ELSEVIER

Journal of Chromatography A, 849 (1999) 71–85

JOURNAL OF CHROMATOGRAPHY A

Windowed mass selection method: a new data processing algorithm for liquid chromatography–mass spectrometry data

Cliona M. Fleming[a,*], Bruce R. Kowalski[a], Alex Apffel[b], William S. Hancock[b]

[a]*Laboratory for Chemometrics, Department of Chemistry, Box 351700, University of Washington, Seattle, WA 98195, USA*
[b]*Biomeasurements Group, Hewlett Packard Laboratories, 3500 Deer Creek Road, Palo Alto, CA 94304, USA*

Received 3 November 1998; received in revised form 9 March 1999; accepted 20 April 1999

Abstract

A number of preprocessing methods are tested on liquid chromatography–mass spectrometry (LC–MS) peptide map data, to determine the best and most efficient way to improve the signal to noise ratio in the data, especially at low analyte concentrations. Three methods are investigated, including an algorithm named ''sequential paired covariance'' (SPC), which was recently reported. An improvement to this algorithm is also reported here. This new, improved method, named the ''windowed mass selection method'' (WMSM), is shown to effectively eliminate random noise that occurs in the data. This method is shown to be particularly useful in improving signal to noise ratios in both chromatographic and mass spectra for data acquired in peptide mapping of recombinant DNA derived proteins.

Keywords: Windowed mass selection method; Data processing algorithm; Chemometrics; Noise reduction; Mass spectrometry; Peptides

1. Introduction

In recent years, advances in biotechnology have resulted in the development of recombinant DNA derived proteins as biopharmaceutical drugs. The analytical challenges associated with these biomolecules are greater than with traditional, synthesized organic pharmaceuticals, because of their large size and complexity. One analysis method that is commonly used in the characterization of proteins is peptide mapping [1,2]. In this analysis method, a proteolytic digest of the protein is carried out, and the resulting mixture of peptide fragments is analyzed by a separation technique such as high-performance liquid chromatography (HPLC) or capillary electrophoresis, or a combination of separation and mass spectrometry, such as liquid chromatography/mass spectrometry (LC–MS) [3]. The amount of protein available for analysis is often very small, placing demands on the instrument sensitivity in addition to the data analysis methods used. Currently, the most popular interface between the HPLC and the mass spectrometer is electrospray ionization (ESI) [4,5]. Traditionally, peptide mapping was carried out using reversed-phase high-performance liquid chromatography (RP-HPLC) with a single channel UV detector, but the use of a mass spectrometer as the detector has become more common since the advent of electrospray ionization (ESI). Electrospray ionization is a relatively soft ionization

*Corresponding author. Fax: +1-206-543-6506.

0021-9673/99/$ – see front matter
PII: S0021-9673(99)00553-1

method, so that fragmentation of analyte molecules can be minimized, and the formation of multiply-charged ions means that large molecular weight peptides can be detected by a quadrupole mass spectrometer [3].

Generating peptide maps with LC–MS presents a number of difficulties that are not an issue with RP-HPLC peptide mapping [6]. There are two types of noise that can present themselves: high background (low frequency) noise and random (high frequency) noise. The high background is generally due to mobile phase components that have a signal over the entire run, while the random noise is due to the electrospray ionization interface.

During the generation of a peptide map, the quadrupole scans at a fixed rate, for example, 1 scan s^{-1}. Because of this high speed, each m/z value is 'seen' by the detector for just a fraction of a millisecond. Efficient ionization is therefore essential in order to ensure constant transmission of eluting analyte ions to the mass spectrometer. Inefficient ionization is a common problem that is often caused by suboptimal flow-rates and/or capillary voltages, and which results in low ion transmission and low signal to noise ratios. In order to explain the reason for this, the electrospray mechanism must be described.

Fig. 1 shows a diagram of an electrospray source. During an LC/MS run, the effluent is split as it exits the chromatographic column, and the smaller portion enters the electrospray through a grounded capillary. An applied potential difference between the capillary and a counter-electrode pulls positive charge toward the liquid front. Repulsive forces between the positive charges overcome the surface tension of the liquid, and small droplets leave the surface and travel toward the counterelectrode.

When the surface tension, flow-rate and electrolyte concentrations are low, the formation of the droplets proceeds without any problems. An increase in any of these parameters requires an increase in the electric field so that droplets can continue to be released from the liquid. This is because too low an electric field produces a low analyte signal. However, too high an electric field causes an electric discharge, resulting in an unstable, noisy spray. Finding the optimal electric field can be difficult.

Because of these difficulties with unassisted electrospray, a number of modifications have been implemented. These include pneumatically assisted electrospray, which consists of an inert gas flow that sweeps evaporated solvent molecules out of the interface thereby reducing noise from the LC solvents. When pneumatically assisted electrospray is used, higher flow-rates and aqueous solutions may be used without the need for high electric fields, and so

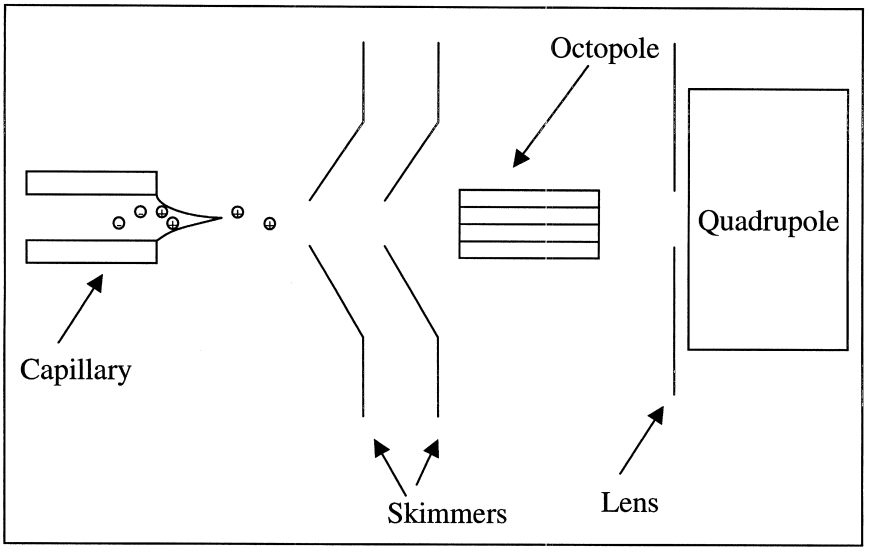

Fig. 1. General schematic diagram of an electrospray ionization source.

electric discharges are reduced. In addition, in some instruments the sprayer is positioned diagonally or at right angles to the source. In this case, operation stability is improved, and the transfer of contaminants into the mass spectrometer is reduced.

Despite these improvements to the original electrospray source design, ESI/MS data is still relatively noisy, as the sources of noise mentioned have been reduced, but not eliminated. In addition, transmission of ions to the mass spectrometer is a factor to be considered. The ions must travel through an atmospheric pressure-to-vacuum interface, and through ion optics, which focus the ions and transfer them to the mass analyzer. The ion optic settings are crucial to the efficiency with which ions are transmitted, and also to mass accuracy. Sub-optimal settings can also result in unwanted fragmentation, and the appearance of unexpected ions.

