Characterization by Tandem Mass Spectrometry of Structural Modifications in Proteins

KLAUS BIEMANN* AND HUBERT A. SCOBLE

Tandem mass spectrometry can be used to solve a number of protein structural problems that are not amenable to conventional methods for amino acid sequencing. Typical problems that use this approach involve characterization of peptides with blocked amino termini or peptides that have been otherwise posttranslationally processed, such as, by phosphorylation or sulfation. The structure and homogeneity of synthetic peptides can also be evaluated. Since peptides can be selectively characterized in the presence of other peptides or contaminants, the need for extensive purification is reduced or eliminated.

IN RECENT YEARS, KNOWLEDGE OF THE PRIMARY STRUCTURE of proteins has become increasingly important. The determination of the structure of a small protein such as insulin was a revolutionary event a few decades ago (*1*). Such work is now limited only by the size of the protein, which determines the time and effort required to determine its amino acid sequence. Two general methods have evolved; first, the automated stepwise sequencing by the Edman technique applied to the protein itself (*2*), and second, the indirect approach through the base sequencing (*3, 4*) of the gene coding for the protein and the translation of the former into the amino acid sequence of the latter through the genetic code. This second methodology is a fast and efficient technique for indirectly deducing the amino acid sequence, particularly of large proteins. However, the final biological product is commonly generated by a number of posttranslational processes that cannot be predicted from the DNA sequence.

The direct determination of the primary structure of a protein by the Edman method overcomes some of these ambiguities but it is also hampered by a number of shortcomings. Most prominent of these is its inability to sequence a peptide or protein that lacks a basic, primary or secondary amino group at the NH_2-terminus and thus cannot react with an isothiocyanate, the first step in the sequential cleavage reaction. The second problem with the Edman technique is its inability to detect certain modified amino acids, either because the chromatographic behavior of the corresponding phenylthiohydantoin (PTH) derivatives is not known, because the modifying group is lost during one of the chemical steps of the Edman procedure, or because the PTH derivative of the modified amino acid is insoluble in organic solvents.

The recent advances in the techniques of protein synthesis bring additional challenges. Genetic manipulations now make it possible to alter proteins routinely, either by modifying the gene coding for the protein or through partial or total synthesis of a modified gene. Alternatively, genes produced by one cell line can be expressed in another cell line that may be easier to culture on a large scale. In these cases one must verify that the products of each cell line are indeed identical or else characterize their differences. Finally, the methodology of chemical protein synthesis has reached a level of sophistication and efficiency that makes the total synthesis of a small protein a reality. In such cases it is important to verify the fidelity of the product.

Clearly, neither DNA sequencing nor the Edman method provide unambiguous answers to all of the questions that may arise and has prompted the need for an alternative approach. Mass spectrometry is the most promising tool to bridge this gap. Recent developments in the ionization of large, polar molecules such as peptides have opened the way to provide the information needed. Conventional electron ionization of suitable derivatives has played an important role in the determination of the nature of peptides with blocked NH_2-termini or the presence of modified amino acids. The discovery of novel ionization methods such as ^{252}Cf-plasma desorption (PD) (*5, 6*) and

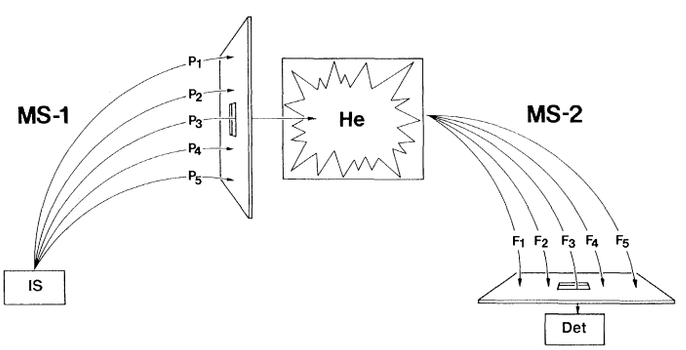


Fig. 1. In FAB-MS a mixture of peptides (corresponding to 0.5 to 1.0 nmol of each peptide) is dissolved in glycerol or a similar polar compound of low vapor pressure, ionized in the ion source (IS) by bombardment with Xe^0 or Cs^+, and the resulting protonated molecular ions $(M + H)^+$, denoted P_1 to P_5, are separated by the first mass spectrometer (MS-1). By adjusting the magnetic field, any one of the $(M + H)^+$ ions can be selected to pass through the collector slit of MS-1 into a collision cell that contains about 10^{-3} torr of helium (He). Collision with a helium atom causes fragmentation of the $(M + H)^+$ ions. The resulting sequence-characteristic fragment ions are mass analyzed by scanning MS-2 and thus MS-2 consecutively transmits the fragment ions through the slit onto the detector (Det). The example schematically indicates the "product" ion (or CID) spectrum that consists of fragments F_1 to F_6 generated from the ^{12}C-component of the "precursor" ion $(M + H)^+$ of peptide P_3.

K. Biemann is professor of chemistry, Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139. H. A. Scoble is visiting scientist, Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, and scientist at Genetics Institute, Inc., 87 CambridgePark Drive, Cambridge, MA 02140.

*To whom correspondence should be addressed.

fast atom bombardment (FAB) (*7*) has eliminated the need for derivatization. The latter technique has now greatly expanded the usefulness of mass spectrometry in this field (*8*).

Conventional Mass Spectrometry

The ability to determine the molecular weights of peptides with an accuracy of better than 1 dalton, even in complex mixtures, has revolutionized certain aspects of protein structure determination, because many questions can be answered by simply determining the molecular weights of peptides generated by specific chemical or enzymatic cleavage of the protein. The verification and, if necessary, correction of the DNA sequence of the gene coding for a particular protein is one of many examples (*9*). Another aspect is the determination of disulfide linkages by the measurement of the molecular weights of the products of proteolytic cleavage of the oxidized protein (*10, 11*). Equally significant is the ability to determine the difference between the predicted molecular weight of a proteolytic peptide and that found on the basis of mass spectrometric data, which indicates a change—either the replacement of one amino acid by another, or the modification of an amino acid present in the peptide. Examples include the identification of hemoglobin mutants (*12*) and the characterization of the active site of an enzyme by photoaffinity labeling (*13*). A modern mass spectrometer measures masses with an accuracy of 1 part in 10,000 up to 12,000 daltons (*14*) and 1 part in 1,000 up to 25,000 daltons (*15*), which is sufficient to reveal almost all such structural changes. Posttranslational processing events can result in the removal of one or more amino acids from the NH_2-terminus, or less frequently from the COOH-terminus, and thus are easily detected by the decrease in molecular weight of NH_2-terminal or COOH-terminal peptides generated by specific cleavage. Furthermore, modifications of any amino acid in the protein manifest themselves in a change (usually an increase) in mass of the modified amino acid (by *N*-acylation, *O*-phosphorylation, *O*-sulfation, *N*- and *O*-glycosylation, or similar processes).

