Published in final edited form as:

Proteomics. 2009 August ; 9(15): 3833–3842. doi:10.1002/pmic.200800030.

A Novel Peak Detection Approach with Chemical Noise Removal Using Short-Time FFT for prOTOF MS Data

Shuqin Zhang^{1,2}, Honghui Wang⁴, Xiaobo Zhou^{1,3,*}, Gerard T. Hoehn⁴, Thomas J DeGraba⁵, Denise A. Gonzales⁴, Anthony F. Suffredini⁴, Wai-Ki Ching⁶, Michael, K. Ng⁷, and Stephen T.C. Wong^{1,3}

¹ HCNR-CBI, Harvard Medical School and Brigham & Women's Hospital, Boston, MA 02215

² School of Mathematical Sciences, Fudan University, Shanghai, 200433, China

³ Bioinformatics Core & Department of Radiology, The Methodist Hospital Research Institute, Weill Cornell Medical College, Houston, TX 77030

⁴ Critical Care Medicine Department, Clinical Center, National Institutes of Health, Bethesda, MD 20892

⁵ Neurology Department, National Naval Medical Center, Bethesda, MD 20889

⁶ Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong

⁷ Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Abstract

Peak detection is a pivotal first step in biomarker discovery from mass spectrometry (MS) data and can significantly influence the results of downstream data analysis steps. We developed a novel automatic peak detection method for prOTOF MS data which does not require *a priori* knowledge of protein masses. Random noise is removed by an undecimated wavelet transform and chemical noise is attenuated by an adaptive short-time discrete Fourier transform. Isotopic peaks corresponding to a single protein are combined by extracting an envelope over them. Depending on the signal-to-noise ratio (SNR), the desired peaks in each individual spectrum are detected and those with the highest intensity among their peak clusters are recorded. The common peaks among all the spectra are identified by choosing an appropriate cut-off threshold in the complete linkage hierarchical clustering. To remove the 1Da shifting of the peaks, the peak corresponding to the same protein is determined as the detected peak with the largest number among its neighborhood. We validated this method using a dataset of serial peptide and protein calibration standards. Compared with MoverZ program, our new method detects more peaks and significantly enhances SNR of the peak after the chemical noise removal. We then successfully applied this method to a dataset from prOTOF MS spectra of albumin and albumin-bound proteins from serum samples of 59 patients with carotid artery disease to detect peaks with SNR ≥ 2 . Our method is easily implemented and is highly effective to define peaks that will be used for disease classification or to highlight potential biomarkers.

Keywords

adaptive short-time discrete Fourier transform; complete linkage hierarchical clustering; peak detection; peak alignment; undecimated wavelet transform

*Corresponding to: XZhou@tmhs.org.

The authors have declared no conflict of interest.

1 Introduction

Mass spectrometry (MS) is increasingly used to detect disease-related biomarkers from human plasma or serum for early diagnosis, prognosis and monitoring of disease progression or response to treatment. Peak detection is a critical early analytical step that will directly influence the identification of potential biomarkers from the MS spectra. Due to the different properties of spectra data from different types of MS instruments, consistency in peak detection methods are essential for data analysis.

Matrix assisted laser desorption ionization-time of flight (MALDI-TOF) is one of the most popular MS approaches for detecting proteins or peptides. The prOTOF 2000 (PerkinElmer), an orthogonal MALDI-TOF MS, has high mass accuracy and resolution over a wide mass range, which overcomes many key performance barriers in traditional MALDI-TOFs. However, prOTOF MS spectra show a unique noise pattern in the background. In addition to the random noise normally observed in other high resolution mass spectra, there are low frequency noise peaks at 1Da intervals [Figure 1(a)]. The exact source of this noise is unknown but it is hypothesized to originate from the matrix since the matrix is also ionized during the data acquisition process [1]. Removal of this “chemical noise” may significantly improve the signal to noise ratio (SNR) in prOTOF MS spectra.

Several methods for peak detection in low resolution MALDI-TOF data have been described [2-5]. However for high resolution data, there are fewer published methods and each is specific to the MS platform used [6-9]. These methods are all based on the *a priori* knowledge of the masses of the peptides being queried and are applied to specifically match the known distributions of isotopic peptides in a spectrum. Yu W. et al [10] applied a Gaussian filter for smoothing and a Gabor quadrature filter to extract the envelope signal of each peptide. The main benefit of this method is that it combines all the peaks corresponding to one peptide together. However, the threshold for defining a peak was empiric and did not consider the “chemical noise” which may mask some small peaks. Jurgen K. et al [11] removed the “chemical noise” by Fourier transform with a fixed window size. This method was applied to the electrospray quadrupole TOF data with a maximum m/z of 2000 Da. However, Fourier transformation with a fixed window size cannot address high-frequency noise at higher m/z values in the operating range (>10 kDa) of the prOTOF 2000.