Our work involves the application of novel chemometric methods to LC–MS peptide maps that extract both qualitative and quantitative information for all components in a sample. However, due to the poor signal-to-noise (S/N) ratio in the data, our attempts have been unsuccessful. We therefore turned to available data preprocessing methods to clean up the data while preserving any quantitative information present.

This paper compares different data preprocessing methods that attempt to reduce the effect of noise on the data, thereby simplifying interpretation. The most effective methods should remove noise without affecting the analyte signal, and this should be achieved in a short amount of time. Following the comparison study, a new preprocessing method is described and shown to effectively separate signal from noise. In this case, the method is applied to peptide map data, allowing rapid identification of peptide fragments.

2. Experimental

2.1. HPLC

Endoproteinase LysC digests of recombinant tissue plasminogen activator (rt-PA) were analyzed. The chromatography was carried out on a Hewlett-Packard (Palo Alto, CA, USA) HP 1100 HPLC equipped with a binary solvent delivery system, automated injection system, heated column compartment and diode-array detector (DAD). Separations were carried out using a HP 250×2.1 mm I.D. 3 μm 300SB-C_{18} column (Hewlett Packard, Wilmington, DE, USA). Mobile phase A consisted of 0.1% trifluoracetic acid (TFA) in water, and solvent B contained 0.09% TFA in acetonitrile. A linear solvent gradient of 0–43.4% B over a 65-min period was used, at a flow-rate of 0.2 ml min^{-1}. The column was thermostated at 45°C.

In order to counteract the signal suppressing effects of TFA on the mass spectrometry, a ''TFA Fix'' [7,8] was used, consisting of a post-column addition of 50% acetic acid in water at a flow-rate of 100 μl min^{-1}. The TFA fix was delivered using a HP 1050 HPLC pump, and was teed into the column effluent after the DAD detector. The tee was connected to the electrospray needle via 0.005-inch I.D. peek tubing. The column effluent was diverted from the MS for the first 5 min of the chromatogram, during which time excess reagents and unretained components eluted.

2.2. Mass spectrometry

Mass spectrometry was performed on a Hewlett–Packard HP1100 LC–MSD with an electrospray source. HP Chemstation software was used to control both the HPLC and the MS.

Peptide mapping was carried out in the scan mode, with the MS scanning from 400 to 2950 Da at an acquisition rate of 1 Hz and a step size of 0.3 Da. Data was filtered in the mass domain with a 0.03-u gaussian mass filter, but was not filtered in the time domain. No thresholding was carried out.

2.3. Materials

HPLC grade water was purified in-house (Barnstaed, Dubuque, IA, USA). Acetonitrile was HPLC grade (Burdick and Jackson, Muskegon, MI, USA). TFA and acetic acid (Sigma, St. Louis, MO, USA) were >99% purity. Recombinant tissue plasminogen activator (rt-PA) was obtained by courtesy of John Frenz at Genentech Inc. (South San Francisco, CA, USA).

The endoproteinase LysC digestion was performed

as follows. Ten nmol of the lyophilized rt-PA samples were reconstituted and denatured in 100 μl 6 *M* Guanidine. HCl. Reduction of each sample was achieved by the addition of 100 μl 50 m*M* dithiothreitol in Tris buffer (pH 7.8), followed by incubation at 37°C for 30 min. An addition of 100 μl of 100 m*M* iodoacetamide in Tris buffer (pH 7.8) followed by 30 min incubation at 37°C served to alkylate the sample. Each sample solution was then added to a vial containing 20 μg endoproteinase LysC (Promega, Madison, WI, USA), and Tris buffer (pH 7.8) was added to a final volume of 750 μl, and a final concentration of 1nmol/75 μl. The sample were allowed to digest for 24 h at 37°C, after which time they were stored at <0°C.

2.4. Data treatment

Mass spectral data were extracted from Chemstation format to ASCII using an in-house program. Chemstation does not store ''zero'' values, therefore in the extraction process missing values were replaced by ''zeros'' to enable matrix computation. An $m \times n$ matrix was produced by this procedure, where m corresponds to ca. 5000 time points and n corresponds to ca. 8500 m/z points. (The m/z range is from 400 to 2950 Da., and at a resolution of 0.3, this gives $(2950-400)/0.3=8500$ points in the mass axis). For ease of computation, each $m \times n$ was divided into 50 smaller matrices, each of size $(m/50) \times n$. Each matrix therefore consisted of approximately 100 spectra. All computations were carried out in the MATLAB programming environment (The Mathworks, Inc., Natick, MA, USA). The data processing algorithms were automated so that the individual matrices were analyzed separately but sequentially. Following analysis, total ion current (TIC) plots were formed by concatenating the TICs from each individual analysis. The computer was equipped with a Pentium Pro 200 processor, 128MB Ram and a Windows NT operating system. The analysis time for all matrices in a single sample under these conditions was typically 10 min.

3. Theory

Four previously reported data preprocessing methods were evaluated in this study. Singular value decomposition (SVD) [9] is a commonly used data reduction algorithm, which will be used in this case to improve the S/N in the data. The Component Detection Algorithm (CODA) [10], Sequential Paired Covariance (SPC) [11] and Higher Order Sequential Paired Covariance (HO-SPC) [12] are all methods that have been developed specifically for use with chromatography/mass spectrometry data.

The singular value decomposition [9] is an eigenvalue-like decomposition for rectangular matrices. An LC/MS peptide map for a given sample consists of a matrix, where one of the axes refers to elution time, and the other refers to m/z value. A matrix $\mathbf{A}$, dimensioned $m \times n$, can be decomposed as follows:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}} \tag{1}$$

where m refers to the number of spectral scans (or time steps) and n refers to the number of m/z increments. $\mathbf{U}$ is an $m \times m$ orthogonal matrix, $\mathbf{V}$ is an $n \times n$ orthogonal matrix, and $\mathbf{\Sigma}$ is an $m \times n$ diagonal matrix. The diagonal entries of $\mathbf{\Sigma}$ are the singular values of $\mathbf{A}$, and the columns of $\mathbf{U}$ and $\mathbf{V}$ contain the corresponding left and right singular vectors. The number of chemical components present in the sample/matrix is p, and this is referred to as the pseudorank of the matrix, since, theoretically, LC/MS data is bilinear. The matrix can be noise filtered by performing SVD and reconstructing the matrix using only the first p singular values and singular vectors.

$$\hat{\mathbf{A}}_{mxn} = \bar{\mathbf{U}}_{mxp}\bar{\mathbf{\Sigma}}_{pxp}\bar{\mathbf{V}}^{\mathrm{T}}_{pxn} \tag{2}$$

The noise and signal are not completely separated into different eigenvectors, however, meaning that some noise ''leaks'' into the first p vectors, while some signal may remain in the discarded singular vectors. In order to estimate the rank in a section of data, the decomposition is carried out as per Eq. (1), and the left and right singular vectors are plotted and examined. When the vectors begin to appear random, the remaining factors are discarded.

The component detection algorithm, CODA [10], is a method that was specifically developed for LC–MS data in order to reduce random noise and high background so as to simplify interpretation of

the data. A similarity index is calculated for each mass chromatogram:

$$S_j = \frac{1}{\sqrt{r-w}} \sum_{i=1}^{r-w+1} a(\lambda)_{ij} a(w,s)_{ij} \quad (3)$$

where $a(\lambda)_{ij}$ is length scaled data at time point i and mass channel j, $a(w, s)_{ij}$ is smoothed, standardized data, w is the window size for smoothing and r is the number of rows in the matrix.

