Determination of the Amino Acid Sequence in Oligopeptides by Computer Interpretation of Their High-Resolution Mass Spectra[1]

K. Biemann, C. Cone, B. R. Webster,[2] and G. P. Arsenault[3]

Contribution from the Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. *Received July 20, 1966*

Abstract: A technique for the amino acid sequence determination of oligopeptides based on the computer interpretation of the high resolution mass spectra of certain peptide derivatives is described. It has been successfully applied to di- through pentapeptides containing glycine, alanine, serine, proline, valine, threonine, leucine, asparagine, α-aminoadipic acid, lysine, glutamic acid, methionine, histidine, phenylalanine, tyrosine, and S-benzylcysteine and with acetyl, trifluoroacetyl, chlorodifluoroacetyl, benzoyl, carbobenzoxy, phthaloyl, chlorophthaloyl, tosyl, and 3-hydroxydodecanoyl as N-terminal substituents. Because of its speed, sensitivity (a few micrograms yield a usable spectrum), relative tolerance toward contaminants, and objectivity of interpretation, the technique should prove useful in the course of the many sequence determinations required for the determination of the structure of a protein.

The determination of the amino acid sequence in oligopeptides is a formidable experimental task which is of crucial importance in three areas: first and above all, in the determination of the primary structure of proteins; second, in the determination of the much smaller nonprotein peptides such as some antibiotics and related metabolites; and third, to verify the sequence (and sequential purity) of a synthetic peptide. While these three cases involve the same common problem, namely elucidation of the consecutive arrangement of amino acids linked by amide bonds, many important practical aspects of these systems differ considerably.

In the course of the determination of the primary structure of a protein molecule of reasonable size a few hundred sequences have to be determined by analyzing the sequences in a very large number of small peptides produced by a series of specific (enzymatic or chemical) or random (partial acid hydrolysis) cleavages performed in the molecule. The major problem is due to the very large number of peptides to be separated, purified, and analyzed. The advantage is the fact that a careful amino acid analysis on the original protein reveals the kind and approximate molar ratios (or even exact number) of amino acids present, and that these are of limited and known structural variability. Furthermore, on the basis of our present knowledge of primary peptide structure, it can be taken for granted that the vast majority, if not all, of the consecutive units involve α-amino acid amide linkages.

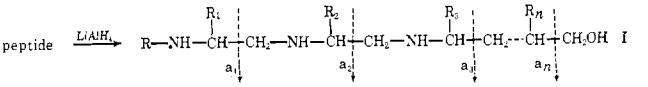
The small, nonprotein peptides found, for example, as microbial metabolites contain a much smaller number of amino acids and peptide bonds, but the acids may be of new or unexpected structure, and the bonding may involve groups other than the amino and carboxyl groups of α-amino acids, or even structural elements differing altogether from amino acids.

Finally, synthetic peptides pose the question whether there are components present which originate from an incomplete or side reaction in one of the intermediate stages of a multistep synthesis, particularly if the product of each single step is not individually purified and characterized. An extreme case is Merrifield's elegant "solid phase" technique,[4] which could conceivably lead to the absence of a particular unit in the final product. Also in this case one deals with a relatively short peptide (as contrasted to a protein) and much is virtually certain about the structure. The complication lies in the fact that partial cleavage into easier to handle oligopeptides is to be avoided if one wishes to find a small amount of by-product of very similar structure.

In the past, conventional techniques, such as specific or nonspecific partial hydrolysis, combined with amino acid analysis of each separated and purified cleavage product, identification of their N-terminal (Sanger) and C-terminal (Akabori) amino acids, and consecutive elimination of amino acids from the N-terminus (Edman) or C-terminus (carboxypeptidase), have been employed to this end with considerable success.

The tediousness and time consumption of the procedures involved made it worthwhile to search for a rather different approach. Mass spectrometric techniques have attracted considerable attention over the past few years, because of the particular specificity of this method for the arrangement of atoms (or groups) along a chain-like molecule, precisely the problem one faces in peptide structure determination.

The first mass spectrometric method proposed for the amino acid sequence determination was based on the exceedingly facile cleavage of the carbon–carbon bond in –N–(R)CH–CH_2–N– groups which form the repetitive unit in the product (I) of reduction of a peptide with lithium aluminum hydride.[5]



I

From the mass of the various ions a_1, a_2, a_3 ... a_n, obtained by single bond rupture along any one of the arrows in individual molecules, the sequence of R_1, R_2, R_3 ... R_n along the chain is deduced.

Another reason for the conversion of the polyamides into polyamine groups was the concomitant increase in volatility of the compound, a point of considerable

(1) Part IV of the series "Computer-Aided Interpretation of High Resolution Mass Spectra." For part III see K. Biemann, W. McMurray and P. V. Fennessey, *Tetrahedron Letters*, 3997 (1966).

(2) Imperial Chemical Industries, Pharmaceutical Division, Alderley Park, Cheshire, England.

(3) On leave of absence from the National Research Council of Canada, Atlantic Regional Laboratory, Halifax, Nova Scotia, Canada.

(4) R. B. Merrifield, *J. Am. Chem. Soc.*, 85, 2149 (1963).

(5) K. Biemann, F. Gapp, and J. Seibl, *ibid.*, 81, 2274 (1959).

importance in mass spectrometry. With the development of experimental techniques for the introduction of the samples directly into the ion source, thus lowering the vapor pressure requirements by about three orders of magnitude, it became possible to obtain mass spectra of peptide derivatives still containing all amide linkages. All later proposals for mass spectrometric sequence determination involve peptides modified only at the C- and/or N-terminus. While the presence of the carbonyl group considerably complicates the fragmentation of these molecules and thus their mass spectra, the greater effort required for their interpretation—in contrast to the exceedingly simple spectra of the polyamino alcohols of type I—may be offset by the elimination of a chemical reaction necessary to produce I from a peptide on a very small scale.

Previously proposed methods involving spectra of such peptide derivatives (N-trifluoroacetyl-,[6,7] N-acetyl,[8,9] and N-acyl ($C_{>10}$) peptide esters[10]) shall briefly be discussed before outlining the technique described in this paper.

The mass spectra of some N-trifluoroacetyl derivatives of methyl esters of a few dipeptides and a tripeptide had been mentioned briefly by Andersson,[11] and the sequence information contained in such spectra was first discussed by Stenhagen.[6] This feature was further elaborated on by Weygand, *et al.*,[7] who had originally made use of TFA derivatives in gas chromatography because of their relatively high volatility.[12]

Various groups, particularly in Germany[8] and Russia,[9] experimented with N-acetyl peptide esters and were able to show that the mass of many of the abundant fragments are due to cleavage of the CO–N bonds and thus carry sequence information. While all these techniques were tested on model substances the first example of the mass spectrometric determination of a partially unknown amino acid sequence is represented by the elucidation of the structure of fortuitine, a naturally occurring peptidolipid.[13]

$CH_3(CH_2)_n$CO-Val-MeLeu-Val-Val-MeLeu-AcThr-AcThr-Ala-Pro-OMe, n = 18 and 20

It appeared that the long acyl chain increased the volatility (as compared to a free peptide of the same sequence). It also aided the interpretation of the spectrum because the heavy N-terminal substituent lets all ions with intact N-terminus (and thus carrying sequence information) appear at high mass, preventing confusion with ions originating from the middle or C-terminal part of the molecule.