Peak alignment is the next step in MS data processing. Several methods have been applied to lower resolution MS data, but these are not applicable to the higher resolution prOTOF data [5,12]. Johnson K. J. et al [13] applied a closest point matching method in peak alignment. This method is limited by false-positive peak detection. Tibshirani R. et al used complete linkage hierarchical clustering that depends on threshold to cut the hierarchical tree [14]. Yu W. et al collected all the possible peaks into a super set to find the common peak set among all peak sets as a standard and then aligned all peak sets to the standard using the robust point matching approach in a sequential manner [10].

Based on these previous analytic approaches, we applied an adaptive, short-time, discrete Fourier transform combined with a wavelet transform to remove “chemical noise” and the higher frequency random noise from the prOTOF MS data. All of the single peaks in each spectrum were detected and those peaks corresponding to the same isotopic cluster were combined into an envelope. The desired peaks were then identified based on the SNR threshold. Complete linkage hierarchical clustering was used for peak alignment along all the spectra. The method was validated with a dataset with a series of peptide and protein calibration standards. Compared to the well known software MoverZ (Genomic Solutions) [15], it can detect small peaks with high signal to noise ratios. With the chemical noise

removal, the peak SNR is greatly enhanced. The proposed peak detection approach (PDA) with chemical noise removal using short-time FFT (stFFT) for prOTOF MS data is abbreviated as PDA_stFFT. Finally, we applied the method to a dataset from patients with symptomatic carotid artery atherosclerotic cardiovascular disease to detect peaks with $\text{SNR} \geq 2$. This paper focuses on the development and validation of PDA_stFFT method for peak detection. The biomarker identification and classification on the patients with carotid artery disease is the subject of a separate communication.

2 Materials and methods

2.1 Sample collection

Serum was obtained for proteomic analysis from a cohort of patients undergoing endarterectomy (EA) for occluded carotid arteries. The study was approved by institutional review board of the National Naval Medical Center, Bethesda, MD. Serum aliquots were stored at -70°C until used for this analysis. Patients from these cohorts were divided into four groups: symptomatic EA group (EAS, $N=12$), asymptomatic EA group (EAA, $N=11$), risk-factor matched control group (RFC, $N=18$), and control group (C, $N=18$). The EAS group was composed of patients who underwent EA surgery with evidence of prior clinical atherosclerotic plaque rupture in the form of a stroke or transient ischemic attack. The EAA group was composed of patients undergoing EA surgery with no clinical evidence of previous plaque rupture. The RFC group included patients without any evidence of atherosclerosis, but with risk factors, such as high cholesterol, smoking history, diabetes and/or hypertension. The C group was patients without any evidence or risk factors for atherosclerosis.

2.2 Sample preparation and MS acquisition for clinical samples

The samples were processed by ProXpression Biomarker Enrichment kit (PerkinElmer) according to the manufacturer's protocol [16]. Albumin and albumin-bound proteins were captured by Cibachron blue dye. After binding and washing, the enriched proteins were eluted, concentrated and applied in triplicate directly on MALDI chip targets (PerkinElmer). Mass spectra were acquired with a prOTOF™ 2000 MALDI-O-TOF MS. Mass range for all spectra was set to 700-10000 Da. The raw data for all spectra were exported in two-column text format for further data processing.

2.3 Sample preparation and MS acquisition for peptide and protein standards

To validate the peak detection method, we prepared a mixture of peptide and protein standards with protein calibration standard I and peptide calibration standard II (Bruker Daltonics). The protein or peptide names with their m/z values are listed in Table 1. Equal volumes of 10pmol/ μl of both standards in dilution buffer (1% TFA in 50% aqueous acetonitrile) were mixed together to prepare stock solution of 5pmol/ μl peptide and protein standards. The stock solution is then diluted 10x with dilution buffer to 500fmol/ μl as starting solution which was further diluted in 3-fold with dilution buffer to prepare a series of 8 diluted standard solutions of 500, 170, 56, 19, 6, 2.1, 0.7 and 0.23fmol/ μl . Mass spectra were acquired with PerkinElmer's prOTOF 2000 within m/z range of 1000 to 20000 Da. The instrument was calibrated first according to manufacturer's protocol. 1 μl of 8 prepared standard solutions was manually spotted on the target in triplicate. Then 1 μl of CHCA matrix is applied to each spot. Total 24 mass spectra were acquired as Ci_01, Ci_02 and Ci_03, $i = 1, 2, \dots, 8$.