The similarity index shown in Eq. (3) is calculated for each mass chromatogram, and has a value between 0 and 1. A high similarity indicates that the mass in question is part of the spectrum of an eluting component, while a low similarity suggests that any signal at that mass is likely to be due to noise. The user specifies a similarity threshold. Any mass chromatogram that has a similarity index greater than the similarity threshold is kept, while those mass chromatograms with smaller indices are ''thrown away''.

Sequential Paired Covariance (SPC) [11] is another method that has been developed specifically for chromatography/mass spectrometry data. The operation consists of an un-scaled, un-normalized correlation between two adjacent mass spectra by multiplying their intensities together mass by mass. Large intensities at each point will arise only if the adjacent spectra have common features. Noise that is un-correlated between neighboring spectra is suppressed by the operation.

$$I_{\mathrm{SPC}} = \sum_{i=1}^{n} (y_k - y_{k,\min})_i (y_{k+1} - y_{k+1,\min})_i \quad (4)$$

where I_{SPC} is the reconstructed total ion current at the kth time point, and $y_{k,i}$ is the original intensity at the kth time channel and ith mass channel. $y_{k,\min}$ is the minimum intensity at that time point.

Higher order SPC [12] is a variation of SPC [11], in that the intensities of more than two mass spectra are multiplied together at a time.

$$I_{\mathrm{HO-SPC}} = \sum_{i=1}^{n} (y_{k,i} - y_{k,\min})(y_{k+1,i} - y_{k+1,\min}) \ldots (y_{k+w-1,i} - y_{k+w-1,\min}) \quad (5)$$

where symbols have the same representations as in Eq. (4), and w represents the order of SPC. Higher order SPC is characterized by the fact that only features that are common to w adjacent mass spectra are kept.

The new preprocessing method reported here consists of two steps to remove random and high background noise, and is based on the assumption that analytes can be distinguished from noise by means of differences in peak width. The first step eliminates random noise by choosing a time window that corresponds to analyte peak width. Any ion that has a non-zero signal over the length of the window is retained. Those ions are selected by calculating $I1$ at time point j and m/z ratio k,

$$I1_{j,k} = y_{j,k} y_{j,k+1} \ldots y_{j,k+w_1-1} \quad (6)$$

where w_1 is the size of the time window to be used and $y_{j,k}$ is the raw ion intensity at time j and m/z k. If $I_{j,k}$ is zero, then the raw ion intensity, $y_{j,k}$, is eliminated. The process begins when j equals *1*, the time window is then moved forward by one time point to $j+1$, and the procedure is repeated until j equals $J - w_1 + 1$, where J is the number of data points in the mass chromatogram. A window of seven indicates that an analyte must elute over at least seven consecutive scans to be retained by the operation. A characteristic of random noise is that it does not have a constant signal over a number of scans, but displays zero intensities intermittently. In other words, it is assumed that random noise will not usually display a non-zero signal over a large number of consecutive scans. Multiplication of the ion intensities at a particular m/z value over a series of scans will result in a zero signal if a zero exists at that mass within those scans, and consequently, random noise will actually be eliminated by this operation. An analyte, on the other hand, will produce a constant signal over its elution time, and the signal will therefore be retained.

Not all LC/MS data will display intermittent ''zero'' intensities, but instead will have a very low, consistent background. This situation requires an additional step before the random noise is removed by the process described above. The mean of each chromatogram, **c**, should be subtracted,

$$\mathbf{c}_{\mathrm{new}} = \mathbf{c}_{\mathrm{old}} - \frac{1}{J} \sum_{j=1}^{J} \mathbf{c}_{j,\mathrm{old}} \quad (7)$$

where j and J have the same meanings as before. The effect of this operation should be to reduce the baseline noise, and in the process, introduce ''zero'' intensities in preparation for the first processing step.

The second processing step removes the high background that can often be caused by mobile phase and column bleed. This time, a time window w_2 is chosen that is much larger than the maximum expected elution time of an analyte ion. At each m/z value within the window, $I2$ is found,

$$I2_{j,k} = y_{j,k} y_{j,k+1} \cdots y_{j,k+w_2-1} \quad (8)$$

In this case, if $I2_{j,k}$ is *non*-zero, every data point within the window is eliminated.

The window is then moved forward to the $j + w$ timepoint, where the process is repeated. Any ion that has a consistent signal over a long period will not be due to an eluting analyte but to a high background, so that the multiplication in Eq. (8) is non-zero, leading to the identification of that ion as unwanted chemical noise.

It must be noted that there are actually several possible methods of implementing this processing technique. For example, since the method is based upon selecting data points within a window depending on whether or not a zero intensity is present, the algorithm could simply run through the windows and delete the signal at any particular m/z where a zero signal was present in the window. The authors have found, however, that implementation in the MATLAB programming environment is fastest when Eqs. (6) and (8) are used. In other programming languages, it is likely that this would not be the case, in which case an alternative implementation, such as that described above, should be used. The proposed name for this technique is the Windowed Mass Selection Method (WMSM).

4. Results and discussion

In order to evaluate the mass spectra before and after signal processing, the multiply-charged ions of the analytes in question must be identified. Electrospray ionization is a soft-ionization technique, and under the conditions used here with low in-source collision energy, very little fragmentation is seen; therefore, any ions that appear in a mass spectrum that are not multiply-charged ions are due either to noise or to an interfering species. Fig. 2(a) displays the total ion current (TIC) plot for the raw data, before any preprocessing was carried out: it is seen that the data is fairly noisy, with a large baseline and some noise spikes toward the beginning and end of the run. (See, for example, the region between 8 min and 25 min in Fig. 2(a).) Fig. 3(a) shows the mass spectrum for peptide A, while peptide B's mass spectrum is shown in Fig. 4(a). The multiply charged ions due to these peptides are indicated, and it is clear that the noise level is high, making identification of the analyte ions difficult. The first method to be implemented was SVD [9], and Fig. 2(b) contains the reconstructed TIC plot, while Figs. 3(b) and 4(b) display the selected mass spectra after the SVD processing. The TIC following SVD filtering (Fig. 2(b)) has a lower baseline than the raw data, but it can be seen that noise spikes are still present (e.g. a large spike remains at 25 min). Also, some regions of data remain noisier than others, as is seen between elution times 70 and 75 min. This is due to the fact that the rank estimation was carried out on small sections of data, and the estimation may have been more accurate in some sections than in others. Overestimating the number of components results in the inclusion of noise in the factors retained, making the reconstructed ion current noisy. Under-estimation of the pseudo-rank, on the other hand, results in the exclusion of components from the processed data. Figs. 3(b) and 4(b) contain the processed mass spectra of peptides A and B. The noise level has been reduced, but the SVD is, in effect, the same as averaging a number of runs, and the same result can be obtained by simply averaging the mass spectra over a chromatographic peak. Also, some negative intensities can be seen, which is due to the fact that the SVD is a mathematical operation with no positivity constraints.