Tandem Mass Spectrometry

Although conventional fast atom bombardment–mass spectrometry (FAB-MS) primarily generates the protonated molecular ion $(M + H)^+$ of the peptides, insufficient energy is transferred in the ionization process to cause significant fragmentation, which is necessary to obtain reliable structural information. Fragmentation can be achieved by collision-induced decomposition (CID), in which the $(M + H)^+$ ion collides with a neutral atom (such as helium) (*16*), or by other methods, such as photon excitation (*17*). In such experiments two consecutive mass analyzers are used, the first (designated MS-1) to generate and select the $(M + H)^+$ ion of the peptide of interest (the precursor ion), which is then passed through a collision cell where fragmentation takes place. The product ions are then mass analyzed in the second mass spectrometer (designated MS-2) and the CID spectrum of the precursor is recorded. Because of this arrangement such instruments are called "tandem mass spectrometers" and the method is also referred to as MS/MS (Fig. 1). It should be emphasized that in CID spectra, unlike conventional FAB spectra, each peak originates from the selected precursor, a feature particularly important when investigating mixtures.

There are in principle various ways in which MS/MS experiments can be conducted. For work on reasonably large peptides of unknown structure, the resolution and sensitivity requirements suggest the use of a four-sector instrument in which both MS-1 and MS-2 are high-mass double-focusing magnetic spectrometers (*18*). For example, the primary structure of a protein has recently been determined exclusively by tandem mass spectrometry (*19*). Quadrupole mass spectrometers can also be used if a lower mass range and lower resolution is sufficient (*20*). Other systems, particularly Fourier transform mass spectrometers (*21*), are promising but have not yet reached the state of development where they are routinely applicable.

We present below a number of typical examples of how structural questions that involve modified proteins can be answered quickly and efficiently by appropriately chosen mass spectrometric measurements.

Peptides with Blocked NH_2-Termini

Human lipocortin (lipocortin I) complementary DNA (cDNA) was recently cloned and expressed in an *Escherichia coli* expression system (*22*). The cDNA codes for a 38.5-kD protein that is approximately 346 amino acids in length. Since the NH_2-terminus of the protein is blocked, the translational start site could not be deduced by the conventional Edman method. For the same reason it was difficult to demonstrate that the native human lipocortin and the *E. coli*–expressed protein have the same sequence in the NH_2-terminal domain.

In order to resolve these questions, the NH_2-terminal peptide produced by tryptic digestion of placental-derived lipocortin I and that from *E. coli*–expressed lipocortin I were isolated by C_{18}

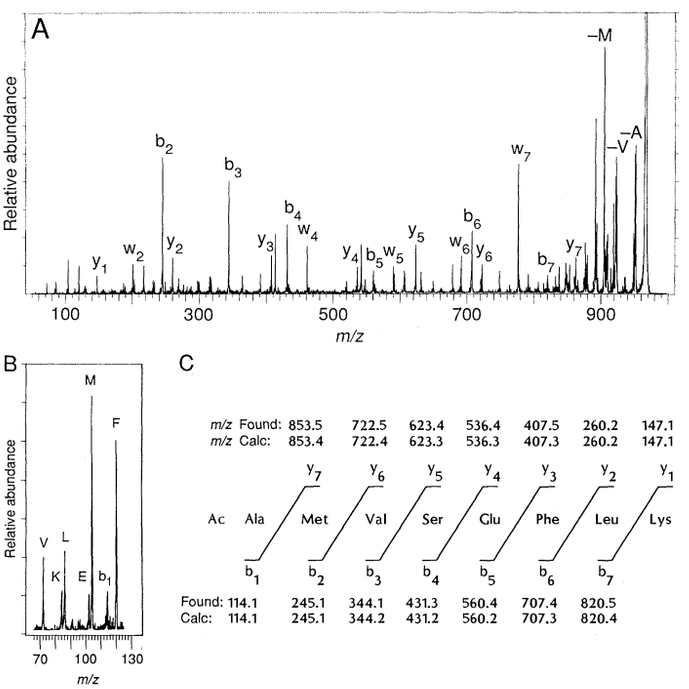


Fig. 2. (**A**) The CID spectrum of the $(M + H)^+$ ion (*m/z* 966.4) of the NH_2-terminal tryptic peptide from lipocortin I and (**B**) the low-mass end of the spectrum (*m/z* 65 to 125). At the low-mass region and the high-mass region peaks are labeled with the single letter code of the amino acid for which they are characteristic (*23*). Lower case letters refer to sequence characteristic fragment ions (see Fig. 3). For clarity only the major ion series are labeled. For example, the peaks 18 daltons below the b_4 and b_5 ions correspond to loss of water from the serine present in these fragments. (**C**) Amino acid sequence as derived from mass spectral data. The y_n and b_n fragment ions are shown including their calculated and experimentally determined masses. Abbreviations: Ac, acetyl; calc, calculated.

Table 1. Monoisotopic residue masses (mass of –NH–CHR–CO–, rounded to the nearest 0.01 dalton) of the common amino acids.