The prOTOF 2000 used for clinical samples was located in the PerkinElmer facility in Boston, MA. The prOTOF 2000 used for peptide and protein standards was located at University of Arkansas for Medical Science, Little Rock, AK.

2.4 Noise modeling and filtering

Assume we have m spectra in study, and each takes on the same equally-spaced TOF t_k with $k = 1, 2, \dots, T$, where T is maximum number of m/z points. The intensity of the spectrum j is modeled as:

$$x_j(t_k) = N_j s_j(t_k) + c_j(t_k) + n_j(t_k), \quad j=1, 2, \dots, m, \quad k=1, 2, \dots, T;$$

where $S_j(t_k)$ is the true intensity of the spectrum j at TOF t_k and $x_j(t_k)$ is the observed intensity at TOF t_k . In this model, there are two different kinds of noise: $c_j(t_k)$, the “chemical noise”, and $n_j(t_k)$, the random noise which mainly results from the limitation of the detector. The “chemical noise” has an interval of approximately 1 Da. In typical *pro*TOF MS, multiply charged peptides are rarely observed, so we can assume that all peptides or proteins carry a single charge; thus the spacing between the isotopic peaks is 1 Da. We assume that the amplitude of the chemical noise is smaller than that of the true peaks in a local region.

Random noise has very high frequency and can be readily removed with a wavelet transform [2]. We use the undecimated wavelet transform (UDWT) to remove the noise $n_j(t_k)$. After computing the wavelet coefficients for the original signal, a hard threshold is applied. The threshold is set to be the product of a threshold parameter and a robust estimate of the noise: the median absolute deviation divided by 0.67. The software used in this study is the Rice Wavelet Toolbox (RWT) [<http://www-dsp.rice.edu/software/rwt.shtml>]. After removing the random noise, the model becomes:

$$\hat{x}_j(t_k) = N_j s_j(t_k) + c_j(t_k), \quad j=1, 2, \dots, m, \quad k=1, 2, \dots, T;$$

where $\hat{x}_j(t_k)$ is the intensity of the smoothed spectrum j at TOF t_k . Now we are left with the smoothed signal with chemical noise.

To identify the true peaks from the smoothed spectrum, the “chemical noise” $C_j(t_k)$ is removed with a short-time discrete Fourier transform (stDFT). Then the processed MS dataset is normalized to correct for the different total amount of proteins desorbed from the sample plate represented by N_j . Finally we identify the locations of the desired peaks in each spectrum according to a defined SNR. One of the goals of this study is to remove the chemical noise so that the peaks with comparatively small amplitude can also be detected with high SNR.

As noted, the chemical noise has the property that the spacing between two neighboring peaks is about 1Da and it has the shape similar to the sinusoidal signal. So we consider using discrete Fourier transform, which is the most popular transform to obtain the information from a signal that is not obviously in the time domain. It decomposes a signal to complex exponential functions of different frequencies and tells whether a certain frequency component exists or not in the signal. But it only applies for the stationary signal whose frequency content is unchanged in time. Since transferring the TOF data into m/z follows a formula like $m/z = a(t - t_0)^2 + b$, where a and b are related to the length of the flight tube and the applied voltage, $t - t_0$ is the flight time, the number of data points in a 1Da region is different and decreasing from lower m/z to higher m/z . The original DFT cannot be applied directly to the smoothed signal. To analyze the non-stationary signal of which some portion of its signal is stationary, DFT is extended to stDFT. In stDFT, the signal is divided into small enough segments, where each segment of the signal can be assumed to be stationary. For this purpose, a window function is chosen in such a way that the width of the window is equal to the segment of the signal where its stationarity is valid. Since the decreasing of the

number of data points in 1Da region with increasing m/z is very slow, and it is found that the number of data points in 1Da region is nearly the same in a certain region within several Daltons, we resort to stDFT. It can be described as:

$$X_j^w(l, u) = \sum_{k=1}^T w(k, u) x_j(t_k) e^{-2\pi i(k-1)(l-1)/T}, l=1, 2, \dots, T;$$

where $w(k, u)$ is the window function which is defined as:

$$w(k, u) = \begin{cases} 1, & \text{if } k - u < \text{window size}, \\ 0, & \text{otherwise.} \end{cases}$$

with u being the starting point of the window. The window of the stDFT for the smoothed signal $\hat{x}_j(t_k)$ is defined as the region in which the number of data points in 1Da is approximately the same. Figure 1(b) shows the number of data points in approximate 1Da region. So we define the stDFT window as the regions with the same number of data points in 1Da shown in the Figure 1(b). To compute DFT, fast Fourier transform is applied which is an efficient algorithm and reduces the computational complexity from $O(T^2)$ to $O(T \log T)$, where T is the number of data points in the signal.