A similarity index of 0.7 was chosen for the CODA analysis. This choice was made after performing CODA and using a number of different similarity thresholds, at which point a similarity threshold of 0.7 was found to retain all analyte signals while removing noise. It must be noted that, because the data was divided into time windows, separate indices were calculated for each section of

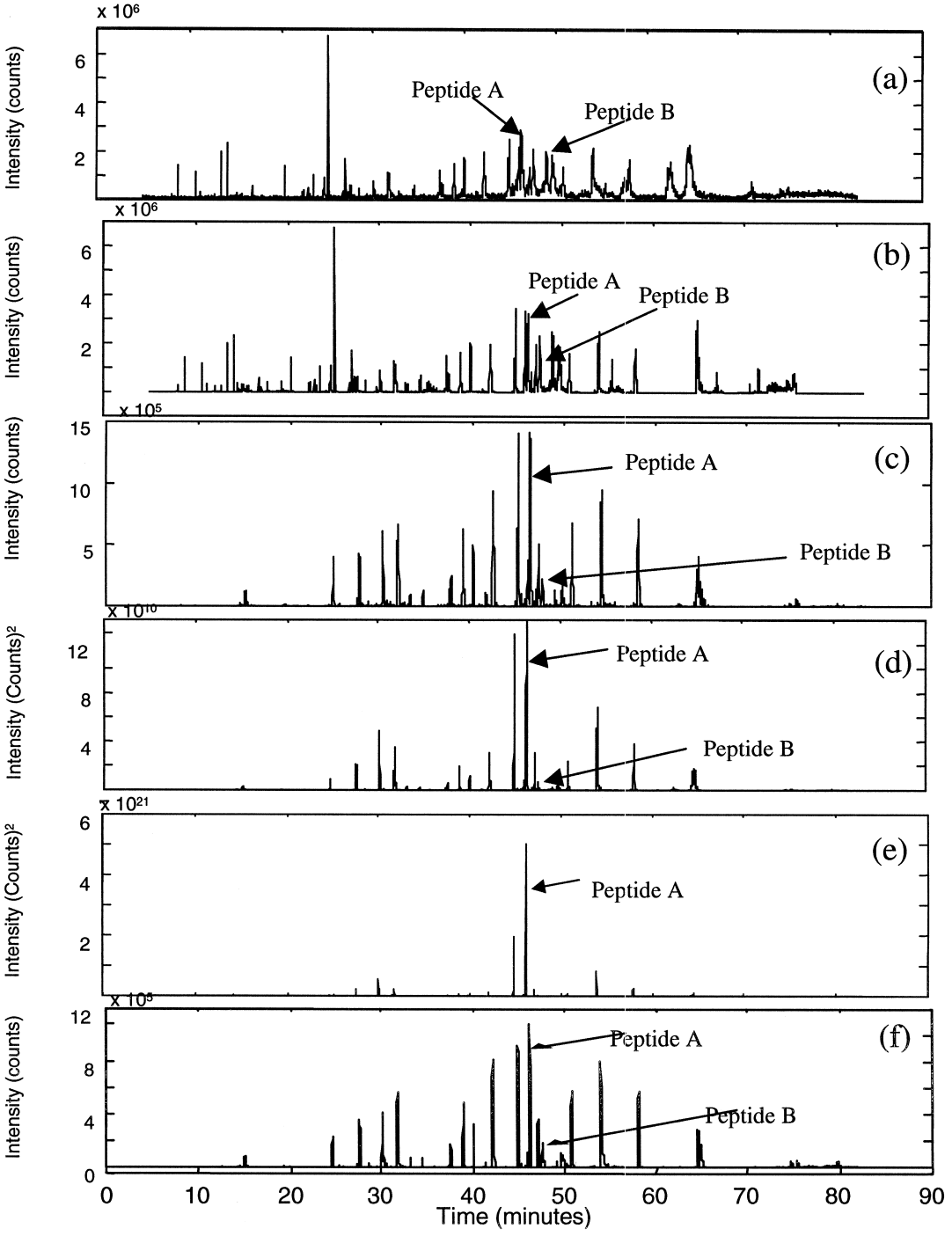

Fig. 2. Data processing results: total ion current (TIC) plots. (a) Raw data, (b) SVD results, (c) CODA results, (d) SPC results (e) 3rd Order SPC results, (f) WMSM, window = seven.

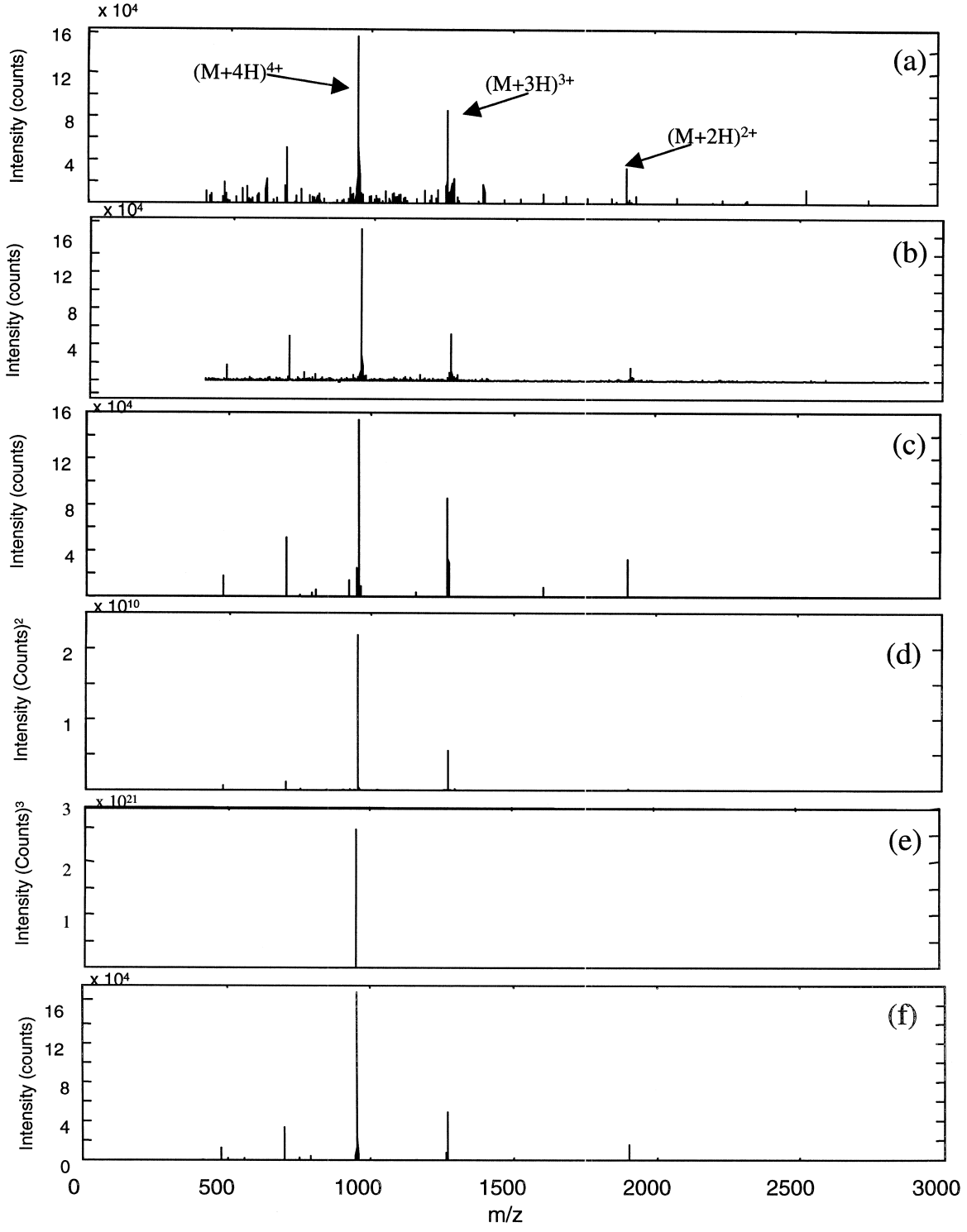

Fig. 3. Mass spectra of peptide A before and after data processing. (a) Raw data, (b) SVD results, (c) CODA results, (d) SPC results (e) 3rd Order SPC results, (f) WMSM, window = seven.