Amino acid	Residue mass (daltons)	Amino acid	Residue mass (daltons)
Gly	57.02	Asp	115.03
Ala	71.04	Lys	128.09
Ser	87.03	Gln	128.06
Pro	97.05	Glu	129.04
Val	99.07	Met	131.04
Thr	101.05	His	137.06
Cys	103.01	Phe	147.07
Ile	113.08	Arg	156.10
Leu	113.08	Tyr	163.06
Asn	114.04	Trp	186.08

reversed-phase high-performance liquid chromatography (HPLC). These two tryptic peptides had the same chromatographic properties and identical molecular weights as determined by FAB-MS. In order to establish their structural identities, the CID spectrum of the $(M + H)^+$ ion at mass-to-charge ratio (*m/z*) 966.4 was generated and revealed the posttranslational modifications that had occurred. The spectrum shown in Fig. 2, A and B, is an example of the structural information that can be obtained from the CID spectrum of a peptide (*23*).

Peaks at the low-mass end of the spectrum represent the immonium ions $(H_2N{=}CHR)^+$ (where R is a side-chain group) that are due to the presence of valine (*m/z* 72), lysine (*m/z* 84, after elimination of NH_3), leucine or isoleucine (*m/z* 86), glutamic acid (*m/z* 102), methionine (*m/z* 104), and phenylalanine (*m/z* 120). The presence of certain amino acids can also be deduced from peaks at high mass that correspond to ions formed by the loss of an amino acid side chain from the precursor ion $(M + H)^+$. The majority of the other peaks in the spectrum is indicative of the sequence (*24*). By searching for consecutive peaks of a mass difference that correspond to the mass of the –NH–CHR–CO– residue of naturally occurring amino acids, one can recognize one or more series of peaks related in this manner (see Table 1 and Fig. 3). In Fig. 2, A and B, they are labeled b_1 to b_7, y_1 to y_7, and w_4 to w_7. For example, the peaks labeled b_2 (*m/z* 245.1) and b_3 (*m/z* 344.1) differ by 99.0 daltons, which corresponds to a valine residue; from b_3 to b_4 the difference is 87.0 daltons, which indicates serine. This can be continued to the admittedly small peak labeled b_7, but this identification is confirmed by the presence of the y_n series of ions. The peaks labeled y_1 and y_2 differ by the same 113.0 daltons (leucine or isoleucine) as those labeled b_6 and b_7. In fact, b_2 to b_7 reads -Val-Ser-Glu-Phe-xLeu- (where xLeu indicates Leu or Ile), whereas the mass differences of the peaks labeled y_1 to y_7 read xLeu-Phe-Glu-Ser-Val-Met-. Clearly, these are different directions of the same sequence. The mass of the peak labeled y_1 indicates that the COOH-terminal amino acid is either lysine or glutamine. The former is the logical choice for the COOH-terminal amino acid of a tryptic peptide, and this was confirmed by reaction with phenylisothiocyanate, which adds 135 daltons to lysine but not to glutamine. Thus, this series must represent the sequence from the COOH-terminus to the NH_2-terminus. The peak labeled b_2 (*m/z* 245.1) must be derived from the first two NH_2-terminal amino acids. Since one of them is methionine (see above), subtraction of its residue mass gives *m/z* 114.1, and there is indeed a small peak at this mass (b_1). This peak could correspond to $H_2N{-}CH(C_4H_9)CO^+$, and indicates a leucine or isoleucine. However, this would result in a free rather than a blocked NH_2-terminus. The only other group that is consistent with this mass assignment is *N*-acetylalanine. Thus the structure of this peptide must be that shown in Fig. 2C.

Although at this point no decision could be made between leucine and isoleucine at position 7, these amino acids can be differentiated on the basis of the third set of peaks labeled w_n in Fig. 2A. These ions are formed by a fragmentation process that ultimately involves cleavage of the bond between the β- and γ-carbons in an amino acid moiety that retains the substituents at the β-carbon (Fig. 3). This permits differentiation of leucine from isoleucine, because for the former the w_2 ion is expected to be at *m/z* 201.1 (as found), whereas the latter would lead to a peak at either *m/z* 215.1 (retention of the β-methyl) or *m/z* 229.1 (retention of the β-ethyl). The w_n series is usually not extensive because aromatic amino acids do not undergo this fragmentation, in agreement with the absence of a w_3 ion in this spectrum. These data indicate that the translational start site of lipocortin I begins at the first of two closely spaced methionines, which is posttranslationally removed, and that the new NH_2-terminal alanine is acetylated. In a similar manner, tandem mass spectrometry has been used to determine the NH_2-terminal sequence of a blocked tetradecapeptide derived from prostatropin, where the first four amino acids were not known (*25*).

Phosphorylated Peptides

Troponin T (TnT), a regulatory muscle protein, is a key component in the troponin-tropomyosin complex that mediates Ca^{2+}-dependent muscle contractility and relaxation. Amino terminal heterogeneity has been reported in muscle from chicken (*26–29*), rabbit (*30*), and in bovine cardiac muscle (*31*). Wilkinson (*28*) suggested that the basis for this variation could be attributed to alternative splicing of the TnT primary RNA transcript, a proposal that has since been confirmed by recombinant DNA analysis of the rat fast skeletal muscle TnT gene (*32*). This work indicated that a number of TnT isoforms differing in their NH_2-termini can be derived from a novel pattern of splicing in the 3′ region of the rat skeletal muscle troponin T gene. The role for this potential hypervariability in muscle contraction is not well understood. The characterization of the NH_2-terminal domain is further complicated by the fact that the NH_2-terminus is blocked, which renders classical Edman sequencing ineffective.

