Sequence Pattern Correlation of Amino Acid in Collision-induced Dissociation Electrospray Ionization Mass Spectrometry

SONG, Hao-Wei[a,b](宋浩威) YUE, Gui-Hua[a,b](岳贵花) LU, Yu[a](陆宇)
YANG, Peng-Yuan*[,a,b](杨芃原) WANG, Hong-Hai*[,b](王洪海)

[a] *Department of Chemistry, Fudan University, Shanghai 200433, China*

[b] *State Key Laboratory of Genetic Engineering, Fudan University, Shanghai 200433, China*

A novel approach of sequence pattern correlation has been applied to predict an expected amino acid sequence from CID ESI-MS spectra. The proposed approach deduces sequence patterns with no help from known protein database such that it is useful to identify an unknown peptide or new protein. The algorithm applies a cross-correlation to match an experimental CID spectrum with predicted sequence pattern generated from fragmentation information. The fragmentation knowledge of both y-series and other non y-series are utilized to generate the predicted sequence patterns. In contrast to the normal *de novo* approach, the proposed approach is insensitive to mass tolerance and non-susceptive to spectral integrality with no need for selection of a starting point.

Keywords mass spectrometry, sequence pattern correlation, amino acid sequence, protein and peptide

Introduction

Sequence determination for peptides or proteins by means of mass spectrometry has been an important research area in recent years.[1] The partial sequences can be obtained using techniques of the orifice-collision induced dissociation(CID)-electrospray ionization-mass spectrometry (CID-ESI-MS) and of the collision activated dissociation(CAD)-tandem MS (CAD-MS/MS). The database searching for a sequence pattern matched with CID/CAD spectra has attracted more attention now[2] in addition to that based on the peptide mapping spectra.[3]

It has been well understood for a known peptide or protein that the library search would be more accurate[4] if a partial amino acid sequence of one or two peptide(s) can be determined for a proteolysis protein. A partial sequence determines a string of residues with a peptide sequence tag containing information of both molecular weight *MW* and specific fragmentation. This interpretation is particularly useful in the identification of a peptide or a protein existing in a database.[4-9] It should be noticed that such a database searching could be a tedious job for simulation of every possible spectrum. In addition, the correlation technique has no way to generate a sequence pattern for an unknown protein.[10,11]

For the sequence of an unknown or new protein from a CID spectrum, *de novo* method has to be used. The *de novo* algorithm is a direct and effective method to interpret the CID mass spectrum.[6] In comparison to the conventional technique of chemical protein sequencing,[12] it is rather effective to apply a *de novo* approach[13] to obtain a partial sequence for a known protein to increase the search accuracy. However, the interpretation of CID/CAD mass spectra is often troublesome especially for an unknown peptide or protein.[5] In addition, the *de novo* method requires a good starting point and a good spectrum.[14] Recently, the *de novo* interpretation for ^{18}O labeling CID spectrum[15] and for electron-capture dissociation spectrum[16] has been also reported with better results than the normal CID technique.

In this work, an approach of sequence pattern corre-

* E-mail: pyyang@fudan.edu.cn and hhwang@fudan.edu.cn; Tel: 86-021-65642009; Fax: 86-021-65641740

Received January 10, 2001; revised October 10, 2001; accepted November 20, 2001.

Project supported by the National Natural Science Foundation of China (No. H29927002) and the Visiting Scholar Foundation of Key Laboratory in Universities of China.

lation (SPC) has been proposed and studied for the amino acid sequence analysis with orifice-CID spectra. The theoretical aspects of the new approach are described and the corresponding algorithm is discussed in detail. Some examples are given to illustrate features of the proposed approach.

Experimental

Instrument and reagents

An ESI-MS instrument (API-165, Perkin-Elmer, USA) was used in this study, equipped with HPLC. Mass spectrometer was calibrated with standard polypropylene glycol (PPG). The calibrated API-165 instrument has a mass accuracy within + 0.5 amu, with a general mass resolution of ~ 3000 for the averaged mass of m/z 300—1800 studied in this work. An IBM-586 compatible PC computer is used for all developed programs for the SPC algorithm written in C^{++} programming language on the window operating system.