After stDFT, we restore the sinusoidal signal that has the highest amplitude in the stDFT frequency domain in each window to estimate the chemical noise. When m/z is less than a certain value (about 4000Da), the smoothed signal $\hat{x}_j(t_k)$ has the property that the spacing between peaks is 1Da. So the estimated chemical noise still has the period of 1Da. But when m/z becomes larger, undecimated wavelet transform smoothed the spectrum. The restored signal becomes an estimate of the long distance smoothed signal. Figure 1(c, d) shows the smoothed signal and the estimate of the chemical noise.

To remove the chemical noise from the smoothed signal, we compared the estimated chemical noise with the smoothed signal. Here a hard threshold is applied. If the net signal, defined as the original smoothed signal minus the estimate of the chemical noise, is less than a certain threshold, it is set to be zero. After this step, we are left with an estimate of $N_j S_j(t_k)$.

2.5 Peak identification

After de-noising the signal, we get an estimate of the true intensity in each individual spectrum. Peak identification in each estimated signal is described below.

Normalization of the processed spectrum—In our model, the parameter N_j represents the amount of proteins desorbed from the sample plate. This number may vary as the total amount of proteins desorbed from a sample plate may differ between experiments. In order to correct for these differences, we define the total ion current as the mean intensity of the processed spectrum, the number N_j . The intensity of the processed spectrum is normalized by dividing it by the total ion current, providing an estimate of the true signal $S_j(t_k)$.

Definition of the signal-to-noise ratio (SNR)—SNR is defined as the ratio of the intensity of the true signal $s_j(t_k)$ to the background noise. Background noise is considered to be represented by the random noise $n_j(t_k)$. The noise level is defined as the mean intensity of the random noise in a certain region.

Peak identification—In the prOTOF data, each protein corresponds to a cluster of peaks that includes all the isotopic peaks. To combine these peaks together, an envelope signal is extracted by fitting a curve to the local maximum points in each 1Da of the signal $S_j(t_k)$. This envelope combines all the peaks corresponding to the same protein into a concave curve. Then the local maximum value of all such concave curves is determined among the cluster of isotope peaks. The m/z of all these local maximums determines the locations of the proteins. This approach differs from some other peak detection methods [6,7] that detect the monoisotopic peak as the location of the protein. By selecting proper SNR threshold, we can detect the desired peaks. Figure 1(e) shows the envelope of the normalized signal and the peaks we detected by setting the SNR to be 2. The curve over the denoised signal is the envelope and the circles label the positions of the peaks.

2.6 Peak alignment

In the peak detection, we locate the peak in the position where it achieves the highest intensity in its peak cluster. Since the position of the peak that has the highest intensity among its isotope peak cluster for all the spectra may vary about 1Da, the detected position of the peaks that correspond to the same protein may shift 1Da left or right too. Thus we need to choose a position that can reflect the proper m/z of the peaks along all the spectra. Before that, we first align the peaks that should correspond to the same m/z since the same peaks can be horizontally shifted within a small region in different spectra. We propose to use the complete linkage hierarchical clustering to do it. This method has been applied in [13], but for our dataset, several points require further explanation.

To do the clustering for all the peaks, the peaks in all the spectra are first ordered into a $M \times 1$ vector, where M is the total number of peaks in all the spectra. The distance between each two peaks is computed. In complete linkage clustering, the distance between one cluster and another cluster is the greatest distance from any member of one cluster to any member of the other cluster. Here, Euclidian distance is applied. Starting by assigning each item to a cluster, we can find the closest pair of clusters and merge them into a single cluster. Then the distance between the new cluster and each of the old clusters is determined. The process will be repeated till all the clusters are merged. Finally, all the items are clustered into a single cluster of size M . In practice, we do not need all the items to be clustered into a single cluster, so we need to define a proper height to cut the tree to get the desired clustering precision. In [13], a predefined threshold 0.005 is applied to cut the tree. This value is too large to get a good estimation of the peak position in our dataset. If a large threshold is applied, some noisy peaks can be included into the true peaks cluster, which results in the shifting of the positions of the true peaks. In the prOTOF data, the m/z shift is very small. Even when m/z is about 4000Da, the shifting is still less than 0.2Da. Thus in our testing, we chose the cutoff threshold to be 0.002.

