The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 222–231

Peak Detection with Chemical Noise Removal Using Short-Time FFT for a Kind of MALDI Data

Shu-Qin Zhang[1,2] Xiaobo Zhou[1,*] Honghui Wang[3]
Anthony Suffredini[3] Denise Gonzales[3] Wai-Ki Ching[2]
Michael K. Ng[4] Stephen Wong[1]

[1] HCNR-CBI, Harvard Medical School and Brigham & Women's Hospital, Boston, MA 02215
[2] Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong
[3] NIH Clinical Center, Bethesda, MD 20892
[4] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Abstract Peak detection is the first step in biomarker extraction from the mass spectrometry data, which significantly influences the results of the following steps. Designing a good method for peak detection greatly depends on the properties of the data. In this paper, we propose a novel automatic peak detection method *without the a priori knowledge on the mass of the proteins* for a kind of MALDI data, which have a regular noise pattern called chemical noise except the random noise. The random noise is removed by using the undecimated wavelet transform. An adaptive short time discrete Fourier transform is proposed to do the chemical noise de-noising. We combine the possible peaks corresponding to one protein by extracting an envelope over them. Depending on the signal-to-noise ratio, the desired peaks that have the highest intensity among their peak clusters in each individual spectrum are detected. We examine the performance of the method in the carotid artery disease data set that shows the efficiency of the method. With the chemical noise removal, the signal-to-noise ratio of the peaks is increased greatly compared to the result without chemical noise removal.

Keywords Peak detection; adaptive short time discrete Fourier transform; undecimated wavelet transforml

1 Introduction

Matrix assisted laser desorption ionization-time of flight (MALDI-TOF) is one of the most popular mass spectrometry (MS) approaches applied presently for detecting the proteins or peptides which could be the disease related biomarkers from human plasma or serum for early diagnosis, prognosis and monitoring of disease progression or response to treatment. We are particularly interested in one kind of

*Corresponding to: zhou@crystal.harvard.edu

MALDI data. To extract the biomarkers from the MS data, a necessary step is peak detection, which directly influences the results of the following biomarker identification steps. Due to the different properties of the data from different kinds of MS instruments, proper peak detection methods should be developed to treat the corresponding data.

In this kind of MS spectra, except the random noise we normally observed in other high resolution MS data, there is a unique noise pattern in the background as shown in Figure 1. The low frequency noise peaks at 1Da apart. It should mainly originate from the application of the matrix since the matrix is also ionized during the data acquisition process [7]. Here we refer this type of noise as chemical noise. Removal of this type of chemical noise may significantly improve the signal/noise ratio the peaks.

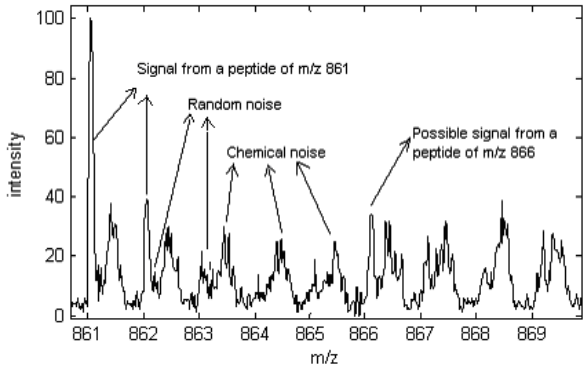

Figure 1: Part of a raw spectrum. The chemical noise has a unique pattern that peaks at 1 Da apart.

There are several published methods [3,4,8,10] for peak detection of the low resolution MALDI-TOF data, or similarly the SELDI-TOF data. But for the high resolution data, there are only a few literatures [1,2,5,9], in which the method depends on the data greatly. These methods are all based on the a priori knowledge of the mass of the peptides or proteins. Different models are applied in these papers to match the distribution of isotopic peptides or proteins one by one to detect all of them in a spectrum. For peptides with m/z less than 4000Da, the isotopic cluster is very easy to recognize. But for this kind of MS data tested here, the maximum m/z is 20000Da. As the m/z increases, the isotopic cluster is becoming more complicated and more difficult to identify. Yu W. et al [11] proposed to use the Gaussian filter to smooth the signal and Gabor quadrature filter to extract the envelope signal of the same peptide. The main benefit of this method is to combine all the peaks corresponding to one peptide together. However, they got the threshold of defining peaks via an empirical approach without considering the noise model, with which some small peaks can be omitted. Jurgen K. et al [6] removed the regular noise (chemi-

cal noise) by Fourier transform with fixed window size. This method is applied to the eletrospray quadrupole TOF data, whose maximum m/z is less than 2000Da. As m/z is becoming larger, Fourier transform with a fixed window size cannot capture the property of the regular noise, which may result in the removal of the true peaks because of the complexity of the protein structure when m/z is large.

