Low-Mass Ions Produced from Peptides by High-Energy Collision-Induced Dissociation in Tandem Mass Spectrometry

A. M. Falick,[*] W. M. Hines,[*] K. F. Medzihradszky,[*] M. A. Baldwin,[†] and B. W. Gibson[*]

[*]Mass Spectrometry Facility, Department of Pharmaceutical Chemistry, and [†]Department of Neurology, University of California, San Francisco, California

High-energy collision-induced dissociation (CID) mass spectrometry provides a rapid and sensitive means for determining the primary sequence of peptides. The low-mass region (below mass 300) of a large number of tandem CID spectra of peptides has been analyzed. This mass region contains several types of informative fragment ions, including dipeptide ions, immonium ions, and other related ions. Useful low-mass ions are also present in negative-ion CID spectra. Immonium ions (general structure $[H_2N{=}CH\text{-}R]^+$, where R is the amino acid side chain) and related ions characteristic of specific amino acid residues give information as to the presence or absence of these residues in the peptide being analyzed. Tables of observed immonium and related ions for the 20 standard amino acids and for a number of modified amino acids are presented. A database consisting of 228 high-energy CID spectra of peptides has been established, and the frequency of occurrence of various ions indicative of specific amino acid residues has been determined. Two model computer-aided schemes for analysis of the amino-acid content of unknown peptides have been developed and tested against the database. *(J Am Soc Mass Spectrom 1993, 4, 882–893)*

High-energy, tandem collision-induced dissociation (CID) mass spectrometry provides a rapid and sensitive means of determining or verifying the amino acid sequences of peptides [1–6]. Mass spectrometric methods have several advantages over more conventional methods. First, they generally work as well for modified or unusual amino acid residues as for the 20 standard ones. Second, the peptide of interest need not be exhaustively purified; the method usually works as well on simple mixtures as on isolated peptides. Finally, acquisition of a tandem CID mass spectrum takes only a few minutes regardless of the composition of the peptide. At present, interpretation of a spectrum of an unknown is less straightforward, but programs for computer-aided interpretation are being developed [7–11]. Interpretation of a peptide CID spectrum relies mainly on the sequence-specific ions present (a_n, b_n, c_n, d_n, x_n, y_n, z_n, w_n, and v_n ions in the nomenclature originally proposed by Roepstorff and Fohlman [12] and elaborated by Biemann and co-workers [1, 7, 13]). However, any additional information about the nature of the peptide that can be inferred, either from the CID spectrum or otherwise, can enhance the performance of any interpretation strategy by reducing the number of candidate sequences to be considered and by resolving ambiguities.

The presence of immonium ions in a normal (2-sector) fast-atom bombardment (FAB) mass spectrum corresponding to various amino acid residues present in a peptide was first reported by Barber et al. [14, 15]. Renner and Spiteller [16] characterized the probable formation pathway for such ions, and Kausler et al. [17] systematically studied the immonium and their related fragment ions produced from amino acid residues in peptides by ion bombardment in a liquid matrix.

Lippstreu-Fisher and Gross [18] reported a few peptide-derived immonium ions in a 3-sector, tandem CID study. Biemann and co-workers [1, 2, 4, 5, 11] and Madden et al. [19] have commented on the presence of such ions in 4-sector, tandem, high-energy CID spectra of peptides. A considerable amount of useful information can be obtained from these published spectra, but no systematic study has yet been undertaken to determine the occurrence and characteristics of immonium and other low-mass ions with a view to eventually incorporating this information into a "second-generation" computer-based scheme for interpretation of peptide CID spectra, using, for example, a pattern-based approach [8]. It is important to study all of the

Address reprint requests to Prof. Arnold M. Falick, Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-0446.

1044-0305/93/$6.00

Received June 21, 1993
Revised July 6, 1993
Accepted July 7, 1993

ions present in CID spectra, because of their potential usefulness and also as a means of uncovering previously uncharacterized fragmentation modes (see ref 20 for example). In the present study, we have somewhat arbitrarily concentrated on the region below mass 300. We have studied a large number of high-energy CID mass spectra of peptides containing both normal and modified amino acids to better correlate the observed low-mass ions with peptide structure, in particular with the presence or absence of various residues in the peptide. As is shown below, a great deal of information can be obtained from a careful examination of the ions of relatively low mass in such spectra.

Experimental

Mass Spectrometry

Tandem mass spectrometry experiments were carried out on a Kratos (Kratos Analytical Instruments, Manchester, U.K.) Concept IIHH 4-sector instrument of EBEB geometry [21], equipped with a liquid secondary ion mass spectrometry source [22, 23] and a multichannel array detector [24] capable of acquiring data simultaneously over a 4% mass window. Precursor ions were generated with an 18 keV Cs^+ primary beam. The acceleration voltage in the first mass spectrometer (MS-I) was 8 kV and the collision energy for CID was 4 keV. The collision gas (He) was used at a pressure sufficient to suppress the precursor ion beam to about 30% of its initial value. The instrument was controlled and data were acquired with a DS-90 data system. Spectra were acquired by stepping, in 4% decrements, from the precursor protonated peptide mass down to the lower mass limit. For simplicity, this process is referred to below as a scan. The lower limit was set to just below mass* 60 for most spectra, but a subset of earlier spectra were scanned to about mass 70. Relatively little useful information is contained below mass 59, and a relatively long time is required to scan to lower masses on the system used for these measurements. (For example, the length of time required to scan from mass 60 to 30 is the same as that required to scan from 600 to 300.) Only peaks below an arbitrary cutoff of mass 300 were analyzed for this study. Calibration and data display were carried out with the aid of a Mach 3 data-processing system. Tandem CID spectra of more than 200 peptides were used in this study. Commercially available peptides were obtained from either Sigma (St. Louis, MO) or Bachem (Torrance, CA) and were used without further purification. (In a tandem CID mass spectrum, molecular ions from compounds of a different M_r than the desired peptide are normally not transmitted through MS-I, although some artifacts have been noted [26].) Numerous other peptides were obtained as by-products of sequence analysis studies of various proteins of known or subsequently verified sequences. In all of the latter cases, the peptides had undergone purification by high-performance liquid chromatography prior to tandem mass spectrometry.