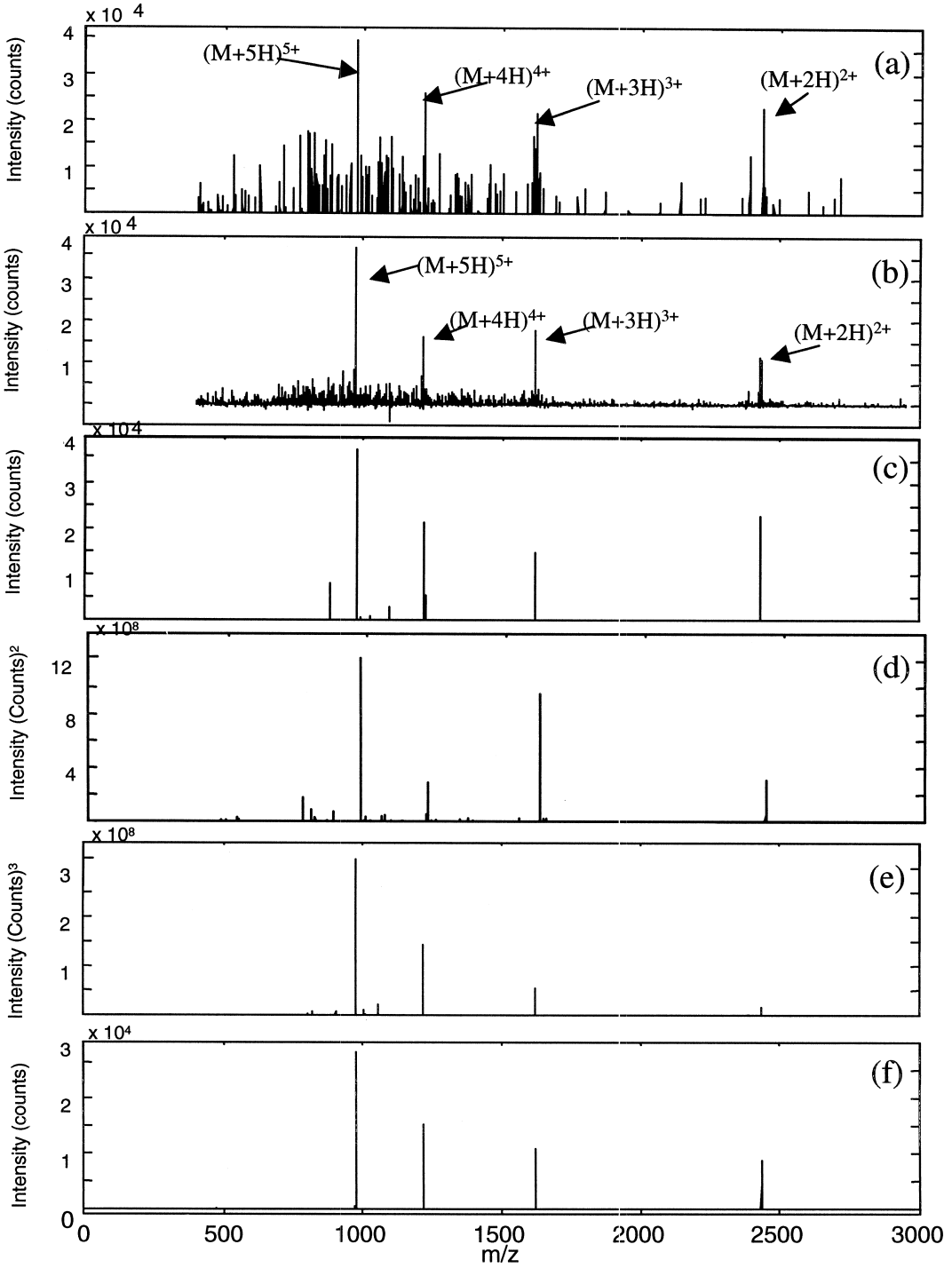

Fig. 4. Mass spectra of peptide B before and after processing. (a) Raw data, (b) SVD results, (c) CODA results, (d) SPC results (e) 3rd Order SPC results, (f) WMSM, window = seven.

each mass chromatogram. The calculations were performed in this way so that sections of the mass chromatograms that had no analyte information could be removed. In the TIC plot, Fig. 2(c), the absence of the baseline is noted, along with the disappearance of the noise spike that was present in the raw data. The mass spectra of peptides A and B (Figs. 3(c), 4(c)) demonstrate the ability of CODA to selectively remove noise while retaining analyte signal. However, it is seen that some peaks remain which are not multiply-charged ions of the peptides.

The next method that was tested was sequential paired covariance (SPC) [11]. The TIC plot following this analysis is shown in Fig. 2(d). While it can be seen that the baseline has been reduced, and the noise spike has been eliminated, it is also apparent that the data has been altered. Some peaks in the chromatogram have become much larger in size, while others have been reduced. Upon examination of the mass spectra (Figs. 3(d), 4(d)), the same effect is demonstrated: some of the analyte multiply charged ions have actually disappeared from the spectra-this is apparent in the mass spectrum of peptide A, where the $(M+2H)^{2+}$ ion has disappeared. The principle behind SPC is that analyte signal can be enhanced while noise can be suppressed through multiplication of adjacent spectra. However, the signal suppression also occurs for ions that are relatively small, while ions that initially have large intensities are enhanced. This method is therefore not suitable for unknown analytes, as it is difficult to tell whether or not any of the peaks in a processed mass spectrum have been suppressed. Although it might appear that applying SPC is similar to simply squaring each data point, Fig. 5 shows that this is not the case. The TIC after squaring, seen in Fig. 5(a), does not resemble the original data, while that after SPC, seen in Fig. 5(b), does. This is due to the fact that large noise spikes present in the original data are amplified by the squaring operation, but are suppressed by SPC as they do not appear in sequential spectra.

For the demonstration of higher order SPC [12], 3rd order SPC was applied to the data. As is to be expected following the discussion of SPC, signal suppression also results from the application of this method. In this case, however, the disappearance of small peaks from the mass spectra is even more evident, while large peaks are enhanced to a greater degree than with SPC. The TIC plot (Fig. 2(e)) shows that almost all of the peptide fragments are difficult to find in the chromatogram. While the protein in question produces 21 peptide fragments during digestion, the TIC plot displays only six clear chromatographic peaks. All others have been suppressed, including that of peptide B. Because it is known that peptide B elutes at 48 min, however, its mass spectrum can be extracted even though it's not visible in the TIC plot. The mass spectrum of peptide A (Fig. 3(e)) shows only the $(M+4H)^{4+}$ ion, while all ions due to peptide B (Fig. 4(e)) are visible in its mass spectrum. These results indicate that this method is therefore not suitable for situations in which either the elution time or the mass spectra of analytes are unknown. In the case of peptide A, the chromatographic signal was visible, but the mass spectrum was incomplete, while the mass spectrum of peptide B was complete, but its signal was not found in the TIC plot.