Sequence measurement

The proteolysis of cytochrome-c or bovine serum albumin was done with the normal digestion procedure by trypsin. Commercially available cytochrome-c (Sigma), lysozyme, bovine serum albumin (Sigma) and trypsin (Sigma) were used. Separation of peptides after proteolysis was carried out on a Zobax C_{18}(2.1 × 150 mm) reverse phase column. The CID experiment was performed at high orifice voltage (OR) of 110 V.

Sequence pattern correlation

The proposed algorithm of SPC can be briefly described as follows. The experimental mass spectrum is defined as X_i ($i = 1, 2, \cdots n$), and the guessed sequence pattern in an estimated order is defined as Y_j ($j = 1, 2, \cdots n$). To make a clear pattern calculation, intensities for all lines are set to unity for Y_j pattern. The cross-correlation coefficient of C_{XY} can then be given by:

$$C_{XY} = \max(C_{XY,l}) = \max(\sum_{i=l}^{n} X_i \times Y_{i+l}), \quad (1)$$
$$l = -n,\ -n+1,\ \cdots n-1,\ n$$

where l is the displacement between the experimental spectrum and the guessed sequence pattern. In this case, a score at lth data point can simply be defined as 1 when the product of corresponding X × Y is greater than 0.

A y-score and a sequence-score are defined to evaluate the predicted sequence pattern from an experimental mass spectrum. The y-score is applied for a sequence pattern matched only in y-series. Once a y-series match is obtained, the sequence-score is utilized to verify the existence of all fragment ions related to this y-series. Only those patterns with both high y- and sequence-scores are reserved as candidates of the predicted sequence pattern.

The SPC procedure is illustrated in Fig. 1. A guessed sequence pattern in a y-series is generated by the algorithm and is applied to correlate with an experimental mass spectrum for y-series and also for other series.[17] If a matching is achieved, this guessed sequence pattern is considered as a potential candidate as a predicted sequence pattern. As shown in Fig. 1, the guessed sequence pattern is started from one-residual pattern and will be extended to two-residual patterns and more, until no longer-sequence pattern is found. When the number of residues h is less than or equal to three, the candidates of matched pattern are only selected for those y-score = $h + 1$ (peak number in an h-residual pattern). When the number of residue is above three, the candidates are selected for those y-score $\geqslant h$ such that one missing line is possibly allowed. This enlarged criterion would save those potential candidates with only a missing line in a y-series pattern.

The sequence-score S_{seq} is calculated as follows. Related fragment ions other than y-series include ions in x-, z-, a-, b-, c-series, and ions formed from those fragment clusters by losing an amine, a carbon monoxide or a H_2O molecule. Only fragment types of y-17, b-, b-17, a- and a-17 are considered because other type barely exist in the CID spectrum[18] for a low energy CID. With an MW information of m for parent ion $[M+H]^+$, and m_y for the matched pattern in the y-series, the mass of those related fragment ions can be calculated with Eq. (2)—(6).[18]

$$m_{y-17} = m_y - 17 \quad (2)$$
$$m_b = m + 1 - m_y \quad (3)$$
$$m_{b-17} = m + 1 - 17 - m_y = m - m_y - 16 \quad (4)$$
$$m_{a-17} = m + 1 - 28 - 17 - m_y = m - m_y - 44 \quad (5)$$

$$m_a = m + 1 - 28 - m_y = m - m_y - 27 \quad (6)$$

Thus, the sequence-score is calculated with Eq. (7).

$$S_{seq} = \sum_{i=1} s_i = \sum_{i=1} \sum_{j=1}^{6} s_{ij}^{R} \quad (7)$$

where s_i is the sequence-score for ith residue in the found pattern in y-series, and s_{ij}^{R} is the sequence-score for all matched residues in above six series for this ith residue. Empirically, the s_{ij}^{R} is weighted to be 0.08, 0.04, 0.08, 0.04, 0.02, and 0.01, for y-, y-17, b-, b-17, a- and a-17 series, respectively. The consideration of y-series in the calculation of s_{ij}^{R} is accounted for a possible residue lost in y-series in the experimental spectrum. In this case, this lost residue will contribute zero sequence-score for the lost y-ion in the sum score of s_i. Also, to use the parent-ion m/z information of $[M+H]^+$, another score of 0.088 for $[M+H]^+$ is added to the total score. Only the top 100 candidates in the score sorting-list are saved in a next iteration.