Data analysis. CID spectra were originally recorded as peak profiles. Automated data reduction of spectra was performed as follows. A peak detection routine, as previously described [8], was applied to each spectrum. For convenience, abundances were taken simply as peak heights. A threshold was applied to the resulting spectrum, which consisted of a list of peak abundances versus nominal masses, as follows. The number of peaks to be retained in the final list was equal to the nominal MH^+ mass divided by 10 (90 peaks for a peptide of MH^+ 900), thus yielding a final average peak density in the spectrum of 1 peak per 10 Da. For all of the subsequent statistical work, except Scheme B, peak abundance data were not used except to establish the existence of a peak at a given mass. For Scheme B, peaks were classified as strong, medium, or weak in the following way. All of the peaks below mass 200 were renormalized such that the largest peak in this range was assigned an arbitrary abundance value of 25. Those peaks with a renormalized abundance of between 5 and 25 were deemed strong, those between 1 and 5 as medium, and the remainder were defined as weak. (This classification is crudely logarithmic.) Computer programs for analyzing spectra were written in C and run on a SPARCstation IPX™ computer (Sun Microsystems, Mountain View, CA). Additional programs were written to drive the prediction software and collate the results. Our database for this study was derived from a larger selection of CID spectra of peptides of known sequence. Some spectra were excluded from the final database either to reduce the number of peptides with highly similar sequences or because of the quality of the spectra (due to either saturation of daughter ions or low signal intensity). However, some spectra of lesser quality were included to increase the diversity of our sample with respect to overrepresented and underrepresented amino acids.

Interpretive programs were designed to suggest amino acid residues that were likely absent ("excluded") or present ("required") in the peptide from which the spectrum under examination was produced. In cases where no prediction could be made, the residue was "allowed." Predictions were restricted to genetically encoded amino acids. Predictions for Gly and Ala were not attempted, as we did not scan low enough to detect their immonium ions. No attempt was made to distinguish between the isomeric pair, Leu and Ile, and the isobaric pair, Lys and Gln.

*We refer to fragment ion mass rather than mass-to-charge ratio in this article because the fragment ions are known to be singly charged. A multiply charged fragment ion formed from a singly charged precursor in the collision cell of a 4-sector, tandem machine under normal operating conditions (floated collision cell) would not appear at all in the fragment-ion spectrum. Such an ion would not be simultaneously transmitted through E2 and B2 of MS-II during the linked scan of these two fields [25].

Results and Discussion

The identifiable peaks observed in the low mass region of the peptide CID spectra we examined can conveniently be divided into three classes. The first class consists of normal sequence ion peaks that happen to fall in this mass range. These were easily identified and were not studied further. Second are dipeptide-ion peaks arising from internal acyl ions or immonium ions (designated by the dipeptide single-letter code or the dipeptide single-letter code − 28, respectively) [4, 13]. The internal acyl ions have masses equal to the sum of the residue masses plus one and correspond to $\mathbf{y}_m\mathbf{b}_n$ cleavages, while the immonium ion mass ($\mathbf{y}_m\mathbf{a}_n$ cleavage) is 28 Da less than that of the respective acyl ion.

In our data, dipeptide ions are often abundant when one of the residues gives rise to a strong immonium ion of its own and/or when cleavage at one of the residues is favored, such as at the N-terminal side of Pro or, to a lesser extent, Gly. Peaks corresponding to losses of water (if the dipeptide contains Thr, Ser, Asp or Glu) or ammonia (from Arg, Lys, Gln, or Asn) from dipeptide ions are sometimes also observed. These are particularly intense when the dipeptide is at the N-terminus (i.e., $\mathbf{a}_2 - 17$ or 18 and $\mathbf{b}_2 - 17$ or 18 ions). On the other hand, peaks corresponding to immonium or acyl dipeptide ions from the two C-terminal residues of a peptide are only rarely observed.

Dipeptide ion peaks can aid in interpretation of CID spectra. It is frequently the case that two more-or-less equally satisfactory interpretations of a spectrum of an unknown peptide differ by a dipeptide inversion, for example, VSWAMFPNGK or SVWAMFPNGK. The presence of a strong dipeptide signal for SW (mass 274) and not for VW (286) was an important clue to the correct sequence in this case [27].

The third class of peaks found in the low-mass region is due to immonium and related ions. Immonium ions have the general structure $[H_2N{=}CH\text{-}R]^+$, where R is the amino acid side chain; their mass is 27 Da less than the residue mass [13]. The related ions are those observed in the CID spectrum when certain residues are present in a peptide. Immonium ions for most of the standard amino acid residues and for many modified residues are prominent in high-energy CID mass spectra and offer the obvious possibility of providing useful information about the amino acid composition of an unknown peptide. Compositional constraints thus determined can be used to assist subsequent sequence determination.

In Table 1 are listed the observed immonium and related ion masses. Table 1 is a qualitative summary of manual examination of several hundred peptide CID spectra. Much of the information contained in Table 1 is in accord with previous work [11]. Figure 1 shows the low mass region of the tandem CID spectrum of the peptide Ac–Gly–Ile–Gln–Glu–Leu–Tyr–Gly–Ala–Ser–NH_2. The peaks in this spectrum are labeled according to the conventional nomenclature [13]. The spectrum contains abundant immonium ion peaks for the residues Ile/Leu (mass 86), Gln (84, 101), and Tyr

Table 1. Immonium and related ions characteristic of the 20 standard amino acids

Amino acid residue	Immonium ion mass	Related ion masses	Comments
Ala	44		
Arg	129	112, 100, 87, 73, 70, 59	129, 73 usually weak
Asp	88		Weak or absent if C-terminal
Asn	87	70	87 often weak, 70 weak
Cys	76		Usually weak
Gly	30		
Gln	101	84, 129	129 weak
Glu	102		Often weak if C-terminal
His	110	82, 121, 123, 138, 166	110 very strong; 82, 121, 123, 138 weak or absent
Ile /Leu	86	72	86 strong, 72 very weak
Lys	101	129, 112, 84, 70	101 can be weak, 70 weak or absent
Met	104	61	104 often weak or absent
Phe	120	91	120 strong, 91 weak
Pro	70		Strong
Ser	60		
Thr	74		
Trp	159	171, 170, 130, 117	Strong except 117
Tyr	136	107, 91	136 strong; 107, 91 weak
Val	72		Fairly strong