Following these experiments, it was apparent that the most effective method for preprocessing peptide map data was CODA, as it effectively removed noise without affecting the signal from analytes. However, the choice of the similarity index seems to be a matter of trial and error. In the original CODA paper, a similarity index of 0.85 was very effective, but when this value was used on the peptide map data, much of the signal was removed. It was subsequently found that the value of 0.7 was the most effective choice. This is more than likely due to the structure of the LC/MS data in this case. CODA works most efficiently when the chromatographic peak shapes are well-defined and there is a consistent baseline noise or a high background noise. This data has many mass chromatograms that appear as shown in Fig. 6, where there are many high intensity noise peaks throughout the run. Because of this, the similarity of the raw and smoothed data is less than would be the case if the analyte signal were not at the level of noise. A low similarity threshold is therefore required to retain analyte ions. However, the need for a low threshold also introduces the possibility that more noise will also be retained. The WMSM method that is introduced here uses a different criterion for the selection of analyte signal. The selection is based on the peak width of an

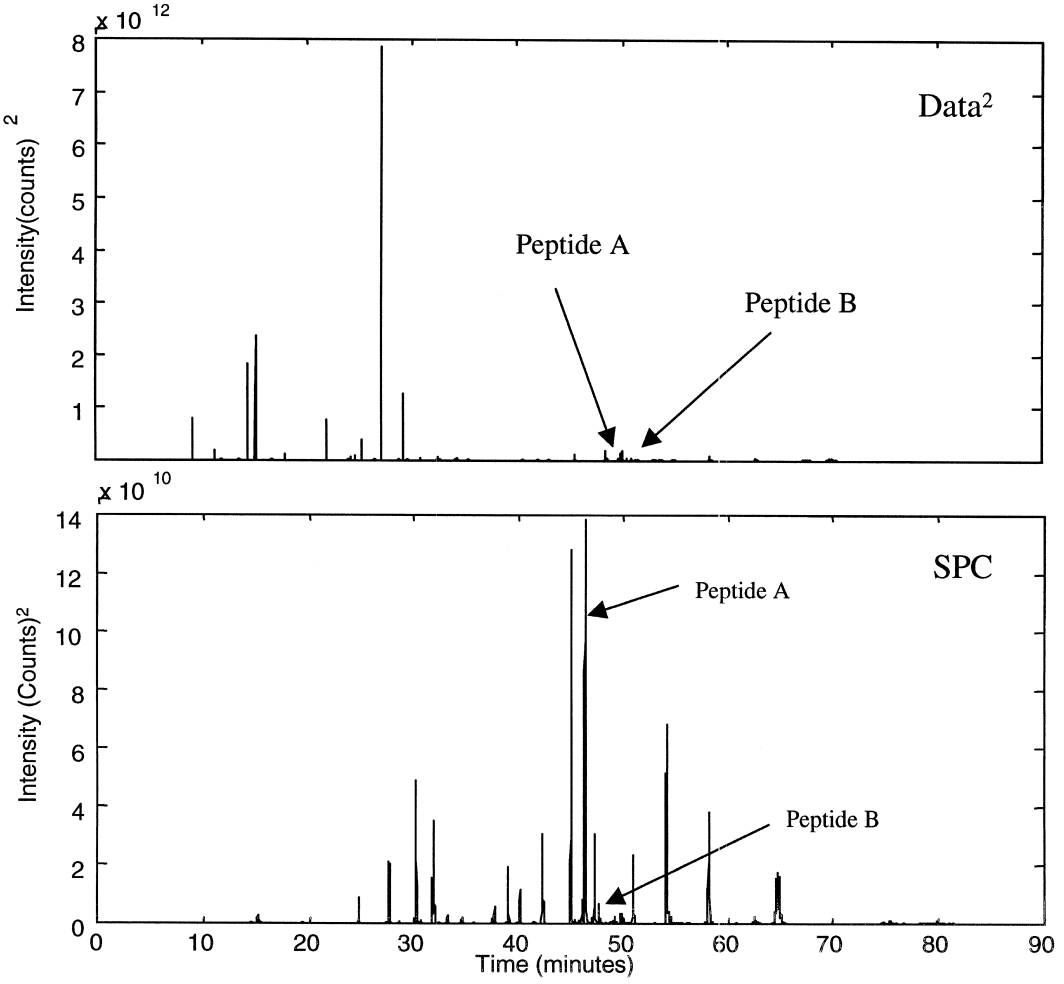

Fig. 5. Comparison of TIC plot after squaring the data (a) versus the TIC obtained from SPC processing (b).

analyte, and does not look at the entire length of either a mass chromatogram or a mass spectrum. This feature is useful for this type of data, which has high noise intermittently throughout the run, and where the peak shapes are not well-defined.

In order to apply the new preprocessing method to this data, a window size was first chosen. The window size is the expected width of a chromatographic peak for a single ion, and the elution profiles for individual ions are examined in order to choose this parameter. It is assumed that the elution times of different analytes will be similar, and therefore the individual ion chromatograms of only one representative analyte are examined. This window should be applicable to all data collected under the same conditions. The extracted ion chromatograms for the multiply charged ions of peptide B are therefore plotted in Fig. 7. (It is known that the elution profiles of this analyte are particularly weak). It can be seen that although the peptides appear to elute over 15–20 scans in the TIC plot, the individual ions have shorter elution times. The choice of window size is based on the peak width of the ion that elutes over the shortest time, in order to ensure that such ions are not eliminated by the method; in this case, it can be seen that the $(M+5H)^{5+}$ ion elutes over approximately nine scans. Since the window should actually be a little smaller than the expected peak width, a