We estimate the position of the peaks using the mean value of all the peaks in the same cluster. If a peak is a noise peak, the cluster that it belongs to should be very small. Thus, a true peak is identified if the size of the cluster is greater than a certain value. The remaining peaks are considered as noise and not considered further. This results in a series of common peaks along all the spectra.

To find a proper value that can reflect the position where the protein achieves highest intensity, we do the following. For a single peak, if there are no peaks with m/z distance about 1Da around it, it is considered as the peak. If there are peaks with spacing about 1Da, we record the number of the peaks and choose the peak that has the largest number as the final peak.

2.7 MoverZ software

MoverZ is a program for viewing and manipulating biopolymer mass spectra, capable of performing a wide variety of manual and automated tasks such as peak labeling, calibration, etc. It is an open software and can be downloaded from:

'<http://bioinformatics.genomicsolutions.com/MoverZDL.html>'. MoverZ contains its own peak labeling algorithm that uses signal to noise and resolution as parameters to determine mono-isotopic peaks. First it attempts to assign average masses to all of the peak distributions that have a root-mean-squared SNR greater than a value set by the user. Then it attempts to find a resolved monoisotopic peak in the appropriate position relative to the average mass. The final peaks are determined by a fuzzy logic decision. Further details about MoverZ can be found at '<http://bioinformatics.genomicsolutions.com/MoverZ.html>'.

3 Results

We first validated our method using a dataset acquired with a series of peptide and protein calibration standards. Next we showed the results of peak detection and alignment for all the four groups of spectra from the patients with carotid artery disease.

3.1 Validation of PDA_stFFT

To validate PDA_stFFT, we tested all the datasets acquired with a series of peptide and protein standards and recorded the SNR of all the known peaks in the peptide and protein standards. The number of peaks detected with PDA_stFFT when SNR is greater than 2 is listed in Table 2. The SNR for the first replicate (C_i_01, $i = 1, 2, \dots, 8$) is shown in Table 3. The other two replicates have the similar results. As described in the Materials and Methods section, the peptide and protein standards were diluted in 3-fold to prepare a series of 8 diluted samples. The peptide and protein concentration for C1 (C1_01 or C1_02 or C1_03) was selected so that we expected to detect all known peaks with prOTOF 2000. As the concentration of peptide and protein was further diluted, we expected the SNR to reduce with dilution. Thus PDA_stFFT will detect fewer peaks as the dilution increases or no peak as the concentration of peptide or protein fails below the detection limit of the instrument. The data in Table 2 reflect this pattern.

Figure 2(a) shows the three raw spectra around the peak with m/z 757Da in sample C1_0i, $i = 1, 2, 3$ and Figure 2(b) shows the SNR of the denoised signal with smoothing and chemical noise removal. Figure 2(c) is the same as Figure 2(b) without the chemical noise removal. From Figure 2(b), we can see that the SNR is enhanced significantly after the chemical noise removal.

3.2 Comparison between PDA_stFFT and MoverZ

To see the efficiency of our method, PDA_stFFT, we compared the total number of the peaks detected with PDA_stFFT and MoverZ.

Table 2 shows the number of peaks with $\text{SNR} \geq 2$ detected with PDA_stFFT and MoverZ. Among all the cases, except in the sample C1_01, our method detects more peaks than MoverZ. This may be due to the removal of chemical noise which enhances the SNR of the peak. The SNR for a peak obtained with our method is much larger than that with MoverZ in most cases. We list the SNR of all the peaks in the sample C1_01, C1_02 and C1_03 in Table 1 as an example. Similar results were achieved with other datasets. In Table 1, "None" means the SNR is smaller than 2 in that position. When m/z is small, the SNR calculated with our method can be several times as that with MoverZ. When m/z becomes very large, MoverZ may detect the peaks with higher SNR than PDA_stFFT. This is because the detection sensitivity of prOTOF 2000 decreases with the increase of m/z , and the peak

intensity of a true signal becomes weaker when m/z is very large. Since the chemical noise in high m/z region is similar to that in low m/z region, after the chemical noise removal, the signal may become even weaker, which may make the SNR smaller.