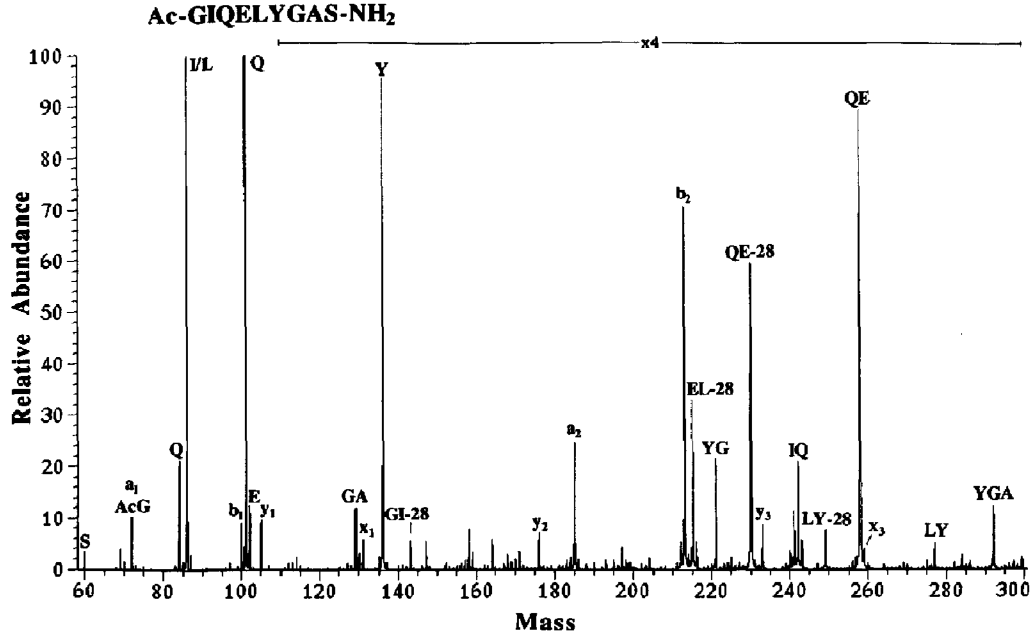

Figure 1. Tandem CID mass spectrum of the peptide Ac-Gly-Ile-Gln-Glu-Leu-Tyr-Gly-Ala-Ser-NH_2. Single amino acid immonium and related ion peaks are marked with their corresponding single letter code. Dipeptide acyl and immonium ion peaks are also indicated (see text) as are the normal sequence ions.

(136); weaker immonium ion peaks can be seen for Ser (60), Glu (102), and Ac-Gly (72). (Note that the peak at mass 72 has the same nominal mass as the valine immonium ion and therefore would be masked if valine were present.) A number of internal dipeptide ions, and one tripeptide ion, are also present.

The usefulness for interpretation purposes of the immonium and related ions varies a great deal for different amino acid residues. For example, the cluster of ions produced by Arg is usually easily recognized, as are the characteristic ions from His, Leu/Ile, Phe, Tyr, and Trp. The apparent stability of these ions can be rationalized on the grounds of favorable resonance effects or inductive effects from alkyl side-chain groups. The immonium ions of a few residues, such as Asp and Glu, are sometimes weak or absent, perhaps because the electron-withdrawing carboxyl group destabilizes the positive ion. Madden et al. [19] have noted that immonium ions from N-terminal residues (a_1 ions) tend to be enhanced in abundance relative to the immonium ion from the same residue located in the C-terminal position. They also suggest that some competition for charge may occur among immonium ions. Our data show quite clearly that the abundance of immonium ions from residues in the C-terminal position is markedly less than for residues in any other position in the peptide. Indeed, the lack of immonium ions for some C-terminal residues is a major obstacle to attempts to automate interpretation.

In some cases, interferences occur. Proline produces a strong ion at mass 70 for which a stable cyclic structure can be drawn and which is an excellent indication of the presence of this residue; however, Arg also gives rise to a strong ion at this same mass, so that the presence of Pro cannot be confirmed in the presence of Arg. Similarly, the mass 87 ion of Asn is masked if Arg is present. In our experience, it is not possible to distinguish between the isobaric pair of Gln and Lys on the basis of their immonium ions: both can produce peaks at mass 84 and 101, although the mass 101 peak due to Lys is sometimes quite weak. A distinction between Lys and Gln can sometimes be made based on clues obtained from sequence ions in the spectrum. However, we believe that, at this stage, these clues are too subtle to attempt the distinction via software. Acetylation of the ϵ-amino group of Lys followed by molecular weight measurement [28] or, if necessary, a second tandem CID spectrum (ϵ-acetyllysine has a distinctive immonium ion, see below), can provide a definitive answer.

The "related ions" listed in Table 1 usually can be rationalized easily. Their masses generally correspond to neutral losses or other logical modifications of the standard immonium ions. For example, the ion produced by Met at mass 61 corresponds to $CH_2{=}SCH_3^+$. Arginine is the source of a large number of low-mass fragment ions, presumably because of its ability to stabilize positive charge at a number of different sites.

The actual structures of these and other ions referred to in this paper are, in general, unknown.

A variety of peptides containing nonstandard or modified amino acid residues has also been examined. One of the most important pieces of information available from the low-mass ions in a CID spectrum is indication or confirmation of the presence of a modified or nonstandard amino acid residue. In fact, the ability to detect and identify such residues is a major advantage of mass spectrometric methods. Table 2 is a list of a number of modified residues and the masses of the associated characteristic positive-ion, high-energy CID peaks observed. The immonium ions generated from modified residues nearly always produce peaks at the expected mass, namely 27 Da less than the modified residue mass. In the case of N-terminal modifications, the immonium ion is of course identical to the $\mathbf{a}_1$ ion. Acetylated N-terminal residues usually also give rise to a peak at the "normal" immonium ion mass (42 Da less). Residues at the C-terminus of a peptide that have a modified carboxyl group (esters, for example) show only the normal immonium ion because the carboxyl group is not present in any case in the immonium ion. Dihexylated [29] C-terminal residues (Asp, Glu), usually show an immonium ion containing one hexyl group as well as the normal immonium ion, which is often weak. Side-chain hexyl esters of aspartate, glutamate, and carboxymethylcysteine in non-C-terminal positions also give immonium ion peaks both with and without the modifying groups.

As an example of a peptide containing modified residues, Figure 2 shows the low mass region of the tandem CID spectrum of the peptide pyroGlu–Tyr–Gly–Phe–Cys*–Lys. The cysteine residue has been modified with 4-vinylpyridine [30] to convert it to S-β-4-ethylpyridyl cysteine. This modified residue gives a very distinctive group of characteristic peaks, dominated by mass 106 ($CH_2{=}CH\text{-}C_5H_5N^+$) (see Table 2). The expected peak at mass 84 from pyroGlu is masked in this case by the Lys peak at the same mass.