Formation procedure for predicted sequence pattern

An example of predicted pattern with only three residues (N, T, and M) is illustrated in Fig. 2. A two-peak sequence pattern of Y_k corresponding to one residue is correlated first with the CID spectrum. Any peaks matched in SPC process are supposed as potential peaks of y-type fragment ion. In this example, if the potential candidate for one-residual pattern is N, or T, or M, it can be presented with a two-peak pattern of (k, Y_k). Thus, only those Y_{i+l} with an SPC score of 2 (two-peak matching) at lth data point are marked for later searching.

Once all mono amino acid residues are found, it would then be able to extend the pattern from one residue to two or more residues with a similar process as shown in Fig. 1. For example, the possible sequence patterns with two residues are NN, NT, NM, TT, TN, TM, MN, MT and MM for a three-peak pattern. These patterns are tested with the SPC, and only the pattern with a high score of 3 is reserved according to Eq. (1). Then a permutation of these residual is checked for their validity in existence of related fragment ions other than y-series and the sequence-score is calculated. In this example, the new predicted sequence pattern for NT would be NTN, NTT and NTM (Fig. 2).

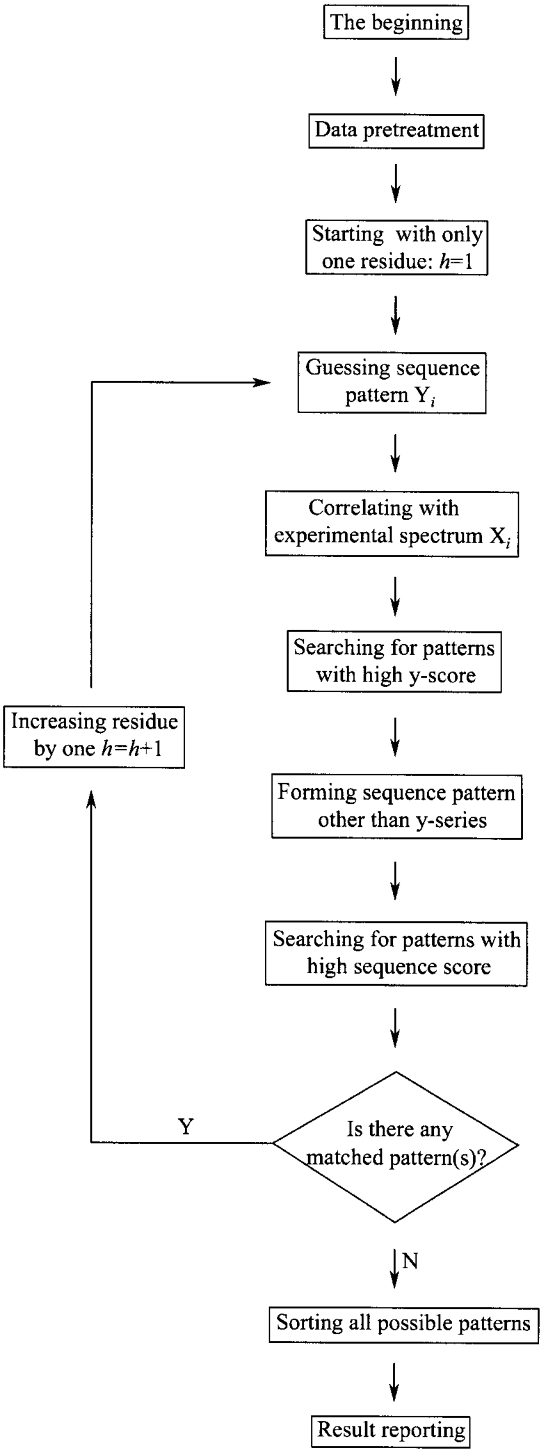
The beginning
Data pretreatment
Starting with only one residue: h=1
Guessing sequence pattern Y_i
Correlating with experimental spectrum X_i
Searching for patterns with high y-score
Forming sequence pattern other than y-series
Searching for patterns with high sequence score
Is there any matched pattern(s)?
Y
Increasing residue by one h=h+1
N
Sorting all possible patterns
Result reporting

Fig. 1 Flow chart of the SPC algorithm.