Based on these findings, the French–English group proposed the use of N-stearoyl peptide esters or related long-chain acyl peptides as a general mass spectrometric method for the determination of the amino acid sequences. Procedures for the N-acylation of free peptides with fatty acids have been reported.[10,14]

The interpretation of the spectra of these peptide derivatives follows the same principle as that outlined for the amino alcohols (I) using mainly fragments due to cleavage at the CO–N bond (along B in structure II below) with retention of the positive charge at the carbonyl group. Many examples of such spectra have been reported but they are peptides containing almost exclusively the small, nonpolar amino acids and sometimes phenylalanine units which do not give complex spectra by themselves and thus result also in simple peptide spectra. Still, in many of the published spectra some pronounced peaks remain unmentioned, while those which are in agreement with the, in most instances known, amino acid sequence are selectively discussed.

During our earlier work on peptide derivatives and free peptides we had become aware of these complications, which become much more pronounced in the presence of polar, polyfunctional amino acids. A recent report on the spectra of aspartyl and glutamyl peptides also points this out.[15] We therefore searched for a method which is independent, as far as possible, of the mass spectrometric idiosyncrasies of the individual amino acids present in the peptide, is primarily independent of the rather subjective assessment of the importance (*i.e.*, abundance) of an ion, and, most importantly, permits speedy, objective, and exhaustive interpretation of the mass spectrum in terms of amino acid sequence.

Before discussing the method, it is necessary to outline the objectives of this approach. Undoubtedly, by far the largest number of sequence determinations which have to be performed in the near future will be part of the determination of primary protein structure and any new technique will have to be generally suited for this task. Conventional techniques of protein degradation involve, in general, specific cleavage by enzymes into large polypeptides which are further degraded by other enzymes, specific chemical reagents, or partial acid hydrolysis. The result is a large number (from each one of the large peptides) of mixtures of rather small peptides, which now have to be carefully separated, analyzed qualitatively and quantitatively for amino acids, and their sequence determined by one of the chemical methods mentioned earlier. It is here that a tremendous effort is necessary to determine the amino acid sequences in all the small peptides which are then to be reassembled to the complete sequence.

We therefore felt that any new technique would have to fulfill the following requirements: (a) it should be fast; (b) it should be applicable to very small amounts of peptides (micrograms); (c) it should not require complete separation or extensive purification of the small peptides; (d) it should not require amino acid analysis on the small peptides; and (e) it should work with a high degree of confidence on all small di- to tetrapeptides (the major products of the final cleavage step) regardless of the nature of the amino acids present.

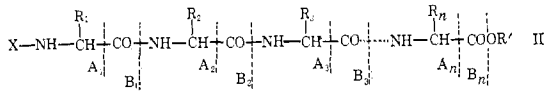
It should be noted that many of these requirements are specific for protein hydrolysates and do not necessarily hold for individually occurring oligopeptides, which, once isolated, can be subjected to amino acid

(6) E. Stenhagen, *Z. Anal. Chem.*, 181, 462 (1961).
(7) (a) F. Weygand, A. Prox, W. König, and H. H. Fessel, *Angew. Chem.*, 75, 724 (1963); (b) F. Weygand, A. Prox, H. H. Fessel, and K. K. Sun, *Z. Naturforsch.*, 20b, 1169 (1965).
(8) K. Heyns and H. F. Grützmacher, (a) *Tetrahedron Letters*, 1761, (1963); (b) *Ann. Chem.*, 669, 189 (1963).
(9) N. S. Wulfson, *et al.*, *Tetrahedron Letters*, 2805 (1965).
(10) E. Bricas, *et al.*, *Biochemistry*, 4, 2254 (1965).
(11) C.-O. Andersson, *Acta Chem. Scand.*, 12, 1353 (1958).
(12) (a) F. Weygand, R. Geiger, and H. Swowdeck, *Angew. Chem.*, 68, 307 (1965); (b) F. Weygand, G. Klipping, and D. Polm, *Chem. Ber.*, 93, 2619 (1960).
(13) M. Barber, P. Jolles, E. Vilkas, and E. Lederer, *Biochem. Biophys. Res. Commun.*, 18, 469 (1965).
(14) A. A. Kiryushkin, *et al.*, *Tetrahedron Letters*, 33 (1966).
(15) N. S. Wulfson, *et al.*, *ibid.*, 39 (1966).

analysis with not too much effort. This analysis then reveals whether there is a chance of obtaining an interpretable mass spectrum and, if so, considerable effort is warranted to determine and to interpret it without prior degradation into smaller peptides followed by reconstitution of the original sequence.

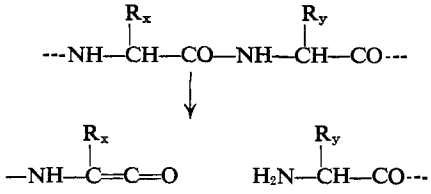
For all these reasons we found it most important to devote our attention to the above stated requirements a–e and developed a technique for the use of a computer in the determination of the amino acid sequence from the high resolution mass spectrum of a small peptide. The main emphasis was placed on general applicability for short peptides regardless of the amino acids present, rather than on longer ones of particularly suitable (aliphatic) amino acids.

While the basic principles of mass spectrometry as well as the results obtained in various laboratories indicate that a peptide chain such as II will, on electron impact, cleave either at the C–CO or CO–N bond one should not assume that the resulting ions (type A or B) will always be the most abundant ones. A superficial consideration of most of the published spectra makes it appear as if they (particularly B) gave rise to the most pronounced peaks, but this is caused by the fact that the examples used for demonstration are mostly peptides derived from small aliphatic amino acids. Functional groups in the side chains introduce other fragmentation modes and the abundance of the resulting ions often overshadows that of type A or B (in II) which complicates any interpretation relying on characteristic nominal differences.[15]