Negative-Ion Peaks

The phosphorylated and sulfated amino acids are the only amino acids documented [31, 32] to give strong characteristic low-mass peaks in negative-ion, high-energy CID spectra of peptides. However, no systematic study has been conducted to determine the presence or absence of characteristic low-mass ions for individual amino acids in the negative-ion mode. Phosphoserine, phosphotyrosine, and sulfated tyrosine all give abundant peaks corresponding to deprotonated immonium ions ($NH{=}CH\text{-}R^-$), where the negative charge is presumably located on the side-chain phosphate or sulfate group. This gives rise to a peak 2 Da lower than the corresponding immonium-ion peak in the positive-ion mode (i.e., mass 214 for phosphotyrosine and sulfotyrosine and mass 138 for phosphoserine). The latter also gives abundant peaks at mass 79 and 97 [32]. Phospho-

Table 2. Immonium and other ions from nonstandard amino acid residues

Amino acid residue	Immonium ion mass	Related ion masses	Comments
Acetylglutamate	144	102	N-terminal
ε-Acetyllysine	143	126	
Acetylproline	112		N-terminal
Acetyltyrosine	178	136	N-terminal
Acetylphenylalanine	162	120	N-terminal
Aminoethylcysteine	120		
3-Carboethoxyhistidine	182	110	
Carboxymethylcysteine	134		
S-β-4-Ethylpyridyl cysteine	181	138, 106, 93	106 strong, 181, 138, 93 can be weak
β-Hexyl aspartate	172	88	88 weak
Hexyl carboxymethylcysteine	218	134	
γ-Hexyl glutamate	186	102	102 weak
Hydroxyproline	86		
3-Hydroxytyrosine	152		
4-Methylphenylalanine	134		
Norleucine	86		
Phosphoserine	140		
Phosphotyrosine	216		Moderate to strong
Pyroglutamic acid	84		N-terminal
Sulfotyrosine	216		

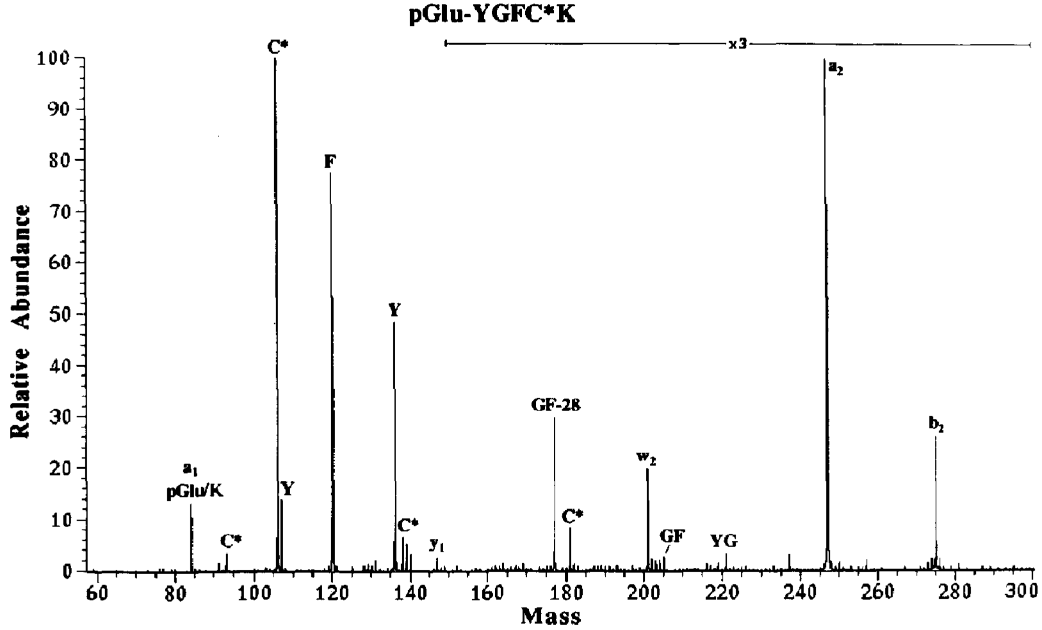

Figure 2. Tandem CID spectrum of the peptide pyroGlu-Tyr-Gly-Phe-Cys*-Lys. Cys* is S-β-4-ethylpyridyl cysteine.

tyrosine and sulfotyrosine give rise to a further series of fragment peaks, shown in Table 3.

The occurrence of characteristic or "indicator" peaks that are associated with the presence of a corresponding residue in a peptide can be quantified. Due to the limited amount of data available for modified residues and in negative mode, the quantitative work has been limited to the positive-ion CID spectra of peptides containing the 20 genetically encoded amino acids. Of these, we have grouped the isomeric pair, Leu and Ile, and the isobaric pair, Gln and Lys, together. We have further neglected Gly and Ala because most spectra were not scanned to low enough mass to detect their immonium ions. Our study sought to determine how strongly the presence or absence of an indicator peak correlates with the presence or absence of the corresponding amino acid. The closely related question, that of whether the presence of an amino acid in a peptide will give rise to the indicator peak in its CID spectrum, is also addressed.

These questions are best answered using a computer-accessible database of CID spectra, which we have generated from our collected data. Initially, peaks were classified as present or absent based on whether their abundances exceeded the calculated threshold value (see Experimental section). It should be noted that peaks that fall below the threshold for a given spectrum may still be useful to a skilled human interpreter. Similarly, peaks that exceed the threshold may be discounted based on experience. This last observa-

Table 3. Characteristic low mass ions found in high-energy negative ion tandem CID mass spectra of peptides containing phosphotyrosine and sulfotyrosine

Residue	Mass of characteristic peak	Suggested structure
Phosphotyrosine	214	$NH{=}CH\text{-}CH_2\text{-}C_6H_4OPO_3H^-$
	199	$CH_2{=}CH\text{-}C_6H_4OPO_3H^-$
	186	$\cdot CH_2C_6H_4OPO_3H^-$
	97	$H_2PO_4^-$
	79	PO_3^-
Sulfotyrosine	214	$NH{=}CH\text{-}CH_2\text{-}C_6H_4OSO_3^-$
	199	$CH_2{=}CH\text{-}C_6H_4OSO_3^-$
	186	$\cdot CH_2C_6H_4OSO_3^-$
	96	$\cdot SO_4^-$
	80	$\cdot SO_3^-$

tion is particularly true for small peptides (less than six residues).