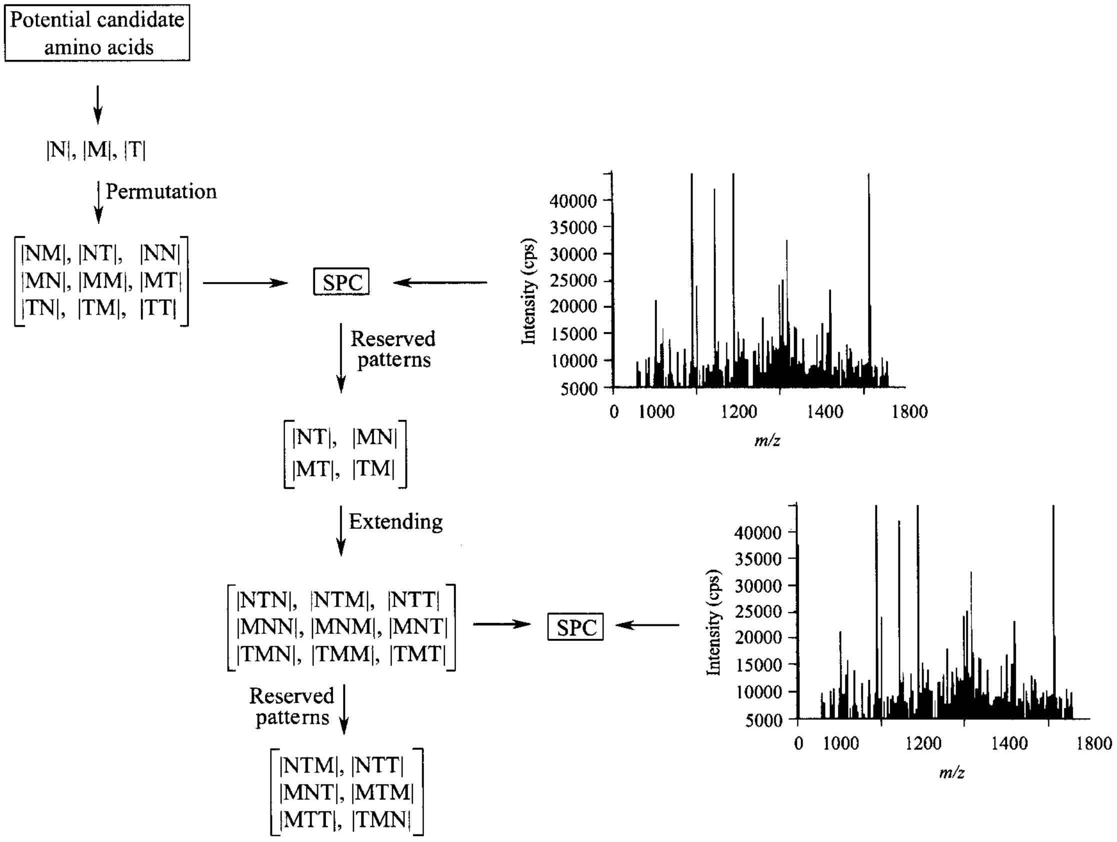

Fig. 2 Example of pattern formation in SPC process. Only three presumed residues of N, M, and T are considered and patterns for three residues are shown.