For this reason it was felt that abundance should not be the primary criterion for a generally applicable method of interpretation. In its place, the elemental composition of the ions appeared to be much more suitable, because the heteroatom content increases by one nitrogen and one oxygen (plus all heteroatoms present in the "side chain" of the amino acid) from one "amine ion" to the next (*i.e.*, from A_1 to A_2 to A_3, etc.), and the same holds for the "aminoacyl ions" (B_1, B_2, B_3, etc.). With the advent of techniques that permit the speedy determination of the elemental composition of all the ions formed from an organic compound in the mass spectrometer,[16] this basic information could be obtained and utilized for deducing the amino acid sequence in an oligopeptide. Since these primary data are already produced by a computer[16] it was obvious that this "interpretation" should be done with the aid of the computer, which is particularly suited for this purpose, since it involves comparison of a set of expected data with those experimentally available.[17,18]

Before outlining the logic of the program, a few technical points need to be made. The free peptides are for a number of reasons rather unsuitable substances for mass spectrometry. Their zwitterionic character decreases their volatility (an important factor in mass spectrometry) and thus severely limits the complexity of peptides of which spectra can be obtained. Furthermore, dipeptides have the tendency to cyclize to diketopiperazines on heating (during the vaporization). The spectra[19] of these do, of course, no longer permit any conclusions concerning the original sequence. Finally, some of the fragmentation processes of peptides and their derivatives involve cleavage of a peptide bond into a free amino group and a ketene, thus liberating a smaller peptide having a new N-terminal amino acid.



To avoid confusion of this electron impact generated N-terminal amino acid with that of the original peptide giving rise to the spectrum, its N-terminus has to be marked by a suitable substituent. This is best combined with the destruction of the zwitterionic character of the peptide and it is for this reason that N-acyl derivatives are the most suitable peptide derivatives for mass spectrometry. To be uniquely recognized, the marking substituent (X in structure II) has to be of unique elemental composition. This leads to type A and B ions of sufficiently unique composition to be easily distinguished from ions not containing the N-terminus and thus not providing N-terminal sequence information. The use of a long-chain fatty acid residue for X, which marks the N-terminus containing fragments by their high mass,[10] is thus no longer necessary. Most suitable are acyl groups containing halogen or aromatic rings or both, but preferably no nitrogen.

Acylation of the amino group leaves the free carboxyl group at the C-terminus which may or may not be esterified. Conversion to a methyl ester further increases the volatility of the peptide derivative, but this difference is of significance only in the cases which are on the borderline of sufficient volatility. Esterification with an alcohol of more unique elemental composition may be useful for similarly following the sequence from the C-terminus, and *p*-fluorobenzyl esters (II, R′ = *p*-C_6H_4F) might be advantageous.

The size of the peptides that can be handled at present is not limited by the mass range covered by the spectrometer[20] but by the volatility and thermal stability of the sample. Excellent mass spectra have been obtained from octa-[10] and nonapeptides[13] exclusively containing only small aliphatic amino acids lacking functional groups (except O-acetylthreonine[13] which cleanly eliminates the functionality as acetic acid). The presence of polyfunctional amino acids, *e.g.*, asparagine, glutamine, histidine, etc., decreases the volatility and increases the thermal lability considerably, presently limiting the size of peptides containing one or more of these amino acids to di- to tetrapeptides. Although the

(16) 12th Annual Symposium of Mass Spectrometry, Montreal, June 1964: (a) P. Bommer, W. J. McMurray, and K. Biemann, p 428; (b) D. Desiderio and K. Biemann, p 433; (c) D. Desiderio, Ph.D. Thesis, Massachusetts Institute of Technology, 1965; (d) K. Biemann, P. V. Fennessey, and J. M. Hayes, Proceedings of the Society of Photo-Optical Instrumentation Engineers, 1966, in press.

(17) For a preliminary report see K. Biemann, C. Cone, and B. R. Webster, *J. Am. Chem. Soc.*, 88, 2597 (1966).

(18) A similar approach was suggested simultaneously; see M. Senn and F. V. McLafferty, *Biochem. Biophys. Res. Commun.*, 23, 381 (1966).

(19) H. J. Svec and G. A. Junk, *J. Am. Chem. Soc.*, 86, 2278 (1964).

(20) Commercially available instruments have presently a mass range of a few thousand mass units (see, for example, H. Fales, *Anal. Chem.*, 38, 1058 (1966)).

sample-handling techniques are continuously being improved, which lets one expect that the range of applicability will steadily increase, we have made an effort to design a method that gives reliable results on small peptides with any amino acid (at least as a suitable derivative, such as S-benzylcysteine, for example,[21]) particularly since the published mass spectra of larger peptides[7b,10,13,14] indicate that the method of interpretation outlined in this paper would also lead to the correct assignment of the amino acid sequence with these substances.

Discussion of the Computer Program for Sequence Analysis

The approach utilized here is based on the above-mentioned fact that any compound of type II will, on electron impact, produce to some extent fragments of type A and/or B (see structure II). These "amine" (A) and "aminoacyl" (B) fragments reflect the amino acid sequence, because the difference in elemental composition between consecutive fragments of the same type indicate the elemental composition of the "side chain" of the next amino acid (the backbone unit, C_2H_2NO, being always the same). With the exception of the isomeric pair leucine and isoleucine, the elemental composition of the side chain identifies any one of the common amino acids. Cyclic amino acids are equally characterized by the difference between $CHNH_2$ and the elemental composition of the ring; thus the side-chain equivalent for proline is C_3H_5.

Theoretically, either series $A_1 \ldots A_n$ or $B_1 \ldots B_n$ alone would suffice as a test, but searching for both represents a double check, and it may be that one or the other cleavage may give rise to an ion of very low abundance which escapes detection. It should be noted that long aliphatic acyl groups (X in II) lead to preferential formation of aminoacyl ions (B), while peptides with aromatic acyl groups (*e.g.*, phthaloyl as an extreme) exhibit spectra in which the amine fragments (A) dominate. Furthermore, some amino acids have a lower tendency to give amine fragments (A) than to give aminoacyl ions (B) (*e.g.*, glycine) and *vice versa* (*e.g.*, proline). The spectrum, *i.e.*, the list of accurate masses of all ions formed, regardless of their abundance, is searched for the presence of a set of ions that represents consecutive amine and/or aminoacyl ions belonging to a unique sequence and that is terminated with the corresponding molecular ion. This requires searching for two sets of up to more than 20 possibilities (one for each amino acid which could be present) for each consecutive peptide bond and is therefore a task which can be performed economically and reliably only by a computer.