The results of our analysis of the 228 spectra in the database are shown in Table 4, which gives the frequency of occurrence (in %) of indicator peaks in the CID spectra of peptides containing each amino-acid residue. For the purposes of the following discussion, it is convenient to define four logical categories of spectra. The designation T+ indicates a spectrum that contained both an amino acid and its indicator peak, and T− indicates spectra containing the amino acid but not a detectable indicator peak (above threshold). Similarly, F+ and F− describe spectra of peptides not containing the residue in question but which do and do not, respectively, contain the corresponding indicator peak.

In most cases, a total of N = 228 spectra were examined, but because some spectra were not scanned low enough, N varies for ions of mass 59, 60, and 61, as indicated in Table 4. All indicator peaks listed show some positive correlation with their respective amino-acid residue. As an example, consider the mass 74 peak for Thr. 82% (71/87) of the spectra with a peak at mass 74 were from peptides containing Thr. Furthermore, only 6% of those spectra without a peak at mass 74 were from peptides that contained a Thr residue. Thus, if our collection of spectra is representative, one could say that if there is a peak at mass 74, the probability that the peptide contains threonine is 82%, and similarly, if there is no peak at this mass, the likelihood that threonine is present is 6%. The situation is somewhat different for mass 70 as an indicator for Pro. In this case, the presence of a peak at this mass only correlates with the presence of Pro in 45% of the cases. This is mainly due to the mass 70 ion that is also associated with Arg. However, if there is no peak at mass 70, then one can be 100% certain (within the limits of the database) that Pro is absent. Negative information of this type is just as valuable as positive information in the context of assisting in the interpretation of the CID spectrum of an unknown peptide.

A different perspective is illustrated in Figure 3. This histogram shows the fraction of the spectra containing each individual indicator peak that can be classified into the four logical categories, T+, T−, F+ and F−. The histogram shows the first three cases; the fourth, F−, is given by the difference between the sum of the first three and unity.

It is evident from Tables 1–3 and Figure 3 that a good deal of information about the amino-acid composition of an unknown peptide can be obtained from a thoughtful examination of the low-mass portion of the CID spectrum, but that this approach is not without pitfalls. We therefore used the database to test various possible schemes for computer-aided interpretation of amino acid composition. This "amino-acid–composition module" was designed ultimately to be used in combination with the computer-aided spectral interpretation program developed in our laboratory [8]. As

Table 4. Frequency of occurrence of indicator peaks and amino acid residues in the data base[a]

Residue	N	n	mass	m	T+/m	T − /(N-m)
			(Da)		(%)	(%)
Cys	228	9	76	5	20	4
Asp	228	84	88	104	62	15
Glu	228	98	102	92	87	13
Phe	228	79	91	52	71	24
	228	79	120	86	90	1
His	228	55	82	64	52	13
	228	55	110	89	61	1
	228	55	121	50	76	10
	228	55	123	33	91	13
	228	55	138	60	62	11
	228	55	166	70	69	4
Lys /Gln	228	134	70	196	56	75
	228	134	84	189	68	15
	228	134	101	98	80	43
	228	134	112	123	54	65
	228	134	129	141	80	24
Ile /Leu	228	191	72	157	80	92
	228	191	86	200	94	14
Met	162	9	61	5	100	3
	228	16	104	22	64	1
Asn	228	59	87	104	31	22
Pro	228	89	70	196	45	0
Arg	94	39	59	27	89	22
	228	85	70	196	43	3
	228	85	73	66	43	3
	228	85	87	137	60	3
	228	85	100	112	71	4
	228	85	112	123	67	2
	228	85	129	141	30	48
Ser	162	63	60	64	61	24
Thr	228	79	74	87	82	6
Val	228	119	72	157	73	6
Trp	228	12	117	23	30	2
	228	12	130	125	10	0
	228	12	159	91	13	0
	228	12	170	51	24	0
	228	12	171	98	24	8
Tyr	228	55	91	52	50	16
	228	55	117	34	85	13
	228	55	136	73	70	3

[a] N = total number of spectra examined, n = number of spectra containing the given amino acid residue, and m = number of spectra containing a peak at the indicator mass. T+/m is the fraction of the spectra containing the peak that also contain the corresponding amino acid, and T − /(N-m) is the fraction of the spectra that did not contain the given indicator peak but came from peptides that did contain that amino acid residue.