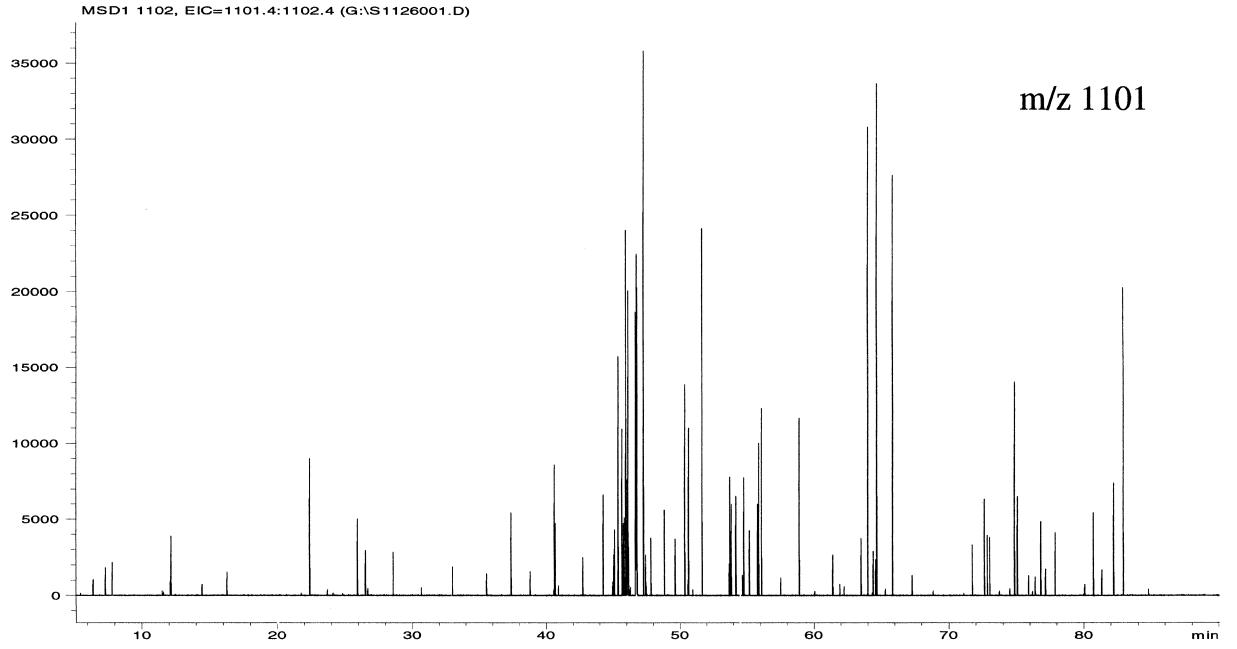

Fig. 6. Mass chromatogram of the ion at m/z 1101. This is an ion that has frequent high intensity noise spikes throughout the entire chromatographic run.

window of seven was therefore chosen for this analysis. Note that examination of the TIC plot alone would have led to a choice of ca. 15 for the window size. The computation time for this method was 5.2 s for a 340×1450 matrix under the conditions described in the experimental section.

The results from the new preprocessing method when the 7th order was used (i.e. $w = 7$) can be seen in Figs. 2(f), 3(f) and 4(f). The TIC plot, Fig. 2(f), appears very similar to that produced by CODA, Fig. 2(c), as the baseline noise has been reduced and the noise spike is also gone. The mass spectra, however, are not identical to those resulting from CODA processing. Peptide A (Fig. 3(f)), for example, retains those mass channels that are due to its multiply charged ions, plus some other peaks, while the mass spectrum of peptide B (Fig. 4(f)) contains its multiply charged ions only. The additional ions in the peptide A mass spectrum are not multiply-charged ions of A, nor are they related to each other. Because these ions were retained by WMSM over a window of seven scans, however, it is likely that they are due to some eluting components at that particular retention time. The extra ions that remain in the CODA mass spectrum of peptide B are probably real: the ions appear and disappear over a range of ca. 5 min around peptide B, never more than four scans in a row, thereby explaining their absence in the WMSM mass spectrum. However, their frequency and intensity in the region of peptide B is greater than in other regions in the chromatogram. The mass chromatogram of one of these ions, at m/z 1101, is displayed in Fig. 6. The spectra have been cleaned up to a large extent in both cases, however, and chemometric analysis should proceed more easily than with the raw data.

As with many processing methods, the choice a parameter, in this case, window size, is an important step in this method. A sub-optimal choice may result in extra noise being included, or analyte signal being lost. In order to demonstrate the influence of window size on the processing results, TIC plots for a range of window sizes are displayed in Fig. 8. It can be seen that as the window size increases, the total ion current decreases, and some peaks become smaller and eventually disappear. The choice of a window of

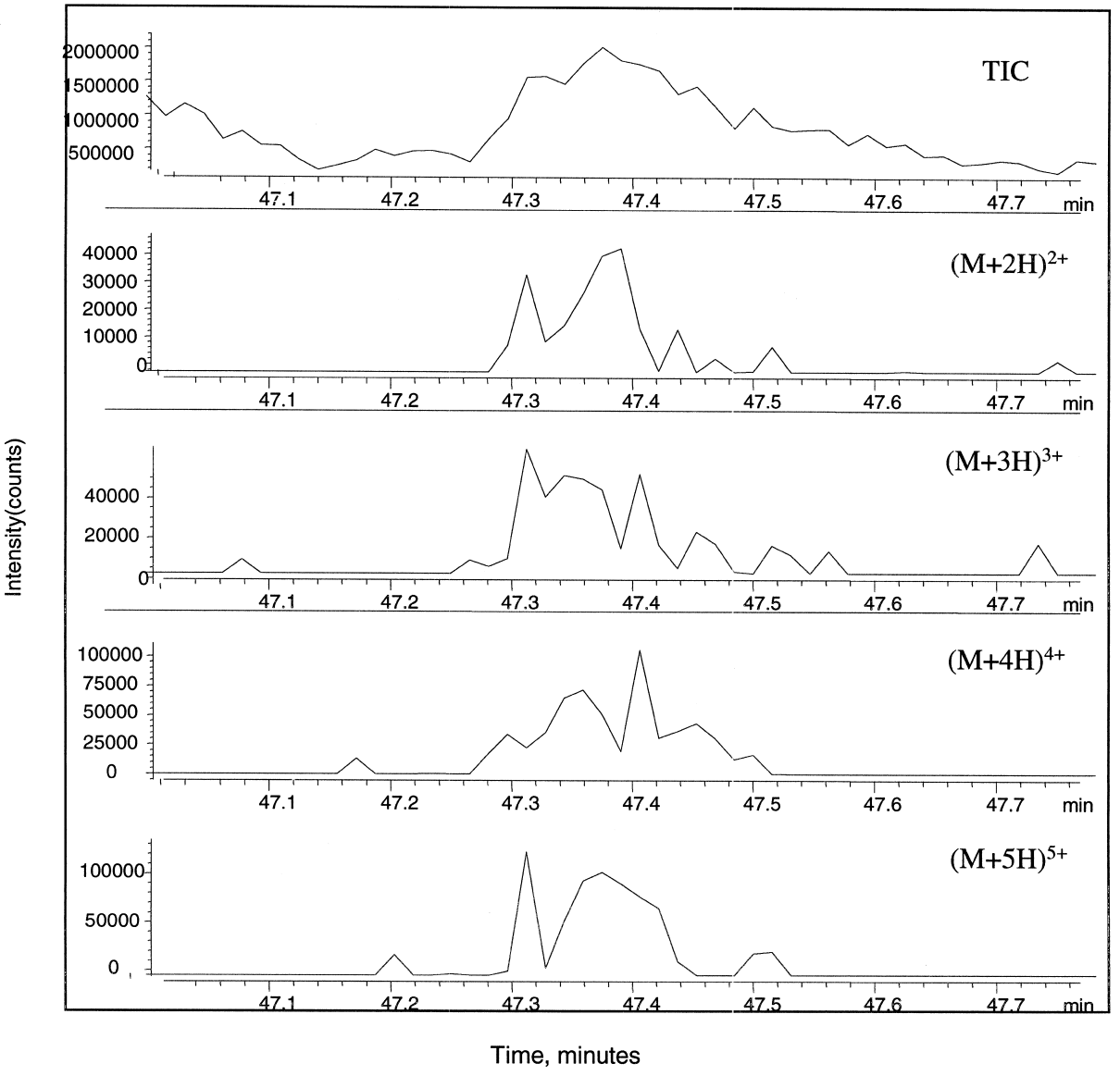

Fig. 7. Extracted ion chromatograms (EICs) are used as an aid in choosing the window size for WMSM. These are the multiply charged ions of peptide B.

seven for this analysis can be seen to be the best choice, as noise has been removed while signal remains. A window of fifteen results in the loss of analyte peaks: for example, the analyte that appeared at a retention time of 25 has disappeared, and the peaks have been dramatically narrowed due to the large window size. On the other hand, when a small window is used, the signal to noise is improved but noise still remains. It is now evident that a window size of fifteen (as was suggested by the TIC plot) would have resulted in the loss of some analyte peaks. It is best to choose a window on the small side, to ensure that signal is not inadvertently eliminated.