Results and discussion

Concrete examples of SPC *algorithm*

Fig. 3 shows the CID spectrum of a peptide of cytochrome-c, with a predicted sequence pattern of (G, I) TWGEETLMEYLE(N, P)K in comparison with the expected one of GITWGEETLMEYLENPK, and two short sequences of (G, I) and (N, P) are guessed ones with less confidence. Three main problems have been found in this sequence searching process, and cause a difficulty to predict a pattern by the traditional *de novo* method. First, there is one mass unit difference between experimental peak (*m/z* at 2011 for M + H ion) and expected peak (*m/z* at 2010) for the parent ion (2009 Da), due to an instrumental calibration error. The second and the third problems are unfortunately arisen from the absence of peaks for the second y_{15} ion (*m/z* at 1952.9) and for the two sequence ions (*m/z* at 147.1 and 244.2) in low-mass end, respectively.

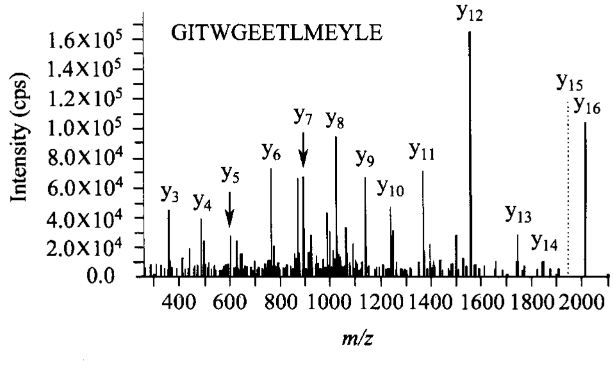

Fig. 3 CID spectrum of a peptide of cytochrome-c with *m/z* of 2011, acquired at an OR voltage of 110 V, and ESI voltage of 5000 V; (G, I)TWGEETLMEYLE(N, P)K, predicted amino acid sequence, (G, I) and (N, P), guessed patterns with less confidence. The peak labels are residual sequence in y-series. The peak for y_{15} is a presumptive one.

A comparison has been made as listed in Table 1, between SPC in this study and the software provided with the PE API-165 instrument for CID sequence analysis.

Table 1 Comparison between SPC and other method for CID sequence analysis

Peptide	Expected sequence	Predicted sequence[a]	Rank[a]	Predicted sequence[b]	Rank[b]
Brady	RPPGFSPFR	RPPGFSPFR[c]	1	RPPTC SPFR	1
Cyt-2011	GITWGEETLMEYLENPK	GITWGEETLMEYLE	1	No match	
Cyt-1264	YPDYSKPVPAK	YPDYSKP	2	No match	
Al-1640	KVPQVSTPTLVEVSR	KVPQVSTPTLV	1	Q VPQVSTSQ* *	5
Al-1511	VPQVSTPTLVEVSR	VPQVST	1	VPQVST* * * *	1
Ly-874	HGLDNYR	HGLDNYR	1	No match	

[a] By this work. [b] By software in API-165 instrument. [c] Matched sequence.

For the commercial instrument, the most probable but wrong pattern was given as CSQQTGFCLCGDNQLVK with a score of 0.431. In contrast, a partial sequence of TWGEETLMEYLE can still be derived from the spectrum directly by the SPC algorithm with a high score of 1.48. In addition, a simple algorithm is used to estimate the partial sequence of (G, I) to fit with the parent ion peak, and to guess the sequence of (N, P) in low-mass side. Because this algorithm can also increase the complexity of calculation, a caution must be taken for possibly incorrect results.

Figs. 4(a—c) show CID spectra of a peptide (1264 Da) for alkaline lipase, of a peptide (1640 Da) for bovine serum albumin, and of a peptide (1305 Da) for bovine serum albumin, respectively. With the similar SPC process as mentioned above, the predicted sequence of PKSYDPY can be found to fit the partial expected sequence of KAPVPKSYDPY [Fig. 4(a)]. Also, the predicted sequence of VLTPTSVQPVK can be closely matched with the partial expected sequence of RSVEVLTPTSVQPVK [Fig. 4 (b)], and that of EDVLH can be partially approached to the expected sequence of KILNQPEDVLH [Fig. 4(c)], respectively. Because the interpretation of CID mass spectra for unknown peptide is obviously more complicated,[19] the SPC approach certainly opens an avenue to solve this problem.