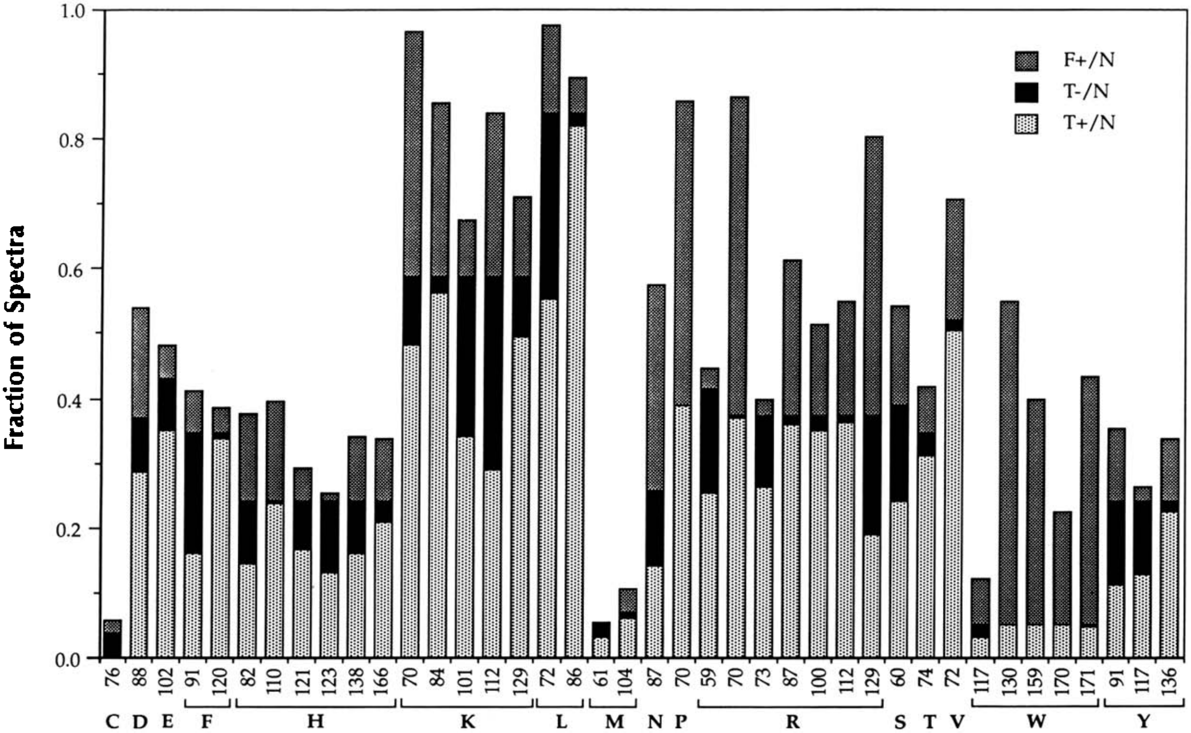

Figure 3. Histogram showing the fraction of spectra in the database that contain certain indicator peaks and the corresponding residues.

mentioned, the database did not contain sufficient spectra to include modified or nonstandard amino acid residues; however, it is perfectly possible to add additional rules in the future.

The logical rules for a rather simple scheme (Scheme A) based mainly on the presence or absence of various immonium ions are shown in Table 5. In this scheme, if the CID spectrum contains a peak (above threshold) at mass 102, for example, all sequences proposed by the interpretation module of the program would be constrained to include at least one Glu residue. Similarly, if this peak is absent, any proposed sequences containing Glu would be excluded. In a slightly more complicated case, the presence of three or more peaks at masses characteristic of His would be sufficient to require the presence of this residue in all candidate sequences, while the presence of fewer than two would exclude it. If exactly two of these indicator peaks were present, no firm conclusion could be made, so that all candidate sequences, whether or not they contained His, would be allowed.

The logic for a more sophisticated scheme (Scheme B) for determining amino-acid composition is shown in Table 6. In addition to taking account of possible interferences between indicator masses from different residues, this scheme also uses (crude) ion abun-

Table 5. Simple set of logical rules (Scheme A) for deducing the presence or absence of amino acid residues based on the number of indicator peaks present in the CID spectrum

AA	Masses	Require if ≥	Exclude if ≤
Simple*	immonium ion mass	1	0
His	82, 110, 121, 123, 136, 166	3	2
Leu /Ile	72, 86	2	0
Lys /Gln	70, 84, 101, 112, 129	3	1
Met	61, 104	2	0
Arg	59, 70, 87, 100, 112, 129	4	2
Trp	117, 130, 159, 170, 171	4	2
Tyr	91, 107, 136	2	0

*Includes Asp, Cys, Glu, Phe, Asn, Pro, Ser, Thr and Val.

Table 6. More sophisticated logic scheme (Scheme B) for determining the amino acid composition of a peptide from its CID spectrum[a]

AA		Test	Prediction
Simple*		Immonium ion mass strong or medium	+
		Immonium ion mass absent	−
		All other cases	0
Asn		87 strong or medium and Arg predicted absent	+
		87 absent	−
		All other cases	0
Pro		70 strong or medium and 87, 100, and 112 all absent	+
		70 absent	−
		All other cases	0
Ser		60 strong or medium and Arg predicted absent	+
		60 absent	−
		All other cases	0
His		110 strong	+
	or	110 medium and at least 2 of 82, 121, 123, 136, and 166 present	+
		110 absent	−
		All other cases	0
Arg		100 strong or medium, 87 and at least two of 59, 70, 112, or 129 present	+
		100 strong or medium, 87 and less than two of 59, 70, 112, or 129 present	0
	or	87 or 100 present and at least one of 59, 70, 112, or 129 present	0
		All other cases	−
Lys /Gln		84 and 101 strong and weak (either order) or both medium	+
	or	84, 101, 112, and 129 all present	+
		84 and 101 both absent	−
		All other cases	0
Trp		130 and 159 strong and weak (either order) or both medium and 171 present	+
		130 and 159 strong and weak (either order) or both medium and 171 absent	0
	or	130 and 159 both present	0
	or	171 present and either 130 or 159 present	0
		All other cases	−
Tyr		136 strong or moderate and at least one of 91 or 107 present	+
		136 absent	−
		All other cases	0

[a] The prediction "+" means that all candidate sequences must contain this residue, "−" means that none may contain it, and "0" means that the sequences may or may not contain it.

*The designation "simple" refers to residues for which only one indicator peak is used, namely their immonium ion. These are Cys, Asp, Glu, Phe (91 not used), Ile /Leu (72 not used), Met (61 not used), Thr, and Val.

dances, rather than simply determining if a peak is above threshold. In this scheme, peaks are defined as being strong, medium, or weak (see Experimental section). To be "present," a peak simply has to exceed the lowest threshold.

Figure 4 illustrates the results obtained when these two logic schemes were applied to the spectra in our database. Figure 4a displays the accuracy of predictions of the presence of each residue (required). Each bar represents the fraction of all spectra of peptides containing the given residue that were predicted to contain that residue. The shaded bars are for Scheme A, while the black bars are for Scheme B. Figure 4b shows the fraction of spectra of peptides *not* containing each residue, for which the logical rules gave a result of excluded. Figure 4c and d show the fraction of incorrect predictions, false negatives and false positives, respectively.

Although Scheme A gives a greater fraction of correct predictions than Scheme B, it also yields a greater fraction of incorrect predictions. This is a consequence of more stringent criteria for the classifications of required and excluded in Scheme B. The result is that a greater fraction of the overall predictions from Scheme B fall into the allowed category, and thus do not appear explicitly in Figure 4. The data shown in Figure 4 for Cys make it clear that an accurate prediction of the presence or absence of this residue is not possible. Similarly, it is not possible to accurately exclude Asp, Ser, or Glu, although their presence is predictable with