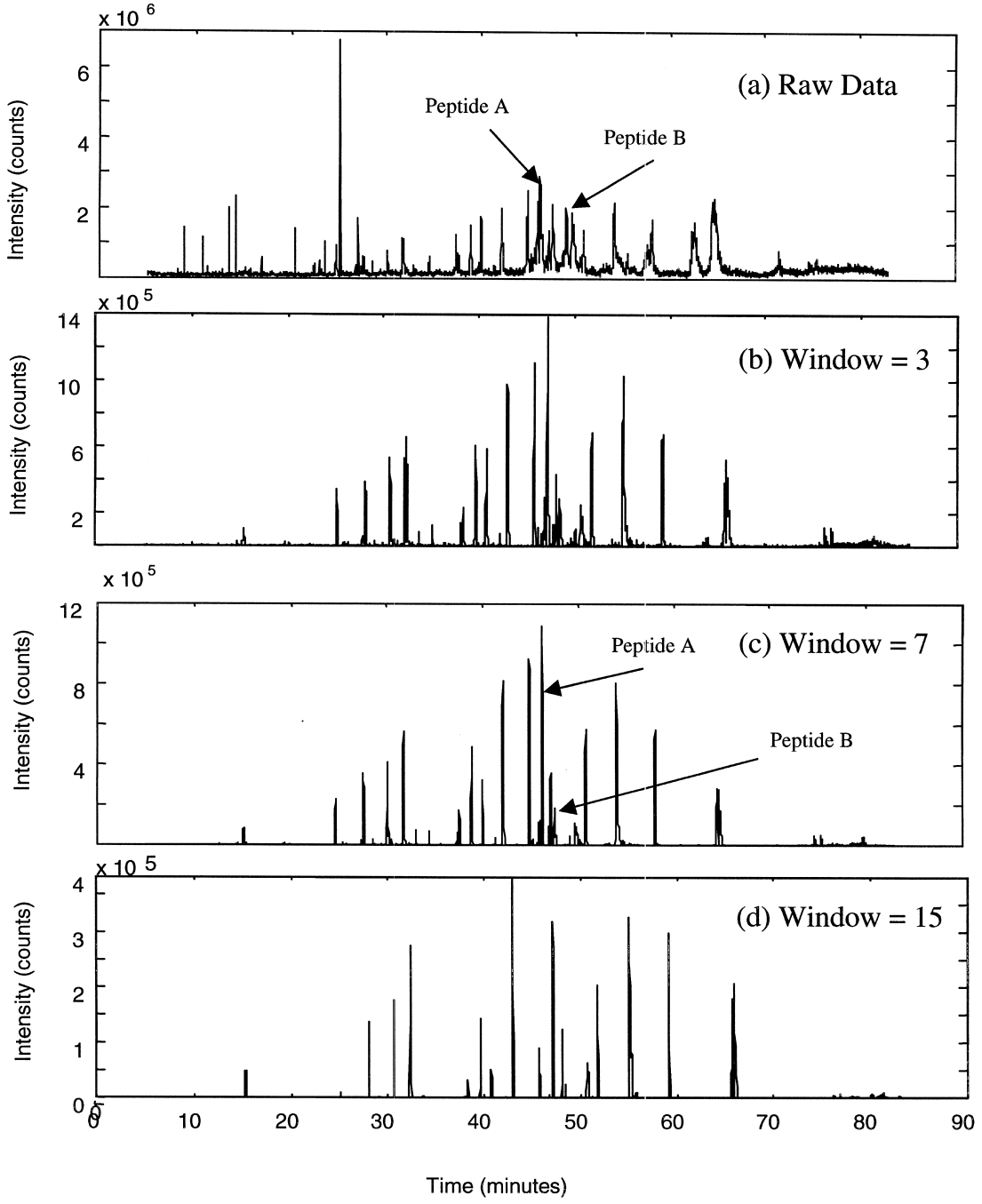

Fig. 8. Demonstration of the use of different window sizes with WMSM. (a) Raw data (b) window of three (c) window of seven (d) window of fifteen.

5. Conclusions

It has been shown that the noise that occurs in peptide map LC–MS data can be effectively removed, thereby simplifying interpretation of the mass spectra. There are two methods that are optimal: the first is the component detection algorithm (CODA), and the second is the windowed mass selection method (WMSM), an extension of higher order SPC that was introduced in this paper. The presence of frequent, high intensity noise peaks in the mass chromatograms of the data meant that, in this case, WMSM performed well. In other circumstances, where the data has a high background signal, and analyte peaks are smooth and continuous, CODA may be the better choice. The choice of processing method will ultimately depend on the structure of the data to be analyzed. It is expected that the use of preprocessing will aid in the interpretation of peptide maps, while also speeding up the process.

Acknowledgements

The authors would like to thank John Frenz of Genentech Inc. for supplying the rt-PA sample. C.M.F acknowledges the financial support of the Endowed Analytical Professorship.

References

[1] V. Ling, A.W. Guzzetta, E. Conova-Davis, J.T. Stults, W.S. Hancock, T.R. Covey, B.I. Shushan, Anal. Chem. 63 (1991) 2909.
[2] D.A. Lewis, A.W. Guzzetta, W.S. Hancock, M. Costello, Anal. Chem. 66 (1994) 585.
[3] W.M.A. Niessen, J. Chromatogr. A 794 (1998) 407.
[4] J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, C.M. Whitehouse, Mass Spectrom. Rev. 9 (1990) 37.
[5] R.D. Smith, T.A. Loo, Mass Spectrom. Rev. 10 (1991) 359.
[6] B. Mehlis, U. Kertscher, Anal. Chim. Acta 352 (1997) 71.
[7] A. Apffel, S. Fischer, G. Goldberg, P.C. Goodley, F.E. Kuhlmann, J. Chromatogr. A 712 (1995) 177.
[8] F.E. Kuhlmann, A. Apffel, S.M. Fischer, G. Goldberg, P.C. Goodley, J. Am. Soc. Mass Spectrom. 6 (1995) 1221.
[9] G. Golub, C. VanLoan, Matrix Computations, The Johns Hopkins University Press, Oxford, 1983.
[10] W. Windig, J.M. Phalp, A.W. Payne, Anal. Chem. 68 (1996) 3602.
[11] D.C. Muddiman, A.L. Rockwood, Q. Gao, J.C. Severs, H.R. Udseth, R.D. Smith, Anal. Chem. 67 (1995) 4371.
[12] D.C. Muddiman, B.M. Huang, G.A. Anderson, A.L. Rockwood, A. Proctor, Q. Wu, S.A. Hofstadler, M.S. Weir-Lipton, R.D. Smith, J. Chromatogr. A 771 (1997) 1.