With these limitations in mind, tandem mass spectrometry was used to determine the NH_2-terminal sequence of several isoforms of TnT from rabbit skeletal muscle and bovine cardiac muscle. In each sequencing experiment, the protein was digested by a specific proteolytic enzyme and the resulting peptide mixture was separated

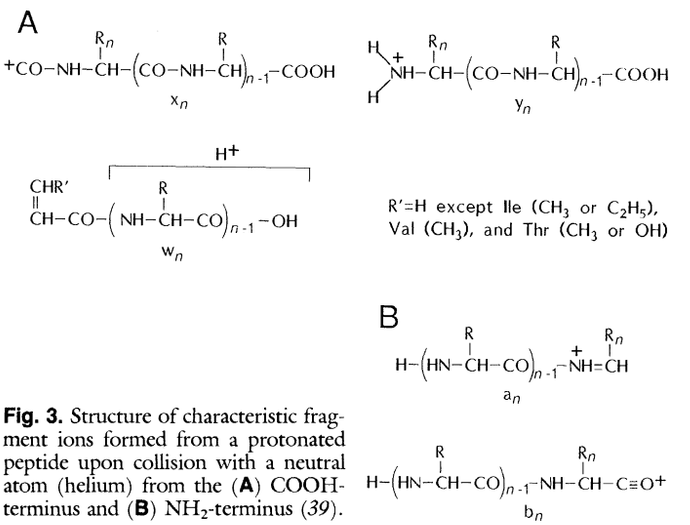


Fig. 3. Structure of characteristic fragment ions formed from a protonated peptide upon collision with a neutral atom (helium) from the (**A**) COOH-terminus and (**B**) NH_2-terminus (*39*).

by chromatography. Those peptides which did not react under Edman conditions were presumed to be NH_2-terminally blocked and were analyzed by tandem mass spectrometry. One such peptide from rabbit skeletal muscle TnT was further purified to apparent homogeneity by reversed-phase HPLC. The conventional FAB spectrum indicated small amounts of several contaminants in addition to the major species, which exhibited an $(M + H)^+$ ion at *m/z* 1323.4. The CID spectrum of this ion (Fig. 4) indicated a series of peaks that differed by 1 of the 18 characteristic masses (Table 1) that correspond to the 20 common amino acids (leucine and isoleucine as well as glutamine and lysine have the same mass). The longest continuous series in the spectrum revealed a sequence -Glu-Val-Glu-His-Val-Glu-Glu-. An additional series of lower abundance indicated a sequence -Glu-Val-Glu-Glu-Asp-. These two sequences must represent an NH_2-terminal series and an overlapping COOH-terminal series, although the respective direction of each cannot be assigned yet.

One interesting feature of this mass spectrum is the unusually abundant ion at *m/z* 1225.4, which is 98.0 daltons less than the precursor ion. This mass loss corresponds to the elimination of H_3PO_4 from a phosphorylated hydroxy amino acid and formation of dehydroalanine or dehydro α-amino butyric acid. The longest continuous series must represent b_n-type ions since the mass difference between the last ion of the series (at *m/z* 1207.4) and *m/z* 1225.4 corresponds to the loss of water from the protonated COOH-terminal carboxyl group to generate an acyl ion (Fig. 3). The nature of the NH_2-terminus (which must be acylated) can be deduced from the mass of the b_2 ion (227.1), which can only correspond to the combination of an acetyl group, a dehydroalanine, and an aspartic acid. From the other series of peaks, which must represent y_n ions and indicate that aspartic acid precedes glutamic acid, one can conclude that the NH_2-terminal amino acid is *N*-acetylphosphoserine and that the structure of this modified peptide is:

$$\begin{array}{c} PO_3H_2 \\ | \\ \text{Ac Ser Asp Glu Glu Val Glu His Val Glu Glu} \end{array}$$

where Ac is acetyl. Ions that result from loss of side-chain groups from the precursor indicate the presence of valine, aspartic acid, and glutamic acid, which provides further support for this sequence. Another unusual feature of this mass spectrum is the presence of the ions at *m/z* 920.3, 1019.5, and 1148.3, which correspond to ions a_7 through a_9 that have not lost the phosphate moiety and thus leave the phosphoserine intact. When this mass spectrum was first determined, it was not known that the peptide was phosphorylated, yet the data clearly revealed the presence of such a group. This group could also be detected by a negative ion tandem mass spectrum,

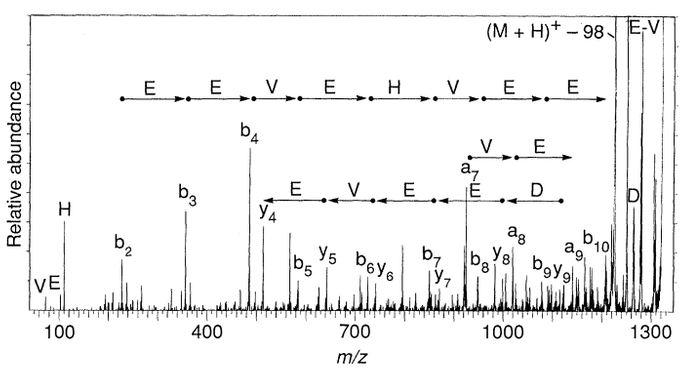


Fig. 4. The CID spectrum of the $(M + H)^+$ ion (*m/z* 1323.4) of the NH_2-terminal peptide of rabbit skeletal muscle TnT (*23*).

Met	Asn	Lys	Trp	Leu	Asn	Thr	Leu	Ser	Lys	Thr	Phe	Thr	Phe	Arg	15
ATG	AAT	AAG	TGG	TTA	AAC	ACA	TTA	TCT	AAG	ACA	TTC	ACT	TTT	CGG	45
Leu	Leu	Asn	Cys	His	Tyr	Arg	Arg	Ser	Leu	Pro	Leu	Cys	Gln	Asn	30
CTT	TTG	AAC	TGT	CAT	TAT	AGG	CGA	TCA	TTA	CCA	CTT	TGT	CAA	AAC	90
Phe	Ser	Leu	Lys	Lys	Ser	Leu	Thr	His	Asn	Gln	Val	Arg	Phe	Phe	45
TTT	TCT	CTG	AAG	AAG	TCG	TTA	ACT	CAT	AAT	CAA	GTC	AGG	TTC	TTT	135
Lys	Met	Ser	Asp	Leu	Asp	Asn	Leu	Pro	Pro	Val	Asp	Pro	Lys	Thr	60
AAA	ATG	AGC	GAT	CTT	GAT	AAT	TTG	CCT	CCA	GTT	GAC	CCA	AAG	ACT	180
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Ile	Glu	Asn	Leu	Lys	Arg	Leu	Lys	Leu	1104						
AAT	GAA	AAC	TTG	AAG	CGT	TTG	AAA	TTG	3312						

Fig. 5. Partial sequence of the gene that encodes valyl-tRNA synthetase from *Saccharomyces cerevisiae*, as determined by Fasiolo *et al.* (*33*). The two putative initiation translation sites are underlined.

which should exhibit a significant peak at *m/z* 81 for the $H_2PO_3^-$ ion. A similar strategy can be employed to investigate sulfated peptides, which are characterized by the loss of 80 daltons (SO_3) from the $(M + H)^+$ ion of peptides that contain sulfated tyrosine.