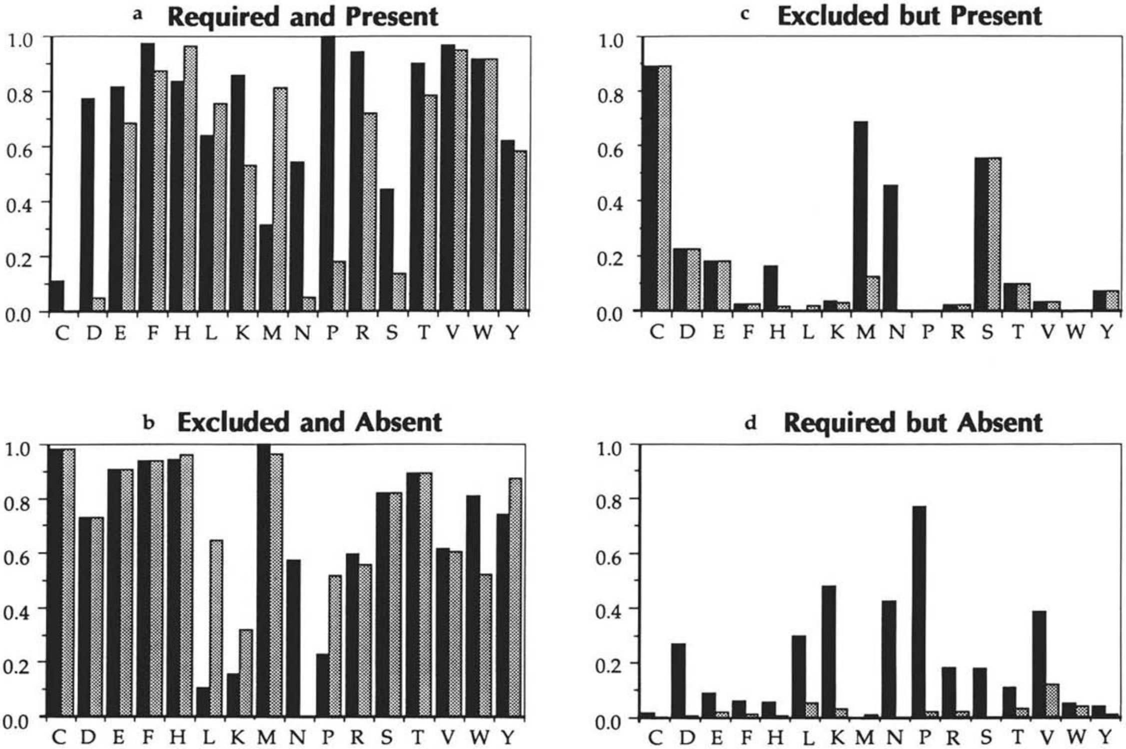

Figure 4. Results of application of Schemes A (black bars) and B (gray bars) to the database. (a) All spectra of peptides in the database containing the residue in question were examined. The bars show the fraction of these spectra that were predicted to contain the residue. (b) Fraction of spectra of peptides *not* containing each residue, for which the logical rules gave a result of "excluded" (case F−). (c) Fraction of false negative (T−) predictions. (d) Fraction of false positive (F+) predictions.

moderate reliability. Therefore, these predictions are never attempted in practice in the absence of additional information.

In developing logic to screen CID spectra for indicator ions, two key points must be borne in mind. First, it is essential to realize that it is not possible to determine all residues in every case with complete certainty. On the basis of our work with various possible logical schemes as well as considerable experience in manual interpretation of such spectra, we conclude that this statement is almost certain to remain true for the foreseeable future for the type of spectra in our database. This conclusion underlines the basic premise of our computer-aided interpretation software, namely, that the computer can be an extremely valuable assistant in the process, but human interaction is essential.

The second key point, which follows from the first, is that one is required to establish a balance between the desired level of accuracy and the fraction of cases for which a definite prediction can be made. At one extreme, perfect accuracy is required, and all or nearly all tests produce the result "allowed." At the other extreme, definite predictions (required or excluded) are demanded in all cases, but many are incorrect. Figure 5a and b illustrate the proportions of outcomes for the two schemes discussed above. In Scheme A, a greater fraction of definite predictions is attempted, with a concomitantly greater number of correct as well as incorrect results (16.5% incorrect). Overall, Scheme B is a better choice because the number of potentially misleading results is significantly reduced (to 4.8%), even though this comes at the price of a decreased fraction of definite predictions. In our laboratory, we use a variation of Scheme B in which Asp, Ser, and Glu are never excluded and no predictions are attempted for Cys. This reduces the percentage of incorrect results to 2.4%, but at the cost of further increasing the number of cases in which no prediction is made (Figure 5c).

Clearly, there are other possible variations of the logical schemes presented above, with which one could achieve any desired balance between comprehensiveness and accuracy. Additional secondary logical rules could be added, and nonstandard amino acids could

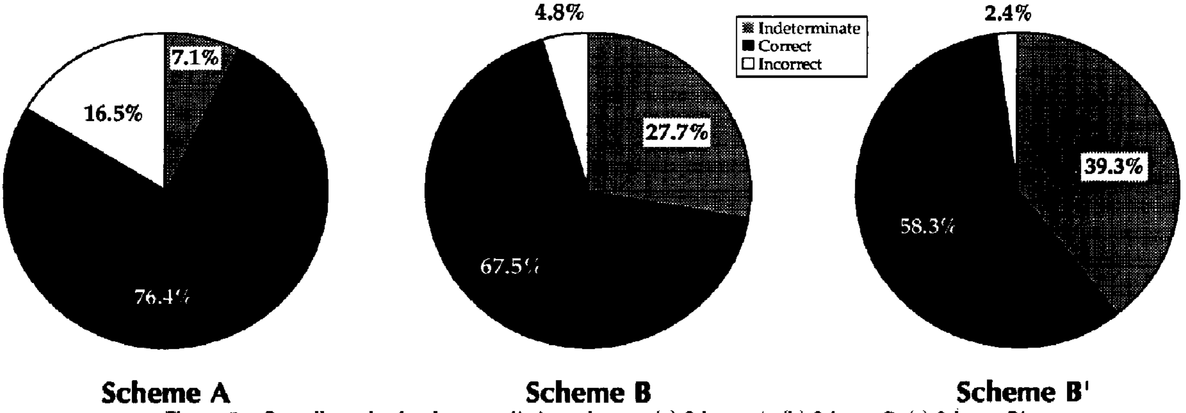

Figure 5. Overall results for three prediction schemes. (a) Scheme A, (b) Scheme B, (c) Scheme B' (same as Scheme B, but Asp, Ser and Glu are never excluded, and no predictions are attempted for Cys).

be added. Scheme B is currently used in our laboratory as shown (with a few omissions, as discussed above), but future improvements are likely as our experience continues to increase.