The input data consist of the following: the accurate mass of the N-terminal marking group plus that of NHCH (*e.g.*, 163.0633 for $C_6H_5CH_2OCONHCH$ or 125.0088 for $CF_3CONHCH$) and that of the C-terminus (*e.g.*, 44.9977 for COOH or 59.0133 for $COOCH_3$) because these are known from the prior treatment of the sample; the mass of the "backbone unit" (CONHCH, 56.0136) which can be modified to allow the use of polyamino alcohols (I) by changing it to the mass of CH_2NHCH; the mass of CO (27.9949), the difference between ions of type A and N; the masses of the "side chains" of all amino acids possibly present; and finally compound data, *i.e.*, the masses and abundances of all ions in the spectrum, determined in the usual manner.[16]

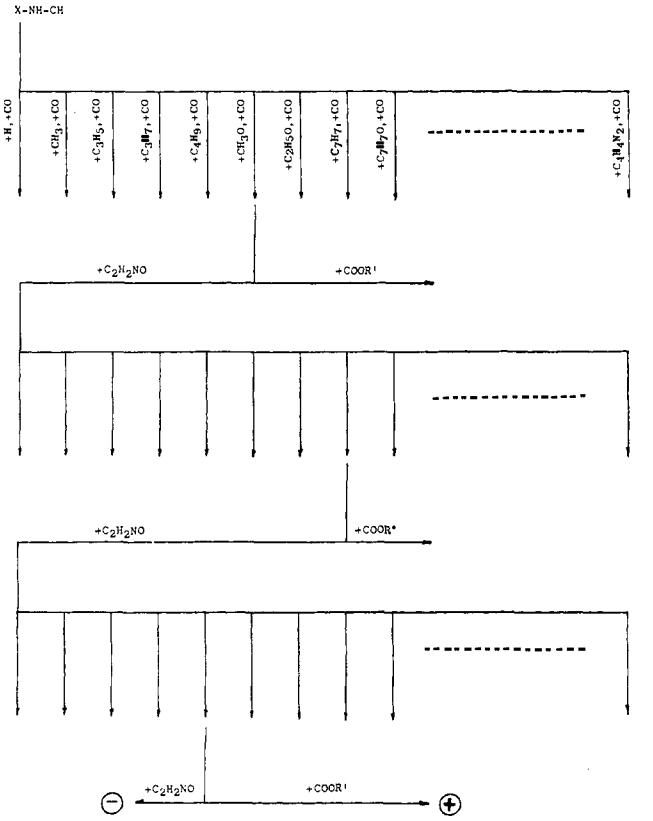


Figure 1. Schematic representation of the basic outline of the program logic of the sequence search.

It should be noted that it suffices to use the exact masses rather than elemental compositions which are directly related to each other. Fortunately, the otherwise sometimes confusing similarity of the masses of different combinations of elements (*e.g.*, N_3 *vs.* C_2H_2O, $\Delta m = 0.0013$ mass unit, or C_3N *vs.* H_2O_3, $\Delta m = 0.0027$ mass unit) causes no problem because of the gradual increase of both nitrogen and oxygen content along the structure of the peptide derivative.

The scheme followed for the search is very schematically illustrated in Figure 1, using the N-carbobenzoxy peptide ester Cbz-Ser-Phe-Leu-OMe as an example (*i.e.*, $X = C_6H_5CH_2OCO$, $R' = CH_3$).

To the initial mass, 163.0633 (X–NH–CH) is added the mass of the first amino acid side chain in the list (the mass of H, 1.0078, for glycine), and the resulting sum (the amine ion for N-terminal glycine) is compared with the masses of the experimentally found compound ions. If one of these agrees with that sum within a certain limit (*e.g.*, ± 0.003 mass unit) glycine is a possibility for the N-terminal amino acid. This is further tested by adding the mass of CO (resulting in the mass of the corresponding aminoacyl fragment) and searching the data for a fit with this sum. If either one of these are found, the search continues for the next amino acid by adding the mass of $C_2H_2NO_2$ and repeating the addition of the side-chain masses and CO (the second and third row of arrows in Figure 1 represent the same amino acids indicated in the first

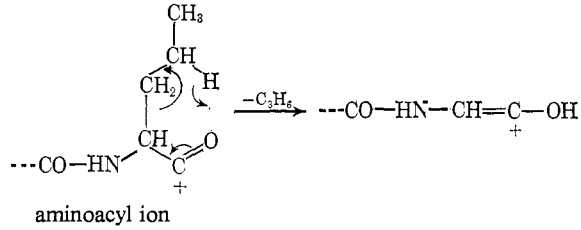
(21) The polarity and basicity of arginine still prevents the use of unmodified arginyl peptides.

row). If neither the amine nor the aminoacyl fragment corresponding to N-terminal glycine was found in the data, the next amino acid (alanine) is tested in the first segment, and so forth. This is done under any circumstances, *i.e.*, even if N-terminal glycine would lead to a complete sequence, to assure that all possibilities are considered (a fit for N-terminal glycine could also arise from N-terminal valine by elimination of C_3H_6, as will be discussed below).

At each stage, the mass of the C-terminal group (*e.g.*, COOH or $COOCH_3$) is added and tested for its presence as this would indicate completion of a possible sequence. Figure 1 represents an example in which a fit for either amine or aminoacyl fragment, or both, was found for serine in the first segment, for phenylalanine in the second, and for leucine in the third but none in the fourth, but for the addition of $COOCH_3$ instead. The sequence Ser-Phe-Leu follows from these data. Since the molecular ion might in some cases be too weak to be recorded, an $M - H_2O$ ion from a free acid or $M - CH_3OH$ from a methyl ester (which is often a more prominent ion) is also considered as indirect evidence for the molecular weight. Not shown in the diagram are some other ions that are present but do not lead to a complete sequence.

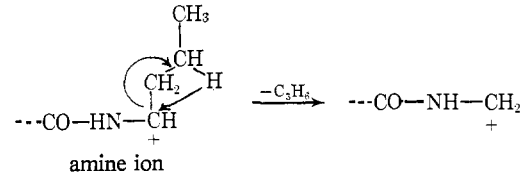
During the process depicted in Figure 1 the relative abundances of all the ions representing a sequence permitted by the data are simultaneously added and stored for later selection of the most probable result (see below). After all combinations of A and B ions have been tested, the partial and completed sequences are printed along with the sum of the intensities represented by the ions indicating this particular sequence. It is logical to assume that that one will be the most probable one which is based on the most abundant amine and aminoacyl ions. The reasons for the appearance of more than one answer will be discussed later. Thus, the computer selects those preliminary results that are based on the most abundant ions (this value can be varied at the discretion of the investigator; *e.g.*, >80% of the highest summed intensities representing a sequence) and gives them a more thorough treatment.

This involves searching for additional ions that can be expected to arise from each one of the amine and aminoacyl ions of the sequence being tested. For example, a fragment containing valine may eliminate the side chain of valine as C_3H_6



aminoacyl ion

or



amine ion

or a fragment containing threonine may eliminate H_2O. This part of the process represents utilization of additional data available for confirmatory purposes, and involves the use of our knowledge of the behavior of ions in the spectrometer. Since this is, however, governed by the nature of the amino acids present, it is deliberately not made part of sequence selection, but represents a separate step, sequence confirmation. Obviously the correct sequence will explain the largest number (in kind and abundance of the ions in the spectrum) which is used as a further criterion in the selection of the final answer. A similar treatment is given to the molecular ion by testing for possible fragment ions thereof (other than those of type A or B), thus providing a further confirmation of the correctly assigned molecular ion (whether or not it appears in the spectrum). These few detailed interpretations are then printed in order of intensities of the summed amine and aminoacyl ions plus the molecular ion, to give the investigator an opportunity to check the final result rather than blindly relying on the computer output. For example, a sample may turn out to be a mixture of two peptides and this can often be seen by inspection of these data, while it would be quite involved to write all the necessary checks into the program.