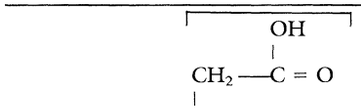
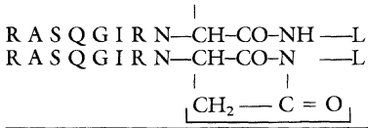
Aminoacyl-tRNA Synthetases

Aminoacyl-tRNA synthetases (tRNA, transfer RNA) play a central role in protein biosynthesis by catalyzing the same basic reaction, namely, the aminoacylation of tRNA by activation of the amino acid followed by transfer to the cognate tRNA. In addition to their various catalytic functions, structural relations, and role in protein biosynthesis, these enzymes have been implicated in the regulation of gene expression. It is necessary to define the primary structure of any enzyme in terms of the amino acid sequence to understand the mechanism and function. In almost all cases this has been accomplished by sequencing the gene that encodes the protein and inferring the primary protein structure. As noted above, this approach is efficient and expeditious but does not provide information about posttranslational processing of the protein to its mature form. The following examples illustrate some of the structural and functional aspects of aminoacyl-tRNA synthetases that we have examined by tandem mass spectrometry.

Recently, Fasiolo *et al.* (*33*) sequenced the gene that encodes valyl-tRNA synthetase from *Saccharomyces cerevisiae* (Fig. 5). The base sequence at the 5′ ends of the transcripts for the *valS* gene showed that there may be two messages transcribed from this region, a short one that contains the downstream ATG and a longer one that contains both the upstream and downstream ATG. The existence of two different messages, each with a different translation initiation site, suggested that the gene may encode two proteins, one of 1058 amino acids and the other of 1104 amino acids. If this single gene encoded both the mitochondrial and cytoplasmic isoenzymes, the 46–amino acid leader could serve as a sequence for importation into the mitochondria. This finding would be analogous to the case of histidine-tRNA synthetase from yeast (*34*) but different from the leucyl-, threonyl-, methionyl-, phenylalanyl-, and tryptophanyl-tRNA synthetases (*35*) from yeast in which the mitochondrial and cytoplasmic isoenzymes are encoded by different genes.

In an effort to identify the initiating ATG, NH_2-terminal sequence analysis of the cytoplasmic valyl-tRNA synthetase by automated Edman degradation was attempted but failed, which indicated that the NH_2-terminus was blocked. The NH_2-terminus of the protein was identified by digesting it with trypsin, partially separating the digest by HPLC, and determining the molecular weights of the peptides in each fraction by FAB-MS. Because the two sites of initiation are separated by 45 amino acids, the detection of tryptic peptides from this intervening region would indicate the upstream methionine as the site of initiation. A total of 80 peptides were

Table 2. Chromatographic and mass spectral data for undecapeptides synthesized by the automated solid-phase method.

Sequence	Retention time (minutes)	$(M + H)^+$ (daltons)
		
R A S Q G I R N—CH–CO–NH —L G	17.40	1186.6
R A S Q G I R N—CH–CO–N —L G	17.92	1168.6
		

characterized by their molecular weights. All but one matched predicted proteolytic peptides beginning after the second initiating methionine, and no peptides from the intervening sequence between the upstream and downstream methionines were detected. The 79 peptides represented approximately 80% of the protein. These data provide a rapid independent corroboration of the correctness of the DNA sequence.

The FAB spectrum (Fig. 6) of fraction 10 from the HPLC separation indicated the presence of at least 11 peptides. All but one of the major peaks correspond to $(M + H)^+$ ions of tryptic peptides predicted for the DNA-derived structure of the protein (minor peaks could possibly correspond to fragment ions and thus are not assigned). However, there was also a strong signal at *m/z* 1351.7 that indicated a peptide of molecular weight 1350.6, which did not correspond to any predicted tryptic peptide from the cytoplasmic enzyme but was in the mass range of a potentially modified NH_2-terminal tryptic peptide if the translational start site was Met^{47}. Thus, a CID spectrum of the ion of *m/z* 1351.7 was recorded and is shown in Fig. 7. Although it is difficult to manually deduce the entire sequence from this CID spectrum, an algorithm that searches and graphically highlights any series of peaks that differ by the mass of an amino acid residue (*36*) simplifies the interpretation process. In this case the partial sequence -Asp-Asn-Leu- (b_3 to b_6) was recognized. It corresponds to the amino acids 51 to 53 of the enzyme. All of the other peaks in the spectrum corroborate the structure Ac-Ser-Asp-Leu-Asp-Asn-Leu-Pro-Pro-Val-Asp-Pro-Lys, which provides further evidence that the initial site of translation is Met^{47}. The identification of this peptide shows that this amino acid had been posttranslationally removed and the new NH_2-terminal serine had been acetylated. These data indicate that the cytoplasmic valyl-tRNA synthetase corresponds to the shorter of the two possible transcripts and that the longer one initiated at the upstream ATG may well be the mitochondrial isoenzyme in analogy with the histidine-tRNA synthetase from yeast.

A different problem was encountered by Regan *et al.* (*37*) while studying the binding of $tRNA^{Ala}$ to alanyl-tRNA synthetase to ascertain the portions of the protein required for recognition by the tRNA. Fragments of *E. coli* alanyl-tRNA synthetase were created by in vitro manipulations of the cloned *alaS* gene. The native protein (875 amino acid, Fig. 8) and analogs, in which from 20 to approximately 500 COOH-terminal amino acids had been deleted, were investigated. All of the fragments that ranged in size from 875 to 385 amino acids were found to specifically bind $tRNA^{Ala}$. Limited digestion of fragment 461N (the segment made from a gene fragment that encodes amino acids 1 to 461) with trypsin generated an NH_2-terminal peptide which, under the conditions of the binding assay, showed no association for $tRNA^{Ala}$. During earlier work on the primary structure of alanyl-tRNA synthetase (*38*), it had been discovered that limited digestion of this protein with trypsin produced an NH_2-terminal fragment, which was termed T-1. This fragment from the native protein and that generated by similar limited tryptic digestion of the 461N fragment appeared to be the same. Analysis of the NH_2-terminal had established that the NH_2-terminal portion of both limited digestion products are identical to the wild-type protein; however, COOH-terminal sequence analysis with carboxypeptidases B and Y was unsuccessful. Because the fragment 385N specifically binds $tRNA^{Ala}$, whereas the T-1 fragment does not, the region bounded by the COOH-terminus of the T-1 fragment and position 385 must be a critical interaction site for tRNA binding. Thus the exact location of the T-1 COOH-terminus was of prime importance.