Comparison of different scoring method

In addition to S_{seq} for fragment ions other than y-series, the intensity score S_{int}, and the combination-score S_{total} of sequence- and intensity-score is also evaluated:

$$S_{total} = S_{seq} + S_{int} \tag{8}$$

To calculate properly an intensity score, the whole mass range is divided into several 100 amu intervals to account for the intensity variation. Only peaks with higher than 5% relative-abundance in each interval is labeled. If a

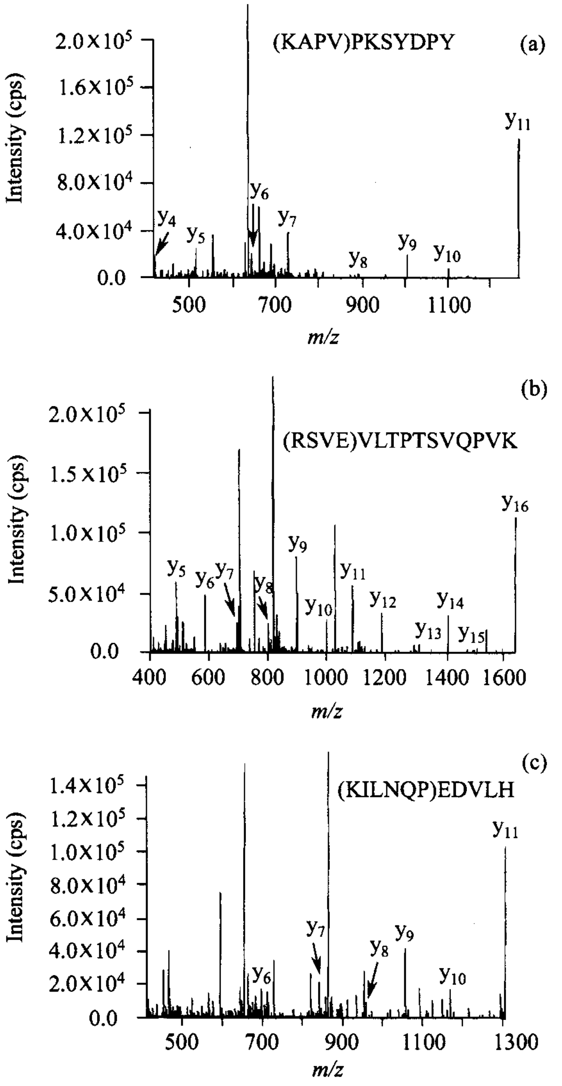

Fig. 4 CID spectra of (a) a peptide of alkaline lipase (*m/z* at 1264), (b) a peptide of bovine serum albumin (*m/z* at 1640), and (c) a peptide of bovine serum albumin (*m/z* at 1305), acquired all at an OR voltage of 110 V. (a), PKSYDPY, the predicted sequence; KAPVPKSYDPY, the theoretical amino acid sequence. (b), VLTPTSVQ-PVK, the predicted sequence; RSVEVLTPTSVQPVK, the theoretical sequence. (c), EDVLH, the predicted sequence; KILNQPEDVLH is the theoretical sequence. See text for detail.

labeled peak is matched in the SPC process, an empirical intensity-score of 0.19 is then added into the total score.

Fig. 5a is the CID spectrum of a peptide (Brady) displaying all fragment ladders with high *SNR*. A good SPC result for the predicted sequence of RFPSFGPPR can be obtained with the S_{int} only. Fig. 5b is the CID spectrum for a peptide (m/z of 1511) from bovine serum albumin and illustrates lost peaks for related b-series at low-mass region and intense peaks for y-series at the high-mass region. In this spectrum the b_1(100.1), b_2(197.1), b_3 (424.3), b_4(511.3) are absent such that the sequence scores for y_{13}, y_{12}, y_{11} and y_{10} are low. Thus, the predicted sequence of VPQVST has ranked as the 120th. In fact, the y_{13}, y_{12}, y_{11} and y_{10} peaks are intense in the high-mass region so that the use of intensity-score will be suitable. With the contribution of S_{int} to S_{total} a high score of 0.19 is obtained for the sequence of VPQVST rank No. 1.