The program permits, of course, the processing of more than one spectrum. About 0.5 to 3 min of main frame time on an IBM 7094 computer is required per spectrum, depending on its complexity and number of individual amino acids to be considered.

It is of significance to point out one other important difference between the approach discussed here and the previously suggested sequence determinations, based on manually interpreted conventional ("low resolution") mass spectra.[7-10,13-15] While they also utilize the ions of type A and B (particularly the latter) the search is conducted in *reverse*, *i.e.*, starting from the molecular ion. Anyone experienced in the mass spectrometry of compounds of unknown structure is aware of the difficulties inherent in the unambiguous identification of a molecular ion, particularly of a compound that has a tendency to produce them in very low abundance (or not at all) and is difficult to purify.[22] That this method has worked in all published cases is mainly due to the fact that they involved well-purified compounds of known structure or known amino acid composition (with the exception of fortuitine[13] where the N-methylleucines had escaped detection). The same criticism applies also to a very recently suggested computerized technique[23] which also involves search from the molecular ion down and utilizes the chemically determined amino acid composition as input data.

The N-terminal amine and/or aminoacyl ions are, however, always quite prominent and this starting point will thus never be missing from the data. Its unambiguous identification is, of course, facilitated by the approach suggested here, namely characterizing it by a unique elemental composition, because abun-

(22) It has even been suggested to use the chemically determined qualitative and quantitative amino acid composition of the peptide under investigation as an aid to identifying the molecular ion, or, if this is not found, to use the molecular weight calculated from this analysis.[7b] As pointed out earlier in this paper, we do not believe that there is much point in suggesting a new technique for the sequence analysis of small peptides to be used for the determination of protein structure if it requires not only complete purification but also qualitative and quantitative amino acid analysis.

(23) M. Barber, P. Powers, and W. A. Wolstenholme, 14th Annual Symposium on Mass Spectrometry, Dallas, Texas, May 1966.

dance alone is a dubious criterion at the lower mass range which may abound with other intense peaks.

At this point it should also be noted that the proposal by Senn and McLafferty[18] which is based on the same principle as ours[17] implies that consideration of the accurate mass (*i.e.*, elemental composition) of the A and B ions alone suffices to uniquely determine the amino acid sequence of a peptide derivative. This is possible only for peptides whose mass spectrum exhibits very intense amine and aminoacyl ions, by setting the threshold for recording the peaks very high. Only in this case can one avoid several alternative answers from the computer search. Such answers arise mainly by the electron impact induced elimination of certain groups which formally correspond to the difference in elemental composition between two amino acids from the molecular ion or a sequence-characteristic fragment ion, or by the occasional appearance of ions whose elemental composition coincides with a possible amine or aminoacyl ion. Examples are represented by the elimination of C_3H_6 from ions containing valine (see above), thus simulating glycine, or by the identity of the contribution which the presence of asparagine or two consecutive glycines (*i.e.*, $C_4H_6N_2O_2$) make to a peptide sequence. These cases always give rise to a mass spectrum which is consistent with more than one sequence, if mass alone is the criterion. This became apparent (already) in some of our earliest experiments and it was realized that the correct solution can be selected only if the abundance of the ions of type A and B is also taken into account. This is based on the fact that ions arising by the loss of groups from other ions generally are of lower abundance than those from which they are derived. Secondly, if there is a Gly-Gly sequence it will contain two A and B ions (those of the first Gly) more than the sequence in which Gly-Gly is thought to represent asparagine, and the sum of the intensities of the A ions, B ions, and the molecular ion is thus lower.

This is best outlined by an actual example taken from our studies in the amino acid sequence of isariin (III), which will be discussed in more detail later. Based on the high resolution spectrum of a derivative, methyl isariate (V), the computer found a considerable number of sequences consistent with the spectrum. Some of these are listed in Table I. It will be noted that they differ in the summed intensities of A, B, and M ions and that the sequence which represents the largest sum is the correct one (see discussion of the data from isariic acid). Close inspection of the differences among these sequences reveals that all the hypothetical ones are due to the elimination of C_3H_6 from one of four ions, namely from the molecular ion, from X-Gly-Val-Leu-Ala-Val(CO), simulating X-Gly-Val-Leu-Ala-Gly-CO, from X-Gly-Val-Leu (CO), simulating X-Gly-Val-Ala (CO), and from X-Gly-Val (CO), simulating X-Gly-Gly (CO). It is apparent that in this case the elimination of the side chain involves only aminoacyl ions, a fact which also may be used to differentiate these artificial sequences from the real one. The last one, simulating the presence of a Gly-Gly sequence, in turn gives rise to ions whose elemental compositions are the same for the presence of asparagine in this sequential position.

Table I. A Few of the Computer-Suggested Sequences for Methyl Isariate[a]

	Sum of intensities
X—Gly $\frac{16^b}{497^c}$ Val $\frac{54}{989}$ Leu $\frac{194}{492}$ Ala $\frac{1}{32}$ Val $\frac{-^d}{2}$ OCH_3 (2)[e]	2279
X—Gly——Val—Leu—Ala—*Gly*[f] $\frac{-}{1}$ OCH_3 (1)	2277
X—Gly——Val—*Ala*—*Leu* $\frac{-}{10}$ Val—OCH_3	1603
X—Gly——Val—*Ala*—*Leu*—*Gly*—OCH_3	1601
X—Gly——*Gly* $\frac{-}{140}$ *Leu*—*Leu*—Val—OCH_3	700
X—Gly——*Gly*—*Leu*—*Leu*—*Gly*—OCH_3	698
X————*Apg*[g]—*Leu*—*Leu*—Val—OCH_3	187
X————*Apg*—*Leu*—*Leu*—*Gly*—OCH_3	185

[a] X = 3-hydroxydodecanoyl. [b] Intensity of amine fragment. When given for a sequence, it is not repeated for another sequence if the same. [c] Intensity of amino acyl fragment. [d] Ion not detected. [e] Intensity of M^+ or $(M - C_3H_6)^+$, whichever is applicable. [f] All "simulated" (see text) amino acids are italicized. [g] Apg = asparaginyl.