One advantage of PDA_stFFT over MoverZ is that it can discern some small peaks from the chemical noise. Figure 2(a) shows a peak at m/z 757 which belongs to the peptide “Bradykinin fragment 1-7 (m/z 757Da)”. PDA_stFFT detected this peak with a high SNR while MoverZ failed to detect it as demonstrated in Table 1. Bradykinin fragment 1-7 in C1_01, C1_02 and C1_03 was detected by PDA_stFFT as a peak with SNR of 101.9, 61.9 and 33.1 respectively (Figure 2(a)). However, MoverZ only detected 757Da peak in C1_01 with SNR of 2.3 and failed to detect 757Da peak in C1_02 and C1_03.

3.3 Results of peak detection on the clinical sample dataset

In the wavelet transform, the threshold parameter was set to 6, allowing smoothing of the signal. During the chemical noise removal, the threshold for determining the chemical noise is set to 80 percentiles of the value obtained by subtracting the estimated chemical noise from the smoothed signal in each adaptive region. If the number of peaks in one cluster is more than 15% of the number of spectra in one group in the peak alignment step, we take it as the true peak, and get the position of the peak by computing the mean m/z over all the peaks in its peak cluster. Otherwise, we consider them noise peaks.

Before the peak identification step, it is necessary to remove the outlier or noise data using the unsupervised information. To find the noise data, the mean of all the spectra is determined. If the mean value of the spectrum ranges between mean minus the deviation and mean plus the deviation, we consider the spectrum acceptable for further analysis. For example, the EAS group has 12 patients and 11 patients have three replicate spectra and one patient has two replicate spectra. Within these 35 spectra, one spectrum was found to be an outlier and was removed from the dataset. A total 894 of common peaks were detected from these 34 spectra after the alignment from 12265 detected peaks. After we combined the peaks that may correspond to the same protein together by choosing the peak with the maximum number in its neighborhood, 519 peaks were defined for further analysis.

While the peaks aligned well (Figure 3(a)), there was a shift in a very small region. We aligned the peaks with the complete linkage clustering and determined the position of each common peak as the average position. To show the peak alignment and detection processes, we focused on the region between 1700Da and 1900Da of the spectra. Figure 3(b) shows all the peaks detected and the aligned peaks between 1700Da and 1900Da. In total, 677 peaks were detected and 41 peaks defined as common peaks. After combining the peaks that correspond to the same protein together, 24 peaks were defined (Figure 3(c) with the solid line). Figure 3(d) shows all the smoothed spectra in this region and the detected peaks. The vertical lines correspond to the peaks detected by PDA_stFFT after alignment and the final determined peaks are pointed out in the Figure 3(d) with label of “Final peak”.

As observed in Figure 3(d), the PDA_stFFT can detect very small peaks, such as m/z 1762, 1833 and 1835 with $SNR \geq 2$. From the two enlarged regions within Figure 3(d), two to three detected peaks (with a vertical line at center of the peak) are 1Da apart and these peaks may correspond to the same protein. Figure 3(e, f) shows the peaks detected in these two small regions. The m/z 1762 peak was determined as the final peak from the peak cluster between m/z 1760 and 1765 and two peaks of m/z 1833 and 1835 were determined as final two peaks between m/z 1831 and 1839. These three peaks are plotted as solid lines in Figure 3(e) and (f). The peaks with dotted lines in Figure 3(e) and (f) are considered to be isotopic peaks.

With the same setting, we finally are able to detect 465 peaks in the EAA group, 589 peaks in the control group, and 334 peaks in the RFC group. These peaks will be used for classification of a disease model to find any potential biomarker or panel of biomarkers that can distinguish difference among the groups.

4 Discussion

Peak detection plays an essential role in biomarker discovery from MS data. The peak detection method however depends significantly on the property of the data. An automatic peak detection method is proposed in this paper for *pr*OTOF data. Random noise is removed by the undecimated wavelet transform. We have designed an adaptive short-time discrete Fourier transform to address chemical noise removal, which increases of the signal-to-noise ratio of the true peaks. The possible isotopic peaks corresponding to one protein are combined together by extracting an envelope over them. Depending on SNR, the desired peaks in each individual spectrum can be identified. In contrast to other methods for identifying peptides, this method is independent of the knowledge of the mass of the peptides. In addition, since we use the short-time discrete Fourier transform to remove the chemical noise, it can be applied to a wide range of mass. The final peaks are determined by choosing the peak which has the largest number among its neighbors with a distance of about 1Da in the aligned common peaks obtained by complete linkage hierarchical clustering for all the peaks detected in all the spectra. We believe this approach can be easily implemented and is very effective.