It is believed that the intensity-score is useful only if all the fragment ions have been recorded with fairly good *SNR*. Thus, the intensity-score can enhance the accuracy of identification when pattern peaks in the low-mass region are absent or missing. However it is noticed that for most of spectra tested, the contribution from S_{int} is only roughly 10% in contrast to that about 90% from S_{seq}.

Direction of pattern extension

It is found that the direction of pattern extension is better to be developed from Y_j (in a higher-mass region) to Y_{j-1}(in a lower-mass region) than that from Y_j to Y_{j+1} due partially to the noisy signal in the low-mass region. A CID spectrum for a peptide with an m/z of 724.8 is shown in Fig. 5c. The expected sequence of CCAADDK is the top one if the pattern extension is carried out from high to low mass. In contrast, if the pattern extension is carried out from a low to high mass, the pattern of CC for two y-fragment ions of y_2 and y_3 would have a relatively lower score and would be removed from a next sequence searching.

Mass tolerance in SPC

Two mass tolerance parameters are applied in the SPC process, defined as E_1 for y-score and E_2 for sequence - score , respectively . The E_1 is used to ensure a pattern-accuracy in the SPC process. Because the quadruple mass spectrometer has a resolving power about ±0.5 mass unit, the E_1 is set as ±0.5 accordingly. It has been verified that this fairly reasonable tolerance is capable to extract any potential sequence patterns matched with a CID spectrum.

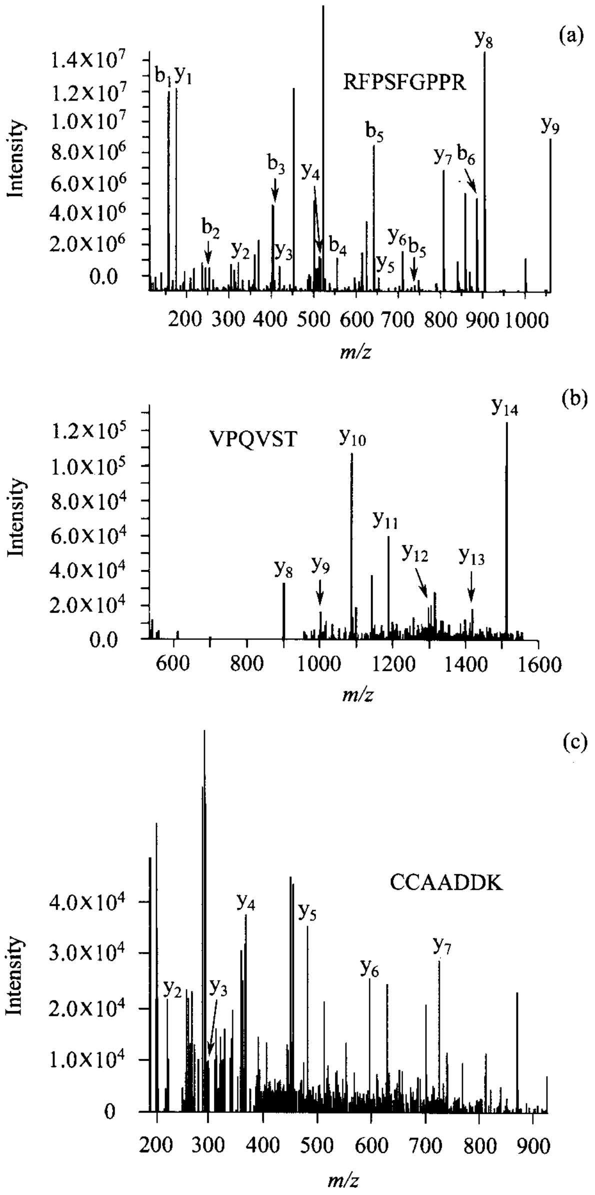

Fig. 5 CID spectra of (a) a peptide named as Brady (copied from the example library in PE Sciex 165 software), (b) a peptide of bovine serum albumin (m/z at 1511), (c) a petide (m/z at 724.8).