There are, in principle, quite a number of such coincidences. The side chain of leucine can be lost as C_4H_8 in a process analogous to that of valine and thus also simulate the presence of glycine; the same holds for the side-chain elimination of serine, threonine, lysine, and, in fact, all amino acids containing a hydrogen in the γ position with respect to the carbonyl carbon (except proline and related cyclic amino acids). Leucine may, however, also eliminate C_3H_6[24] and thus simulate alanine. In analogy with Gly-Gly, Ala-Gly or Gly-Ala will simulate glutamine.

The use of a high threshold during the recording of the data would give great difficulties with peptides containing amino acids that direct fragmentation away from the peptide bonds leading to A and B ions of very low relative abundance.

Examples. It has already been pointed out that all published spectra of peptide derivatives are quoted to contain either the aminoacyl and/or amine peak of each segment, as well as the molecular ion; their interpretation by the computer as outlined in this paper would thus produce the correct answer for the sequence.

For reasons outlined earlier, we have thus paid particular attention to peptides of which one could expect the sequence characteristic fragments to be of low abundance or which contain amino acids other than those covered in the literature. The choice was somewhat limited by the availability of the samples. The compounds used (a few micrograms suffice for a spectrum) ranged from di- to pentapeptides of the amino acids listed in Table II.

The N-terminal substituents used in our work are listed in Table III. Good results were obtained with all of them and since it is, in principle, independent of the N-terminal substituent (except its accurate mass) it eliminates the need of a thorough and detailed investigation of the mass spectrometric behavior of each particular type of peptide derivative, as was necessary for the conventional interpretation of trifluoroacetyl derivatives.[7b]

Of the substituents listed in Table III we found the carbobenzoxy and trifluoroacetyl peptides most useful

(24) K. Biemann, J. Seibl, and F. Gapp, *J. Am. Chem. Soc.*, **83**, 3795 (1961).

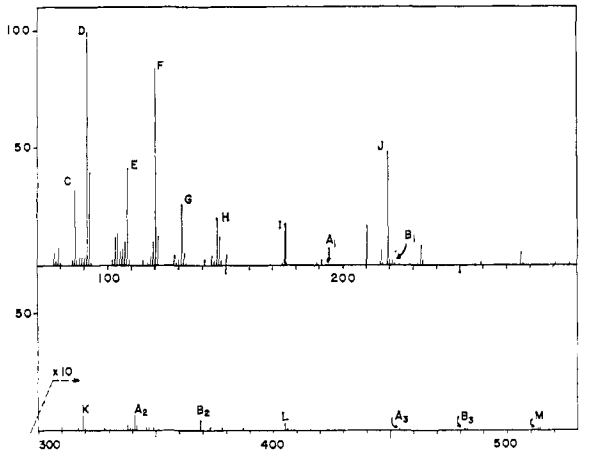


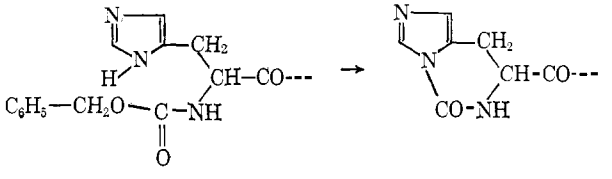
Figure 2. Conventional representation of the mass spectrum of Cbz-Ser-Phe-Leu-OMe. For explanation of the letters see Table IV and formula II.

from the point of view of a compromise involving volatility, abundance of the higher amine and aminoacyl ions as well as molecular ion, and uniqueness of elemental composition. N-Acetyl peptides have the disadvantage of an N-terminal substituent contributing

Table II. Amino Acids Present in Peptides Investigated

Amino acid	Elemental composition and mass of side chain	
Glycine	H	1.0078
Alanine	CH_3	15.0235
Serine	CH_3O	31.0184
Proline	C_3H_6	41.0391
Valine	C_3H_7	43.0548
Threonine	C_2H_5O	45.0340
Leucine	C_4H_9	57.0704
Asparagine	C_2H_3NO	59.0133
α-Aminoadipic acid	C_4H_5O	69.0340
Lysine	$C_4H_{10}N$	72.0813
Glutamic acid	$C_3H_5O_2$	73.0290
Methionine	C_3H_7S	75.0269
Histidine	$C_4H_5N_2$	81.0453[a]
Phenylalanine	C_7H_7	91.0548
Tyrosine	C_7H_7O	107.0497
S-Benzylcysteine	C_8H_9S	137.0425

[a] In the case of histidine, the hypothetical mass −27.0122 ($C_4H_5N_2$ − $C_6H_5CH_2OH$) should also be used, as N-carbobenzoxy histidyl peptides have the tendency to eliminate benzyl alcohol from the N-terminal position.



little to the elemental composition and thus involves the risk of finding too many other ions the composition of which corresponds to a possible but actually absent ion of type A or B. Some 2,4-dinitrophenyl derivatives were also tried but seem to be of relatively low volatility and show low abundance of the ions of higher mass.

The fact that the technique is not dependent on a particular substituent makes it possible to confirm a result obtained by repeating it with another one. This

Table III. N-Terminal Substituents of Peptide Derivatives Investigated

Substituents	Elemental composition of X-NH-CH	Mass
Acetyl	C_3H_5NO	71.0371
Chlorodifluoroacetyl	C_3H_2NOClF	140.9793
Trifluoroacetyl	$C_3H_2NOF_3$	125.0088
Benzoyl	C_8H_7NO	133.0530
Carbobenzoxy	$C_9H_9NO_2$	163.0633
Phthaloyl[a]	$C_9H_5NO_2$	159.0320
Chlorophthaloyl	$C_9H_4NO_2Cl$	192.9931
Tosyl	$C_8H_9NSO_2$	183.0336
3-Hydroxydodecanoyl	$C_{13}H_{25}NO_2$	227.1885
Dodecenoyl[b]	$C_{13}H_{23}NO$	209.1780

[a] X = >N-CH. [b] By electron impact induced dehydrogenation of 3-hydroxydodecanoyl.

may prove particularly desirable if one works with peptide mixtures.

Our preliminary communication contained already one detailed example (N-carbobenzoxyvalylasparaginylleucine methyl ester). It seems to be worthwhile to show here another one which outlines one of the points discussed earlier in this paper. The spectrum of N-carbobenzoxy-L-seryl-L-phenylalanyl-L-leucine methyl ester will be discussed for this purpose. This compound contains serine which has the tendency to eliminate H_2O or CH_2O upon electron impact, as well as phenylalanine which facilitates elimination of a carboxamido group and formation of a cinnamoyl moiety. The latter gives rise to a number of very intense ions, and so do a number of other fragments of phenylalanine. Thus, a spectrum results in which the original amine and aminoacyl ions (A and B) carry very little of the ion current, *i.e.*, are of minor abundance compared with other peaks. This is best illustrated by the conventional representation of its mass spectrum (in "low resolution" form) in Figure 2. The sequence-determining peaks A_1, A_3, B_1, and B_3 are barely visible if at all (since in the mass range below 300 the drawing shows only peaks $\geq 1\%$ of the highest peak). The more abundant peaks are due to fragments which do not reveal the sequence (except J, which is a C-terminal fragment). Most of them are due to the presence of phenylalanine or the carbobenzoxy group (Table IV).