We validated our PDA_stFFT method with a dataset acquired with an analysis of the peptide and protein calibration standards. The results are consistent with our experiment settings. The SNR for known peaks is significantly increased after chemical noise removal. Compared to the popular software MoverZ, our method can detect more peaks in diluted samples and can detect peaks with low intensity with high SNR. Finally, we successfully applied our method to detect the peaks from the spectra of a large dataset derived from clinical serum samples.

Acknowledgments

The authors wish to thank Dr. Lisa Sapp in PerkinElmer (Boston MA) for acquiring all *pr*OTOF MS for clinical samples. This work is partially funded by IBIS award (Zhou) and TMHRI scholarship award (Zhou). This research was supported in part by the Intramural Program of the NIH, Clinical Center as well. The authors would like to thank Dominik Beck for proofreading the manuscript.

5 References

1. Krutchinsky AN, Chait BT. On the nature of the chemical noise in MALDI mass spectra. *J. Am. Soc. Mass Spectrom.* 2002; 13:129–134. [PubMed: 11838016]
2. Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, et al. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics.* 2005; 5:4107–4117. [PubMed: 16254928]
3. Morris JS, Coombes KR, Koomen J, Baggerly KA, Kobayashi R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics.* 2005; 21:1764–1775. [PubMed: 15673564]
4. Du P, Kibbe WA, Lin SM. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics.* 2006; 22:2059–2065. [PubMed: 16820428]

5. Yasui Y, McLerran D, Adam B, Winget M, et al. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *J. Biomed. Biotechnol.* 2003; 4:242–248. [PubMed: 14615632]
6. Berndt P, Hobohm U, Langen H. Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis.* 1999; 20:3521–3526. [PubMed: 10612278]
7. Gras R, Muller M, Gasteiger E, Gay S, et al. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis.* 1999; 20:3535–3550. [PubMed: 10612280]
8. Breen EJ, Hopwood FG, Williams KL, Wilkins MR. Automatic poisson peak harvesting for high throughput protein identification. *Electrophoresis.* 2000; 21:2243–2251. [PubMed: 10892735]
9. Samuelsson J, Dalevi D, Levander F, Rognvaldsson T. Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics.* 2004; 20:3628–3635. [PubMed: 15297302]
10. Yu W, Wu B, Lin N, Stone K, Williams K, Zhao H. Detecting and aligning peaks in mass spectrometry data with applications to MALDI. *Comput. Biol. Chem.* 2006; 30:27–38. [PubMed: 16298163]
11. Jurgen K, Marc G, Keith R, Matthias W. Noise filtering techniques for electrospray quadrupole time of flight mass spectra. *J. Am. Soc. Mass Spectrom.* 2003; 14:766–776. [PubMed: 12837599]
12. Yasui Y, Pepe M, Thompson ML, Adam BL, et al. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics.* 2003; 4:449–463. [PubMed: 12925511]
13. Johnson KJ, Wright BW, Jarman KH, Synovec RE. High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *J. Chromatogr. A.* 2003; 996:141–155. [PubMed: 12830915]
14. Tibshirani R, Hastie T, Narasimhan B, Soltys S, Shi G, Koong A, Le Q. Sample classification from protein mass spectrometry by peak probability contrasts. *Bioinformatics.* 2004; 20:3034–3044. [PubMed: 15226172]
15. <http://bioinformatics.genomicsolutions.com/MoverZ.html>
16. Lopez MF, Milkulskis A, Kuzdzal S, Bennett DA, et al. High-resolution serum proteomic profiling of Alzheimer disease samples reveals disease-specific, carrier-protein-bound mass signatures. *Clin. Chem.* 2005; 51(10):1946–1954. [PubMed: 16081505]

Figure 1.

Process of peak identification in a single spectrum. **(a)** A sample raw spectrum. Peptide signal, random noise and chemical noise are indicated with arrows. The chemical noise has a unique pattern that peaks at 1Da apart. **(b)** Approximate number of data points in 1Da region as a function of m/z . **(c)(d)** Examples of estimating the chemical noise with stDFT. These two figures show the smoothed signals of the two enlarged regions in (a). The smooth curves show the estimate of the chemical noise. **(f)** Example of envelope extraction and peak detection. The curve over the signal shows the envelopes extraction. Each concave part of the curve corresponds to the same protein. The small circles label the peaks detected by our method.

Figure 2. SNR enhancement with chemical noise removal. (a) The three raw spectra around the peak with m/z 757Da in sample C1_0i, $i = 1, 2, 3$. (b) The SNR of the denoised signal with smoothing and chemical noise removal. (c) Same as 2b without the chemical noise removal.

