 [0.9]
	31.01839 (OCH_3)		
445.1829 [0.5]	445.18244		

[a] Error in millimass units.

This conclusion would then lead to finding leucine as the second amino acid. However, the computer search of all possible amino acid combinations for fragments A_1 and B_1 also finds correspondence for leucine as the N-terminal amino acid. Thus the misleading peak suggesting glycine is actually the peak from the rearrangement loss of C_4H_8 from the leucine side chain and the computer output indicates the correct sequence CF_3CO-Leu-Gly-.

CH_3CO-Pro-Gly-Phe-Gly-OCH_3 (III). The computer solution of the sequence is shown in Figure 3. Corroboration is provided by dipeptide fragments at mass 154 for Pro-Gly and at mass 204 for Phe-Gly, and by a very intense peak at *m/e* 120 corresponding to the amine fragment of phenylalanine.[8]

This spectrum shows an interesting indication of the importance of mass-measuring accuracy. Small peaks are found corresponding in mass to CH_3CO-Ala and CH_3CO-Ala-Pro with errors of 4 and 6 mmu, respec-

(8) H. F. Grützmacher and K. Heyns, Advances in Mass Spectrometry, Proceedings of a Joint Conference, Vol. 3, W. L. Mead, Ed., Institute of Petroleum, London, 1966, p 655.

Table IV

IV	V	VI	VII	Fragment
234.0505 (1.5)[b] His	234.0476 (1.4) His	194.0445 (1.6) Pro	143.0888 (1.1) Pro	A_1
248.0657 (1.9)	248.0642 (0.4)			$A_1 + CH_2$
206.0534 (0.7) (His)[c]	206.0523 (1.8) (His)	166.0484 (0.5) (Pro)	115.0952 (0.2) (Pro)	B_1
220.0694 (0.6)	220.0678 (1.9)			$B_1 + CH_2$
331.1003 (1.5) Pro	*a*	341.1114 (0.1) Phe	290.1568 (1.5) Phe	A_2
345.1157 (1.7)	*a*			$A_2 + CH_2$
303.1059 (1.0) (Pro)	337.0965 (1.9) Met	313.1183 (1.9) (Phe)	262.1640 (0.5) (Phe)	B_2
317.1229 (0.4)	351.1083 (1.9)			$B_2 + CH_2$
a	*a*	*a*	*a*	A_3
	Asp			
a	508.1482 (0.5) OCH_3	*a*	441.2341 (1.2) His	$A_3 + CH_2$
a	*a*	450.1772 (2.0) His	*a*	B_3
a	*a*	464.1929 (1.9)	413.2364 (1.6) (His)	$B_3 + CH_2$
		a	*a*	A_4
		a	*a*	$A_4 + CH_2$
		a	*a*	B_4
		a	526.3202 (1.8) Leu	$B_4 + CH_2$
			a	A_5
			a	$A_5 + CH_2$
			a	B_5
			a	$B_5 + CH_2$
525.1848 (1.3) Tyr	*a*	*a*	*a*	$M \cdot^+$
539.1995 (0.4)	539.1645 (1.6) OCH_3	636.2880 (0.3) Leu	698.4178 (1.6) Leu	$M \cdot^+ + CH_2$

[a] No peak found corresponding to this fragment. [b] Number in parentheses indicates error in millimass units. [c] Abbreviation in parentheses indicates duplicate assignment of amino acid fragment.

tively. This structural evidence would have caused considerable confusion if it had not been eliminated by the substantially smaller error limits of the exact mass determination. Use of some other derivative such as CD_3CO in place of CH_3CO generally decreases the possibility of such ambiguities.

```
1 AMINO ACID SEQUENCE ANALYSIS

-SAMPLE NUMBER 66-1

 N-ACETYL METHYL ESTER

- THE FOLLOWING ARE THE SEQUENCE PEAKS FOUND IN THIS PEPTIDE

   SEQUENCE          FOUND          CALC          ERROR
   IDENTITY

    A-1 PRO      112.074300     112.076240      1.94
    B-1 PRO      140.071199     140.071150     -0.05

    A-2 GLY      169.095800     169.097748      1.95
    B-2 GLY      197.090799     197.092659      1.86

    A-3 PHE      316.164799     316.164299     -0.50
    B-3 PHE      344.160500     344.159210     -1.29

    A-4 GLY      373.189301     373.187046     -2.25
    B-4 GLY      401.182598     401.181957     -0.64

    MOL-ION      432.200401     432.200985      0.58

- AMINO ACID SEQUENCE IN THIS PEPTIDE IS

        - AC-PRO *GLY *PHE *GLY *
```

Figure 3. Output from the computer interpretation of the mass spectrum of CH_3CO-Pro-Gly-Phe-Gly-OCH_3.

Histidine-Containing Peptides. The spectra of all of the oligopeptides containing histidine, IV–VII, exhibit anomalous peaks 14 mass units above the molecular weight. (Only IV gives a peak corresponding in mass to the expected molecular ion.) Exact mass measurements show that this difference corresponds to CH_2, indicating the replacement of a hydrogen atom by a methyl group. To ascertain the location of this postulated methyl group, the sequence determination was also made searching for masses corresponding to CH_2 above the key fragments of sequences A and B. The results, shown in Table IV, indicate that the extra methyl group is located on the histidine moiety. The abundances of the sequence peaks containing the nonmethylated histidine are too low to be detected in many cases.