From earlier structural studies (*38*) it was known that the COOH-terminus should be located between the region bounded by Phe^{366}, the amino acid closest to the COOH terminal as previously defined in T-1, and Lys^{383}, the tryptic site closest to the COOH terminal in 385N. It was believed that the exact location could be pinpointed by the mass spectrometric molecular weight determination of a suitably sized COOH-terminal peptide. For this purpose the T-1–like fragment derived from 461N was treated with cyanogen bromide, which causes cleavage at the COOH-terminal side of methionine. From the known position of the nine methionines in 385N, one could predict the molecular weight of the ten peptides expected from the cleavage reaction, with the exception of the COOH-terminal peptide, for which there are 18 possibilities that range in mass from 2846.4 daltons (Gly^{342} to Phe^{366}) to 4725.5 daltons

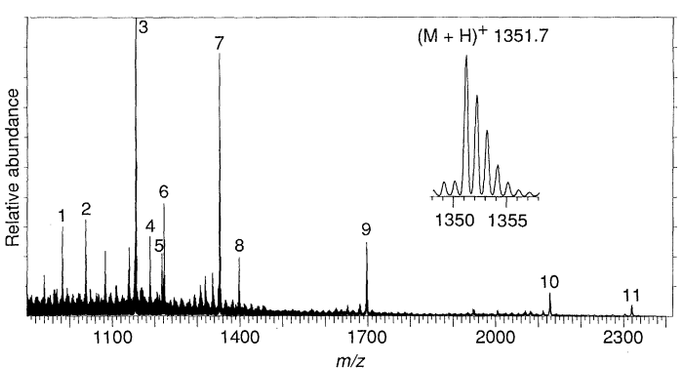


Fig. 6. The FAB spectrum (determined with MS-1 of a tandem instrument) of fraction 10 of an HPLC separation of 2 nmol of tryptic digest of valyl-tRNA synthetase (*23*). Numbers indicate $(M + H)^+$ ions of 11 peptides present in this fraction. The region around peak 7 is expanded to show the well-resolved isotope cluster. The single peak centered around *m/z* 1351.7 (the ^{12}C-component) was selected by MS-1 and the ion beam subjected to collision-induced dissociation, which produced the spectrum recorded by MS-2 and shown in Fig. 7.

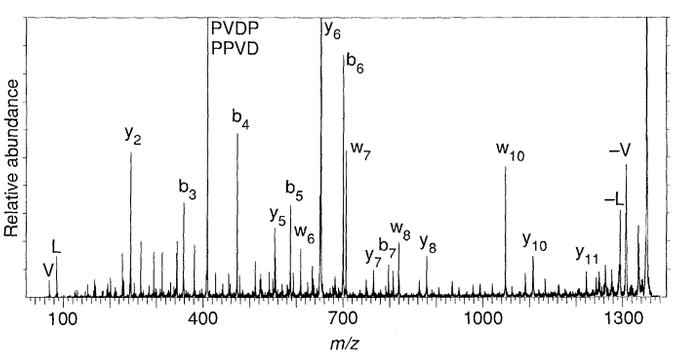


Fig. 7. The CID spectrum of the ion at *m/z* 1351.7 from Fig. 6 (*23*). The high abundance of the peaks at *m/z* 409.2 and 652.4 is typical for cleavage at proline. The notation PVDP and PPVD refers to internal sequence ions where both NH_2-terminal and COOH-terminal fragments have been lost.

Ser	Lys	Ser	Thr	Ala	Glu	Ile	Arg	Gln	Ala	Phe	Leu	Asp	Phe	Phe	15
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Lys	Glu	Thr	Phe	Phe	Tyr	Lys	Leu	Val	Gly	Pro	Leu	Ile	Asp	Val	340
Met	Gly	Ser	Ala	Gly	Glu	Asp	Leu	Lys	Arg	Gln	Gln	Ala	Gln	Val	355
Glu	Gln	Val	Leu	Lys	Thr	Glu	Glu	Glu	Gln	Phe	Ala	Arg	Thr	Leu	370
Glu	Arg	Gly	Leu	Ala	Leu	Leu	Asp	Glu	Glu	Leu	Ala	Lys	Leu	Ser	385
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Asn	Ala	Met	Ile	Arg	Val	Asp	Ser	Ala	Ser	Glu	Phe	Lys	Gly	Tyr	460
Asp	His	Leu	Glu	Leu	Asn	Gly	Lys	Val	Thr	Ala	Leu	Phe	Val	Asp	475
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Ser	Ala	Lys	Leu	Gln	875										

Fig. 8. Partial sequence of the gene encoding alanyl-tRNA synthetase. Met^{341} (in bold) is the most COOH-terminal methionine in the T-1 fragment, so that Gly^{342} must represent the beginning of the COOH-terminal cyanogen bromide peptide. This peptide was found to be that represented by the underlined sequence and thus defined Arg^{368} as the COOH-terminal amino acid of the T-1 fragment.