Table IV. Origin of Prominent Peaks in the Mass Spectrum of Cbz-Ser-Phe-Leu-OMe (see Figure 2)

Peak	Origin
C	$(C_4H_9-CH-NH_2)^+$ (from Leu)
D	$C_7H_7^+$ (from Phe or Cbz)
E	$(C_6H_5CH_2OH)^+$ (from Cbz)
F	$(C_6H_5CH_2CHNH_2)^+$ (from Phe)
G	$(C_6H_5-CH=CH-CO)^+$ (from Phe)
H	$(C_6H_5-CH-C(NH_2)=C=O)^+$ (from Phe)
I	$(A_1-H_2O)^+$
J	$(C_6H_5-CH=CH-CO-NH-CH_2-CO_2Me)^+$
K	$(M - A_1)^+$
L	$(M - C_6H_5CH_2OH)^+$

The computer, disregarding all ions incompatible with the elemental compositions of possible amine or aminoacyl fragments in its sequence-selecting search, correctly identifies the ions A_1–A_3 and B_1–B_3 as well as the molecular ion. The result is shown in Figure 3,

```
SEQUENCE FOUND FOR SAMPLE  574-14-

   CARBOBENZOXY-SER -PHE -LEU -OME -

BASED ON FOLLOWING DATA

AMINE FRAGMENTS
INTENSITY          1    5    1    1
ERROR             -.4  -.9  2.9 -1.9
AMINO ACYL FRAGMENTS
INTENSITY         10    3    1
ERROR            -1.5   .8   .1
SUM OF INTENSITIES IS-     22     TOTAL INTENSITY =  1347
AVERAGE ERROR IS-  1.207

   INT       FRAGMENT LOST FROM R-SER-
     1       H2O

   INT       FRAGMENT LOST FROM MOLECULAR ION
     1       SER
     1       SER-H
     1       PHE
     1       LEU-H

SUM OF INTENSITIES WITH FRAGMENTS LOST      27
```

Figure 3. Reproduction of the result of the sequence analysis of a synthetic peptide derivative.

indicating that the correct result was found although these ions represent less than 2% (22 units out of 1347) of all ions of mass higher than that of $C_6H_5CH_2OCO$-NHCH.

A number of sequence-confirming ions (fragments lost from amine or aminoacyl ions as well as from the molecular ion) are also identified. Some of these do not appear in Figure 2 because they are of less than 0.1% relative abundance (the intensities printed in Figure 3 are rounded up).

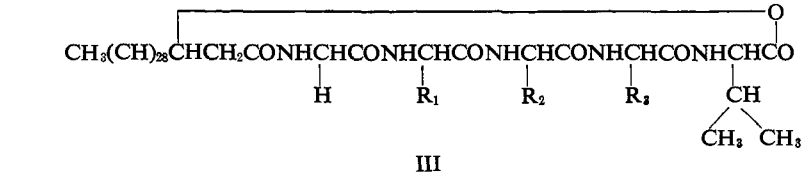
The data are listed in terms of intensity and deviation (in millimass units) of calculated from determined mass (error) for the amine fragments first and for the aminoacyl fragments below those. It should be noted that in Figure 3 under "Fragment lost . . ," the notations Ser, Phe, Leu refer to the side chains of these amino acids (*i.e.*, CH_2OH, $C_6H_5CH_2$, C_4H_9); in Figure 4 empirical formulas were used to avoid any ambiguity.

```
SEQUENCE FOUND FOR SAMPLE 584-26

   DODECENOYL-GLY -VAL -LEU -ALA -VAL - OH -

BASED ON FOLLOWING DATA

AMINE FRAGMENTS
INTENSITY         82   11   98   10    1    1
ERROR              .1 -1.5   .1   .5  2.2   .2
AMINO ACYL FRAGMENTS
INTENSITY        422  635  248   23    1
ERROR            1.2  1.9   .9  2.0  -.9
SUM OF INTENSITIES IS-  1532     TOTAL INTENSITY =  4414
AVERAGE ERROR IS-  1.056

   INT       FRAGMENT LOST FROM R-GLY-VAL-
     24      C3H6

   INT       FRAGMENT LOST FROM R-GLY-VAL-LEU-
     17      C3H6

   INT       FRAGMENT LOST FROM R-GLY-VAL-CO -
     55      C3H6

   INT       FRAGMENT LOST FROM R-GLY-VAL-LEU-CO -
     11      C3H6

   INT       FRAGMENT LOST FROM MOLECULAR ION
      2      CO2
      1      C3H6
      1      C4H8

SUM OF INTENSITIES WITH FRAGMENTS LOST    1644
```

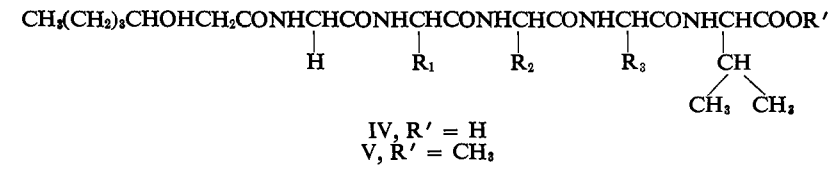
Figure 4. Reproduction of the result of the sequence analysis of isariic acid (IV) using the dehydrated side chain.

mination of the amino acid sequence in isariin by computer interpretation of the high resolution mass spectra of isariic acid (IV), methyl isariate (V), and the polyamino alcohol (VI) obtained by $LiAlH_4$ reduction of isariin was undertaken.

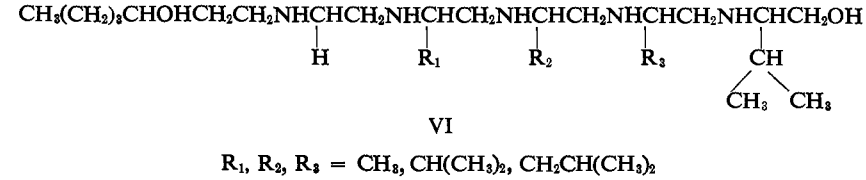
Figure 4 shows the computer output obtained from the amino acid sequence search of the high resolution mass spectrum of isariic acid, based on the dodecenoyl group as N-terminal substituent (produced by electron impact induced dehydration). A similar search was also carried out using 3-hydroxydodecanoyl as the side chain and gave identical results. The combined



III



IV, R′ = H
V, R′ = CH_3



VI

R_1, R_2, R_3 = CH_3, $CH(CH_3)_2$, $CH_2CH(CH_3)_2$

Rather than listing more of the results obtained with model peptides, the use of the technique for the determination of the structure of a naturally occurring compound, isariin, shall be briefly discussed. It is a metabolite of the fungus *Isaria cretacea* van Beyma and was isolated by Vining and Taber who suggested it to be a depsipeptide having structure III.[25] Determination of the amino acid sequence in isariin by computer interpretation of the high resolution mass spectra of isariic acid (IV), methyl isariate (V), and the polyamino alcohol (VI) obtained by $LiAlH_4$ reduction of isariin was undertaken.