These anomalous peaks may arise from some sort of a transmethylation which occurs on heating the samples in the mass spectrometer; similar phenomena have been observed.[9] Alternatively, the esterification to form the methyl ester may have also alkylated a nitrogen atom on the histidine ring. Further investigation to elucidate this apparent anomaly is in progress. However, the effect does not appear to be unique to our esterification procedure, as compound IV was supplied to us as the methyl ester. Esterification of compound V was repeated using the same procedure, but with the substitution of CD_3OD for CH_3OH. The B_3 sequence peak at mass 464 (Y-Pro-Phe-His) is increased in mass to correspond to the trideuterio species (error 2.2 mmu), and the $(M + 14) \cdot^+$ at mass 642 is increased in mass to correspond to the hexadeuterio species (error 0.4 mmu).

As an example of the value of high resolution, compound V gives a substantial peak at *m/e* 465, whose nominal mass could indicate the expected CF_3CO-His-Met-AspOCH_3– (no extra methyl group on the histidine). However, the exact mass measurement of the peak disagrees with that expected for this ion by 14 mmu; the elemental composition actually corresponds to the loss of C_3H_6S from the molecular ion of the peptide containing the additional CH_2.

(9) D. W. Thomas and K. Biemann, *J. Am. Chem. Soc.*, 87, 5447 (1965).

Preparation of Peptide Derivatives

Difficulty was encountered with early derivative preparations which resulted in discoloration of derivatives, thermal decomposition of samples in the ion source, and poor reproducibility of spectra. Most of these difficulties were overcome by using purified reagents and solvents, and by avoiding any contact with metal or dust particles.

Esterification. Approximately 0.1–0.6 μmole of the peptide is dissolved in 40 ± 10 μl of 2–2.5 *N* HCl in absolute CH_3OH in a reaction flask made by blowing a 6-mm bulb in the end of a 2-mm capillary melting point tube. The top of the capillary is sealed and allowed to stand overnight at room temperature. The capillary then is opened and the HCl–CH_3OH evaporated *in vacuo*.

N-Acylation. The residue is dissolved in 40 ± 10 μl of the following: 1:1 acetic acid–acetic anhydride for the preparation of N-acetyl derivatives; CD_3COOH[10]–CH_2Cl_2 for N-trideuterioacetyl derivatives; and 1:2 CF_3COOH–$(CF_3CO)_2O$ for N-trifluoroacetyl derivatives. The solution is transferred to an unmodified melting point capillary, sealed, and allowed to stand 4–6 hr at room temperature. The capillary is then opened, and the reagents are evaporated off *in vacuo*.

After evaporation, the product generally is distributed on the sides as well as the bottom of the capillary tube. Cutting the capillary tube can usually give two parts, each with sufficient sample for a mass spectrum.

This sample tube is directly introduced into the ion source of a CEC 21-110 high-resolution mass spectrometer. The ion source temperatures necessary for adequate volatilization depend upon the nature of the sample (see Table I). Thermal decomposition can be a problem at higher molecular weights; this can usually be detected by recording a separate spectrum at an increased temperature. Using a resolving power of approximately 12,000, spectra are recorded directly on a photoplate with perfluorokerosene as an internal mass standard. Optical densities at 0.25-μ intervals across each ion line are measured automatically with a Grant-Datex microdensitometer–comparator.[7] From these data computer calculation of exact masses is possible to an error limit of ±2.5 mmu.

(10) From Merck Sharp and Dohme of Canada.

Conclusions

Much needs to be done before this technique can be used as a routine tool for amino acid sequence determinations, but the prospects are promising. At present the method appears to be applicable to the majority of amino acids. It is capable of giving information not previously possible, such as the presence of methylated histidine. Sample requirements are now below 0.1 μmole, and current work indicates that this can be substantially reduced. The classical sequence characterization of an oligopeptide requiring a number of degradations and amino acid analyses can be replaced in this scheme by one mass spectral determination, and the mass spectral measurements and calculations appear to be amenable to complete automation. Our current research program in this area includes a search for suitable derivatives of peptides containing the remaining common amino acids, use of other features of the mass spectrum and use of the mass spectra of other derivatives as corroborative evidence for the sequence found, and application of these techniques to completely unknown oligopeptides isolated from protein hydrolysates.

Acknowledgments. The authors are indebted to Dr. A. Light of this department for helpful discussions; to R. D. Board for technical assistance; to the following for valuable supplies of peptide samples: Drs. G. Anderson, Lederle Laboratories, F. M. Bumpus, Cleveland Clinic, K. Hofmann, University of Pittsburgh, P. G. Katsoyannis, Brookhaven National Laboratory, C. H. Li, University of California, Berkeley, E. L. Smith, University of California, Los Angeles, B. Witkop, National Institutes of Health; and to the National Institutes of Health (GM 12755) and the National Aeronautics and Space Administration (NGR 15-005-021) for generous financial support.