The E_2 has been applied for calculation of the sequence-score in locating related fragment ions other than y-series. In contrast to E_1, E_2 is utilized in a relatively short mass region as indicated in Eqs. (2)—(7). Thus, it is reasonably required that E_2 should be more precise and smaller than E_1 in order to enhance the accuracy. A value between 0.1—0.2 was found to be suitable under experimental conditions.

Removal of noise-peak

Results listed in Table 2 demonstrate that the removal of noisy peak is important to identify a reliability sequence of a peptide. Noisy peaks can be found when the background level for noisy peak variation is defined first, and then is subtracted from the experimental spectrum, in a similar way to remove all but the 200 most abundant-ions and to re-normalize the intensities of remaining ions to 100.[6] The resulting and simplified peaks are then used to search for potential sequence patterns. It is noticed in Table 2 that a number of predicted patterns are ranked as top first or second one compared with that of raw data set.

Conclusion

A novel SPC approach has been demonstrated to predict an expected sequence pattern from CID ESI-MS spectra. It is concluded that the new approach is insensitive to the mass tolerance of a spectrometer and also unsusceptible to spectral complexity. It can be carried out without help from a database and other assistant information such as the starting point or missing peaks. Future work will focus on the accuracy of the proposed approach by optimizing algorithm and by accumulating additional cleavage information.

Table 2 Results of SPC with and without the noise filtering

Peptide	Expected sequence	Predicted sequence[a]	Rank[a]	Predicted sequence[b]	Rank[b]
Brady	RPPGFSPFR	RPPGFSPFR	1	RPPGFSPFR	12
Cyt-2011	GITWGEETLMEYLENPK	TWGEETLMEYLE	1	TWGEETLMEYLE(PP)	20
Cyt-1634	IFVQKCAQCHTVEK	IFVQKCAQCHT	2	/[c]	
Al-1640	K VPQVSTPTLVEVSR	K VPQVSTPTLV	1	/[c]	
Al-1511	VPQVSTPTLVEVSR	VPQVST	1	VPQVS(RES)	67
Ly-874	HGLDNYR	HGLDNYR	1	HGL(TE)YR	43

[a] The sequence is obtained by SPC with a pretreatment of background noise level. [b] Raw data set. [c] The expected sequence is missed.

References

1 Roth, K. D. W.; Huang, Z. H.; Sadagopan, N.; Watson, J. T. *Mass Spectrom. Rev.* **1998**, *17*, 255.

2 Yates III, J. R. *Electrophores* **1998**, *19*, 893.

3 Jenson, O. N.; Podtelejnikov, A. V.; Mann, M. *Anal. Chem.* **1997**, *69*, 4741.

4 Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390.

5 Yates III, J. R.; McCormark, A. L.; Eng, J. K. *Anal. Chem.* **1996**, *68*, 534A.

6 Eng, J. K.; McCormack A. L.; Yates, III, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, 976.

7 Yates III, J. R.; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426.

8 Yates III, J. R.; Eng, J. K.; McCormack, A. L. *Anal. Chem.* **1995**, *67*, 3202.

9 Yates III, J. R.; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998**, *70*, 3557.

10 Owens, K. *Appl. Spectrosc. Rev.* **1992**, *27*, 1.

11 Lorenz, S. A.; Maziarz III, E. P.; Wood, T. D. *Appl. Spectrosc.* **1999**, *53*, 18A.

12 Henry, C. *Anal. Chem. News Feat* **1998**, *70*, 401A.

13 Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067.

14 Fernandez-de-Cossio, J.; Gonzalez, J.; Betancourt, L.; Besada, V.; Padron, G..; Shimonishi, Y.; Takao, T. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 1867.

15 Qin, J.; Herring, C. J.; Zhang, X. L. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 209.

16 Axelsson, J.; Palmblad, M.; Hakansson, K.; Hakansson, P. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 474.

17 Taylor, J. A.; Walsh, K. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 679.

18 Parayannopoulos, I. A. *Mass Spectrom. Rev.* **1995**, *14*, 49.

19 Yates III, J. R.; McCormack, A. L.; Link, A. J.; Schieltz, D.; Eng J.; Hays, L. *Analyst* **1996**, *121*, 65R.

(E0101101FZ FAN, Y. Y.)