(Gly^{342} to Lys^{383}). Among the cyanogen bromide cleavage products a $(M + H)^+$ ion of *m/z* 3074.3 was found, which corresponded to a peptide with a mass of 3073.3 daltons. This is consistent with the region from Gly^{342} to Arg^{368} (calculated mass 3073.5 daltons) and is in agreement with formation of the peptide, because position 341 is Met and the COOH-terminus is a tryptic cleavage site. Although this presents a strong argument that T-1 ends at Arg^{368}, it was important to show that this peptide indeed represents the COOH-terminus rather than a random hydrolysis product, as there are five other peptides with a mass of 3074.3 ± 0.5 daltons that could be formed by nonspecific cleavage of T-1. For this reason a CID spectrum of the ion at *m/z* 3074.3 was measured. While the spectrum did not permit the determination of the entire sequence, a series of peaks at *m/z* 1341.5, 1440.7, 1569.6, 1698.0, 1796.9, 1910.2, 2038.1, and 2139.4, which differ by the mass increments 99.2, 128.9, 128.4, 98.9, 113.3, 127.9, and 101.3, indicated a partial sequence . . . -Val-Glu-(Lys or Gln)-Val-(Ile or Leu)-(Lys or Gln)-Thr-. . . . This agrees very well with positions 355 to 361 of alanyl-tRNA synthetase and further proves that this peptide represents the COOH-terminus of T-1. Because the T-1 fragment does not bind $tRNA^{Ala}$, the establishment of Arg^{368} as the COOH-terminus suggests that an important part of tRNA binding by aminoacyl-tRNA synthetase involves a peptide that is short (the 17 amino acids 369 to 385) compared to the size of the protein (*37*).

Synthetic Peptides

The chemical synthesis of peptides has become highly automated and peptides of considerable length can be made by the stepwise addition of amino acids. While these reactions proceed in high yield and generally with little complication, the integrity of the final product must be verified. This is generally done by amino acid analysis and often also by Edman degradation. However, problems may arise that require a different approach and tandem mass spectrometry is again a most suitable technique. The following is a typical example of such a case.

An undecapeptide, R-A-S-Q-G-I-R-N-D-L-G, which has been implicated in the autoimmune disease systemic lupus erythematosus (SLE), was synthesized by the automated solid-phase method. Reversed-phase chromatographic analysis of the synthetic product revealed two clearly resolved peaks. Surprisingly, amino acid analysis and gas-phase Edman sequencing indicated that both had identical composition and the expected sequence. In order to resolve this inconsistency, the FAB spectra of the two components were determined. The data revealed that the early eluting peak produces an $(M + H)^+$ ion at *m/z* 1186.6 as expected, whereas the $(M + H)^+$ ion of the second peak is at *m/z* 1168.6, that is, 18 daltons less. In addition, the mass spectrum of the early eluting compound showed a small peak at *m/z* 1129.6. The CID spectrum of the $(M + H)^+$ ion at *m/z* 1186.6 confirmed that it had the expected structure but that the $(M + H)^+$ ion at *m/z* 1168.6, although similar to the one from the *m/z* 1186.6 ion, showed a shift of 18 daltons to lower mass for some of the ions and the absence of other ions. Both spectra exhibited identical a_n ions up to a_8, which indicated that these two peptides were identical from the NH_2-terminus through the Asn at position 8, and that the structural changes in the peptide of lower mass must have occurred beyond this point. Figure 9 shows that section of the CID spectra of both peptides from which the structural differences can be deduced. The absence (Fig. 9B) of the a_9 and b_9 ions as well as the shift of 18 daltons to lower mass for the a_{10} and $a_{10} - 42$ ions indicate that the modification involves the loss of water from the aspartic acid at position 9 by cyclization at the carboxyl group to form a cyclic imide (Table 2). This is corroborated by the fact that all of the COOH-terminal ions are 18 daltons lower than in the spectrum of the unmodified peptide.

The power of the mass spectrometric approach is further demonstrated by the information obtained from the CID spectrum of the previously mentioned minor peak at *m/z* 1129.6 in the FAB spectrum of the early eluting component. It corresponded to the peptide that lacked the COOH-terminal glycine. Since this peptide coeluted with the major component and is present in relatively low concentration, conventional methods such as amino acid analysis and Edman degradation would not have revealed this impurity in the desired synthetic peptide.

It should be noted that the same conclusion would have been reached if the product of the peptide synthesis had been analyzed directly by mass spectrometry without prior separation by HPLC. A single FAB spectrum would have revealed the inhomogeneity of the material and permitted the determination of the CID spectra of the three components.

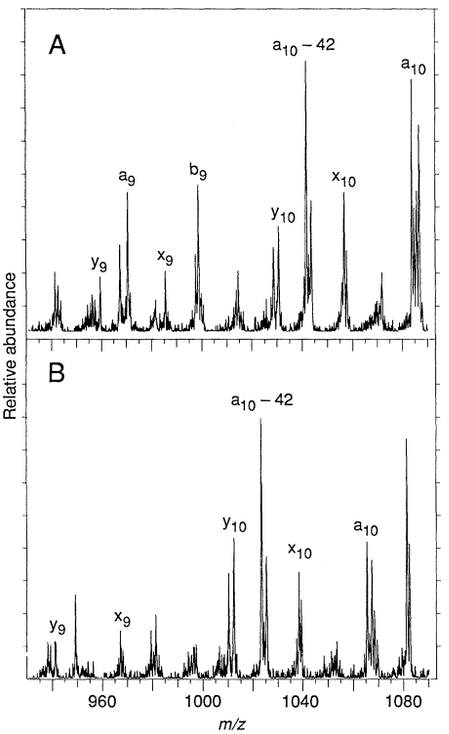


Fig. 9. The CID spectra of synthetic peptides of $(M + H)^+$ of (**A**) *m/z* 1186.6 and (**B**) *m/z* 1168.6 in the mass region from *m/z* 930 to 1090 (*23*).

Conclusions

These examples demonstrate the power of mass spectrometry, particularly tandem mass spectrometry, in solving a variety of structural problems that are not amenable to conventional methods generally used for the structure determination of peptides and proteins. Although the cases discussed deal with only a few structural modifications, it is evident that the same approach can be used for many others, as long as they lead to changes in the mass of the molecule or of fragments formed in the collision cell of the mass spectrometer. These mass differences may be caused by changes of a single amino acid produced by mutation or recombinant techniques, by markers or adducts introduced in the study of active sites of enzymes or receptors, or by posttranslational processing events. Most important is the fact that these mass spectrometric techniques can be applied to mixtures of peptides, which eliminates the time- and material-consuming separation and purification steps. Although the examples discussed here involve tandem mass spectra of peptides that range in molecular weight from 900 to 1400 daltons, complete sequence information has been obtained from peptides of molecular weight up to 2500 daltons (*18*). Since this approach is still in the early stages of application, new developments that extend this range and applicability are to be anticipated.