Figure 3.

Peak alignment and detection processes for the EAS group. **(a)** All peaks detected in EAS group when SNR is 2. **(b)** Peak alignment between 1700Da and 1900Da in EAS group when SNR set to 2. The lines show the peak points after the alignment based on the detected peaks from all the spectra. **(c)** Peak determination between 1700Da and 1900Da. Both solid lines and dotted lines show the aligned peaks in (b). The solid lines show the final determined peaks by choosing the peak that has the largest number among its neighborhood with distance about 1Da. **(d)** All 34 smoothed spectra from 1720Da to 1900Da. Both solid lines and dotted lines are corresponding to the peaks in (c). Two small regions near m/z 1762 and 1834 were zoomed in to see the details. **(e)(f)** Peak detection in the two small regions near m/z 1762 and 1834 shown in (d). It shows all the detected peaks in each spectrum in these two small regions. The solid lines show the final determined peaks.

Table 1

Comparison of SNR from PDA_stFFT and MoverZ for three replicates of peptide and protein standards at concentration of C1 (500fmol/μl)

Peptide/Protein Standard	(M+H)	s/n with MoverZ			s/n with PDA_stFFT		
		Cl_01	Cl_02	Cl_03	Cl_01	Cl_02	Cl_03
Peptide standard II							
Angiotensin II	1046.5	77.7	56.7	29.5	832.7	618.3	424.5
Angiotensin I	1296.7	34.7	28.8	18.8	391.8	293.3	217.6
Substance P	1347.7	45.9	35.4	23.9	436.4	360.0	314.6
ACTH Clip 1-17	2093.1	36.0	27.1	22.0	256.4	178.8	159.3
ACTH Clip 18-39	2465.2	69.8	58.1	38.3	269.8	219.9	172.6
Somatostatin28	3147.5	26.7	27.5	24.0	122.6	103.8	106.1
Bradykinin fragment 1-7	757.4	2.3	None	None	101.9	61.9	33.1
Renin Substrate	1758.9	12.1	10.3	7.3	110.6	81.4	68.1
Protein standard I							
Insulin	5734.6	3.5	3.2	3.0	6.1	6.4	4.8
Ubiquitin I	8565.9	4.2	4.4	4.7	6.4	6.1	6.7
Cytochrome C	12361, 6181	2.7 None	3.1 2.1	3.1 2.0	None 2.4	None 3.2	2.1 3.0
Myoglobin	16952, 8476	None None	None None	None None	None None	None None	None None

Number of peak detected with our method and MoverZ in a serial of peptide and protein standards, $Ci_01.i = 1, 2, \dots, 8$

Table 2

Data		C1_01	C2_01	C3_01	C4_01	C5_01	C6_01	C7_01	C8_01
NO of peaks	PDA_sFFT	11	8	8	8	4	4	1	2
	MoverZ	11	7	4	1	0	1	0	0
Data		C1_02	C2_02	C3_02	C4_02	C5_02	C6_02	C7_02	C8_02
NO of peaks	PDA_sFFT	11	8	8	8	4	2	1	1
	MoverZ	10	7	5	1	0	1	0	0
Data		C1_03	C2_03	C3_03	C4_03	C5_03	C6_03	C7_03	C8_03
NO of peaks	PDA_sFFT	11	8	8	8	3	3	2	1
	MoverZ	10	7	5	0	1	0	0	0

Table 3

The SNR values calculated with PDA_stFFT for a serial of peptide and protein standards, Ci_01, i = 1, 2, ..., 8.

Peptide/Protein Standard	C1_01	C2_01	C3_01	C4_01	C5_01	C6_01	C7_01	C8_01
Angiotensin II	832.7	124.7	40.1	23.9	9.2	6.7		
Angiotensin I	391.8	59.1	14.9	10.5	2.3			2.1
Substance P	436.4	81.6	17.1	10.9		5.0		
ACTH Clip 1-17	256.4	31.8	5.4	6.0				
ACTH Clip 18-39	269.8	45.4	10.1	11.1				
Somatostatin 28	122.6	18.8	2.7	2.7				
Bradykinin Fragment 1-7	101.9	24.2	13.1	7.4	4.4	5.3	9.1	27.1
Renin Substrate	110.6	17.3	4.6	3.4	2.5	5.0		
Insulin	6.1							
Ubiquitin I	6.4							
Cytochrome C, m/z 6181	2.4							