Figure 4 shows the computer output obtained from the amino acid sequence search of the high resolution mass spectrum of isariic acid, based on the dodecenoyl group as N-terminal substituent (produced by electron impact induced dehydration). A similar search was also carried out using 3-hydroxydodecanoyl as the side chain and gave identical results. The combined sum of intensities of both searches, including fragments lost, account for nearly the total intensity of the spectrum from *m/e* 209 to the molecular ion. The computer search (with CH_2NHCH as the backbone unit, as mentioned above) of the high resolution mass

(25) L. C. Vining and W. A. Taber, *Can. J. Chem.*, 40, 1579 (1962).

spectrum of the polyamino alcohol VI fully substantiated the sequence shown in Figure 4. Thus, isariin, isariic acid, and methyl isariate are represented by III, IV, and V, respectively, with $R_1 = CH(CH_3)_2$, $R_2 = CH_2CH(CH_3)_2$, and $R_3 = CH_3$.

While this work was done the investigation of the structure of isariin was also in progress elsewhere. It involved a more conventional mass spectrometric and degradative approach, and the results which are in full agreement with ours have just been published.[26]

The computer search of the high resolution mass spectrum of methyl isariate showed that the sequence accounting for the largest sum of intensities was the same as that shown in Figure 4 for isariic acid. However, an unusually large number of fits was found. Some of these were due to the secondary fragmentation processes shown in and discussed in connection with Table I. Others could only be accounted for by assuming the presence of two compounds, namely methyl isariate (V) and another component amounting to 5–10% of the total. The latter must be a higher homolog since it gave consistently sequences like VII.

X-Gly-Val-Leu-Ala-Val-OMe
V

X-Ala-Val-Leu-Ala-Val-OMe
VII
$X = C_{12}H_{33}O_2$

Indeed the published line spectrum of methyl isariate[26] also contains all the peaks suggestive of the presence of VII except for the absence of its molecular ion (*m/e* 683) which seems to have been omitted since it is present in our low-resolution scans as well as in the high-resolution spectrum recorded on a photographic plate. Sequence VII was not mentioned in the previous work[26] probably because it escaped detection. The computer search, however, found it because of its thoroughness and impartiality in seeking all possible fits rather than terminating after finding the first one that is in agreement with all other (in this case incomplete) chemical data.

A more detailed consideration of the data finally revealed the actual cause for this surprising result. Whenever X–Ala was found, the data always showed the presence of $C_{13}H_{25}O_2$ and $C_{13}H_{23}O$, an acyl ion and an acyl ion less water, both with 13 carbon atoms rather than 12 as in isariin. Isariic acid (IV), which did not give rise to results with N-terminal alanine, did, however, not produce ions having elemental compositions $C_{13}H_{25}O_2$ and $C_{13}H_{23}O$. Sequence VII should thus be rewritten as sequence V but with $X = C_{13}H_{25}O_2$.

Treatment of isariic acid with diazomethane must have resulted in methylation of the hydroxyl group in the side chain to a small extent. This suggestion is supported by the presence of ions having elemental compositions $C_{10}H_{21}O[CH_3(CH_2)_8CHOH]^+$ and $C_{11}H_{23}O[CH_3(CH_2)_8CHOCH_3]^+$, the latter being stronger, in the mass spectrum of methyl isariate. Of these two ions, only $C_{10}H_{21}O$ is present in the spectrum of isaric acid (IV). These results eliminate the possibility that isariin itself may be a mixture of two compounds differing by one carbon atom in the hydroxyacyl side chain.

While isariin is a naturally occurring peptidolipid and might thus not be considered a good analog of a peptide produced by partial hydrolysis of a protein, it does illustrate the flexibility of the method with respect to the N-terminal marking group. More importantly, the spectrum of methyl isariate demonstrated that the method not only yields the correct amino acid sequence in the presence of a related contaminant but even permits elucidation of the latter. We feel that this relative insensitivity of the technique to incompletely separated mixtures of peptides is one of the major and perhaps decisive advantages of the computer-aided mass spectrometric approach to amino acid sequence determination. Consideration of the logic of the program outlined earlier in this paper suggests that the determination of the amino acid sequence of a few peptides in a mixture will be easiest if their amino acid composition differs widely, and most difficult if they are related to each other as are the combinations listed in Table I. Conversely reasonable amounts of nonpeptidic contaminants will hardly ever interfere, because the chance that the elemental composition of ions produced from them will coincide with a series of peptide fragments is remote. Direct eluates of paper or thin layer chromatograms have been used successfully.

Experimental Section

Preparation of Methyl Isariate (V). To 0.5 mg of isariic acid (IV) was added 1 ml of a freshly distilled ethereal solution of diazomethane. Reaction was instantaneous and the product precipitated. The solvent was vaporated to dryness and the residue was dissolved in methanol and transferred to sample capillaries for determination of the mass spectrum.

Preparation of the Polyamino Alcohol (VI) from Isariin (III).[5] A solution of 0.6 mg of isariin in freshly distilled tetrahydrofuran was added to a solution of 250 mg of lithium aluminum hydride in 5 ml of tetrahydrofuran. The mixture was heated to 65° overnight in a sealed tube. Aqueous potassium hydroxide (50%, 1 ml) was added, and the product was extracted with tetrahydrofuran. The solution was dried with anhydrous sodium sulfate and evaporated to dryness in sample capillaries.

Mass Spectrometry. The mass spectra were determined with a CEC 21-110-B mass spectrometer using source temperatures of 150–300°, an ionizing potential of 70 ev, and an ionizing current of 400 μa. The samples (1–20 μg) were introduced through a vacuum lock directly into the ion source. The photographic plate was used for recording the spectra. The method of calculating exact masses from the position of lines on the photographic record has been described earlier.[16]

Acknowledgments. We are indebted to Professors R. B. Woodward and L. C. Vining for samples of the peptides discussed in detail in this paper, to Mrs. V. Beecher for writing the computer program, to the Massachusetts Institute of Technology Computation Center for the use of their facilities, and to the National Institutes of Health for financial support (Grant GM-09352 and GM-05472).

(26) W. A. Wolstenholme and L. C. Vining, *Tetrahedron Letters*, No. 24, 2785 (1966).