REFERENCES AND NOTES

1. F. Sanger and H. Tuppy, *Biochem. J.* **49**, 463 (1961).
2. R. M. Hewick, M. W. Hunkapiller, L. E. Hood, W. J. Dryer, *J. Biol. Chem.* **256**, 7990 (1981).
3. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
4. A. M. Maxam and W. Gilbert, *ibid.*, p. 560.
5. D. F. Torgerson, R. P. Skowronski, R. D. Macfarlane, *Biochem. Biophys. Res. Commun.* **60**, 616 (1974).
6. B. Sundqvist *et al.*, *Biomed. Mass Spectrom.* **11**, 242 (1984).
7. M. Barber, R. S. Bordoli, R. D. Sedgwick, A. N. Tyler, *Chem. Commun.* **1981**, 325 (1981).
8. K. Biemann and S. A. Martin, *Mass Spectrom. Rev.* **6**, 1 (1987).
9. B. W. Gibson and K. Biemann, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1956 (1984).
10. H. R. Morris and P. Pucci, *Biochem. Biophys. Res. Commun.* **126**, 1122 (1985).
11. R. Yazdanparast, P. C. Andrews, D. L. Smith, J. E. Dixon, *J. Biol. Chem.* **262**, 2507 (1987).
12. Y. Wada *et al.*, *Biochim. Biophys. Acta* **749**, 244 (1983).
13. S. Chen, T. D. Lee, K. Legesse, J. E. Shively, *Biochemistry* **25**, 5391 (1986).
14. K. Biemann, *Anal. Chem.* **58**, 1289A (1986).
15. B. Sundqvist *et al.*, *Science* **226**, 696 (1984).
16. F. W. McLafferty *et al.*, *J. Am. Chem. Soc.* **95**, 2120 (1973); F. W. McLafferty, Ed., *Tandem Mass Spectrometry* (Wiley, New York, 1983).
17. W. D. Bowers, S.-S. Delbert, R. L. Hunter, R. T. McIver, *J. Am. Chem. Soc.* **106**, 7288 (1984).
18. K. Biemann *et al.*, in *Mass Spectrometry in the Analysis of Large Molecules*, C. J. McNeal, Ed. (Wiley, Chichester, 1986), pp. 131–149.
19. R. S. Johnson and K. Biemann, *Biochemistry* **26**, 1209 (1987).
20. D. F. Hunt, J. R. Yates III, J. Shabanowitz, S. Winston, C. R. Hauer, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 6233 (1986).
21. R. B. Cody, I. J. Amster, F. W. McLafferty, *ibid.* **82**, 6367 (1985).
22. B. P. Wallner *et al.*, *Nature (London)* **320**, 77 (1986).
23. All CID spectra shown were determined with an instrument with two consecutive double-focusing JEOL HX110 mass spectrometers and a collision cell between them (*14*). The mass range was 14,500 daltons at full accelerating voltage (10 keV). In general the resolution of MS-1 was set to pass the ^{12}C species of the $(M + H)^+$ isotopic cluster into the collision cell; MS-2 was set to allow measurement of the product ions with ≤0.3 dalton accuracy. All mass spectra are shown as raw peak profiles and represent single scans from *m/z* 50 to the precursor ion in 1 to 2.5 minutes. Approximately 0.5 to 1.0 nmol of peptide was used for each spectrum. The normal FAB spectrum (recorded by MS-1) shown in Fig. 6 was recorded in 2.5 minutes from *m/z* 100 to 4000. A JEOL DA5000 data system was used to acquire and process the data. Abbreviations used for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
24. S. A. Martin and K. Biemann, *Int. J. Mass Spectrom. Ion Processes*, in press.
25. J. W. Crabb *et al.*, *Biochemistry* **25**, 4988 (1986).
26. S. V. Perry and H. A. Cole, *Biochem. J.* **141**, 733 (1974).
27. J. M. Wilkinson, *ibid.* **169**, 229 (1978).
28. ———, A. J. Moir, M. D. Waterfield, *Eur. J. Biochem.* **143**, 47 (1984).
29. I. M. Bird, G. K. Dhoot, J. M. Wilkinson, *ibid.* **150**, 517 (1985).
30. M. M. Briggs, R. E. Klevit, F. H. Schachat, *J. Biol. Chem.* **259**, 10369 (1984).
31. N. B. Gusev *et al.*, *Biochem. J.* **213**, 123 (1983).
32. R. E. Beitbart *et al.*, *Cell* **41**, 67 (1985).
33. F. Fasiolo *et al.*, unpublished results.
34. G. Natsoulis, F. Hilger, G. R. Fink, *Cell* **46**, 235 (1986).
35. L. K. Pape, T. J. Koerner, A. Tzagoloff, *J. Biol. Chem.* **260**, 15632 (1985); A. M. Myers and A. Tzagoloff, *ibid.*, p. 15371; L. K. Pape and A. Tzagoloff, *Nucleic Acid Res.* **13**, 6171 (1985); J. M. Schneller, C. Schneller, R. Martin, A. J. C. Stahl, *ibid.* **3**, 1151 (1976); J. M. Schneller, C. Schneider, A. J. C. Stahl, *Biochem. Biophys. Res. Commun.* **85**, 1392 (1978).
36. H. A. Scoble, J. E. Biller, K. Biemann, *Fresenius Z. Anal. Chem.* **327**, 239 (1987).
37. L. Regan, J. Bowie, P. Schimmel, *Science* **235**, 1651 (1987).
38. S. D. Putney *et al.*, *ibid.* **213**, 1497 (1981).
39. This nomenclature is a slight variation of that proposed by P. Roepstorff and J. Fohlman, *Biomed. Mass Spectrom.* **11**, 601 (1984).

40. We are indebted to S. A. Martin and P. Schimmel for stimulating discussions and J. E. Biller for the graphics display algorithms. We thank M. M. Briggs and F. H. Schachat for the TnT sample and K.-S. Huang for the lipocortin I sample. The work at MIT was supported by NIH grants RR00317 and GM05472.