Conclusions

The low-mass region of high-energy, tandem CID spectra of peptides contains a wealth of information, mainly about the amino-acid composition of the peptides. These data definitively show the presence or absence of many amino acid residues, although the complete amino acid composition cannot be deduced in most cases. The information obtained is extremely valuable in generating or narrowing the choice of possible candidate sequences for the determination of the sequence of an unknown, either manually or by using a computer-aided scheme. Rules can be developed to provide computer-assisted evaluation of the amino-acid content of an unknown peptide from its CID spectrum. This process can be integrated into an overall computer-aided spectrum-interpretation scheme or simply used as an aid in manual interpretation. In either case, unknown peptide sequence determination is enhanced.

Acknowledgments

We are indebted to Fred C. Walls for acquiring most of the spectra used for this study. This work was supported by grants from the Tobacco Related Disease Program (to B .W. Gibson), the National Institutes of Health (RR01614) and the National Science Foundation (DIR 8700766) (both to A. L. Burlingame).

References

1. Biemann, K.; Scoble, H.A. *Science* **1987**, *237*, 992–998.
2. Biemann, K.; Martin, S. A. *Mass Spectrom. Rev.* **1987**, *6*, 1–76.
3. Ashcroft, A. E.; Derrick, P. J. In: *Mass Spectrometry of Peptides*; Desiderio, D. M., Ed.; CRC: Boca Raton, FL, 1991; pp 121–138.
4. Martin, S. A.; Johnson, R. S.; Costello, C. E.; Biemann, K. In: *The Analysis of Peptides and Proteins by Mass Spectrometry*; McNeal, C. J., Ed.; Wiley: New York, 1989; pp 135–150.
5. Biemann, K. In: *Methods in Enzymology*, Vol. 193; McCloskey, J. A., Ed.; Academic: San Diego, 1990; pp 455–479.
6. Stults, J. T. In: *Biomedical Applications of Mass Spectrometry*; Suelter, C. H.; Watson, J. T., Eds.; Wiley: New York, 1990; pp 145–201.
7. Johnson, R. S.; Martin, S. A.; Biemann, K. *Int. J. Mass Spectrom. Ion Processes* **1988**, *86*, 137–154.
8. Hines, W. M.; Falick, A. M.; Burlingame, A. L.; Gibson, B. W. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 326–336.
9. Zidarov, D.; Thibault, P.; Evans, M. J.; Bertrand, M. J. *Biomed. Environ. Mass Spectrom.* **1990**, *19*, 13–26.
10. Burlingame, A. L.; Millington, D. S.; Norwood, D. L.; Russell, D. H. *Anal. Chem.* **1990**, *62*, 268R–303R.
11. Johnson, R.; Biemann, K. *Biomed. Environ. Mass Spectrom.* **1989**, *18*, 945–957.
12. Roepstorff, P.; Fohlman, J. *Biomed. Mass Spectrom.* **1984**, *11*, 601.
13. Biemann, K. In: *Methods in Enzymology*, Vol. 193; McCloskey, J. A., Ed.; Academic: San Diego, 1990; pp 886–887.
14. Barber, M.; Bordoli, R. S.; Garner, G. V.; Gordon, D. B.; Sedgwick, R. D.; Tetler, L. W.; Tyler, A. N. *Biochem. J.* **1981**, *197*, 401–404.
15. Barber, M., Bordoli, R. S.; Sedgwick, R. D.; Tetler, L. W. *Org. Mass Spectrom.* **1981**, *16*, 256–260.
16. Renner, D.; Spiteller, G. *Biomed. Environ. Mass Spectrom.* **1986**, *13*, 405–410.
17. Kausler, W.; Schneider, K.; Spiteller, G. *Biomed. Environ. Mass Spectrom.* **1988**, *17*, 15–19.
18. Lippstreu-Fisher, D. L.; Gross, M. L. *Anal. Chem.* **1985**, *57*, 1174–1180.
19. Madden, T.; Welham, K. J.; Baldwin, M. A. *Org. Mass Spectrom.* **1991**, *26*, 443–446.
20. Thornburg, K. R.; Schey, K. L.; Knapp, D. R. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 424–427.
21. Walls, F. C.; Baldwin, M. A.; Falick, A. M.; Gibson, B. W.; Kaur, S.; Maltby, D. A.; Gillece-Castro, B. L.; Medzihradszky, K. F.; Evans, S.; Burlingame, A. L. In: *Biological Mass Spec-*

trometry; Burlingame, A. L.; McCloskey, J.A., Eds.; Elsevier: Amsterdam, 1990; pp 129–146.

22. Aberth, W. A.; Straub, K. M.; Burlingame, A. L. *Anal. Chem.* **1982,** *54*, 2029–2034.
23. Falick, A. M.; Wang, G. H.; Walls, F. C. *Anal. Chem.* **1986,** *58*, 1308–1311.
24. Cottrell, J. S.; Evans, S. *Anal. Chem.* **1987,** *59*, 1990–1995.
25. Boyd, R. K. *Int. J. Mass Spectrom. Ion Phys.* **1987,** *75*, 243–264.
26. Falick, A. M.; Medzihradszky, K.; Walls, F. C. *Rapid Commun. Mass Spectrom.* **1990, 4,** 318–322.
27. Medzihradszky, K. F.; Hall, S. C.; Maltby, D. A.; Hines, W. M.; Burlingame, A. L. In: *Techniques in Protein Chemistry II*; Villafranca, J. J., Ed.; Academic: San Diego, 1991; pp 435–440.
28. Medzihradszky, K. F.; Gibson, B. W.; Kaur, S.; Yu, Z.; Medzihradszky, D.; Burlingame, A. L.; Bass, N. M. *Eur. J. Biochem.* **1992,** *203*, 327–339.
29. Falick, A. M.; Maltby, D. A. *Anal. Biochem.* **1989,** *182*, 165–169.
30. Lee, T. D.; Shively, J. E. In: *Methods in Enzymology*, Vol. 193; McCloskey, J. A., Ed.; Academic: San Diego, 1990; pp. 361–374.
31. Gibson, B. W. In: *Biological Mass Spectrometry*; Burlingame, A. L.; McCloskey, J. A., Eds.; Elsevier: Amsterdam, 1990; pp 315–336.
32. Gibson, B. W.; Cohen, P. In: *Methods in Enzymology*, Vol. 193; McCloskey, J. A., Eds.; Academic: San Diego, 1990; pp 480–501.