In this paper, we propose to use adaptive short time discrete Fourier transform combined with wavelet transform to remove the chemical noise and the random noise. All the single peaks are detected and those peaks that may correspond to a same protein or belong to the same isotopic cluster are combined into an envelope. Finally, the desired peaks are detected based on the signal-to-noise ratio. Numerical experiments based on our data show the efficiency of the method.

2 Method

2.1 Formulation of the problem

Assume we have m spectra in study, and each takes on the same equally-spaced time-of-flight (TOF) t_k with $k = 1,2,...T$. The intensity of the spectrum j is modeled as:

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

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. Note that this model is different from the model in [2], where there is no the term: $c_j(t_k)$. For our data, there are two different kinds of noise: $c_j(t_k)$ and $n_j(t_k)$. $c_j(t_k)$is the so-called chemical noise of spectrum j at TOF t_k. Although such noise has been widely recognized, and some experimental methods have been proposed to reduce it, there are few informatics methods to handle it. As mentioned in the introduction, the chemical noise has the property that the spacing between the two neighboring peaks is about 1Da. $n_j(t_k)$ is the random noise which mainly results from the limitation of the detector. Figure 2 shows an example of the spectrum. In such a spectrum, multiple charged peptides are rarely observed, so we can assume that all the peptides or proteins carry a single charge; thus the spacing between the isotopic peaks is 1Da. We assume that the amplitude of the chemical noise is smaller than that of the true peaks in a certain local region. With the increasing of m/z, the signal of chemical noise becomes weaker. The random noise has very high frequency and it can be easily detected in both figures.

To remove the random noise of the individual spectrum, wavelet transform is a good choice and it has been applied widely to do de-noising [8]. Here we just use the method presented in [8]. After removing the random noise, the model becomes:

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

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 that correspond to the possible biomarkers from the smoothed spectrum, a very useful step is to remove the chemical noise $c_j(t_k)$. We

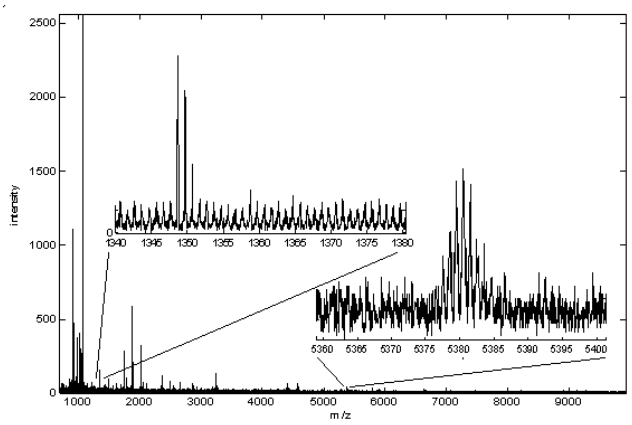

Figure 2: A sample spectrum. The largest m/z of it is more than 9000Da. The two enlarged regions show the detailed signals when m/z is relatively small and large in this spectrum. When m/z is small, the pattern of the chemical noise can be seen clearly, while when m/z is becoming larger, it is weaker.

propose doing it with short time discrete Fourier transform (STDFT). Then, we normalize the processed signal to correct the different total amount of ions desorbed from the sample plate. The locations of the desired peaks are identified in each spectrum depending on the defined signal-to-noise ratio(SNR) finally. Now our main purpose is to remove the chemical noise, such that the peaks with comparatively small amplitude can also be detected with high signal-to-noise ratio.

2.2 Remove the chemical noise with short-time discrete Fourier transform

As we introduced in the last section, 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 get the information from the signal that is not obvious 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 points in a 1Da region is different and decreasing from lower m/z to higher m/z. So 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 the short time DFT (STDFT). In STDFT, the signal is divided into small enough segments, where the segments 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 is very slow, and from the data it is found that the number of points in 1Da region is nearly the same in a certain region within several Daltons, we resort to STDFT. It can be described as:

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

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

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

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 points in 1Da is approximately equal. Figure 3 shows the number of points in approximate 1Da region (Starting from a point, find the first point that is greater than the starting point by 1Da). So we define the STDFT window as the regions with the same number of points in 1Da shown in the figure. 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 doing 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 $\hat{x}_j(t_k)$ keeps the property that the spacing between neighboring peaks is 1Da. So such an estimate of the chemical noise also has the period 1Da. But when m/z becomes larger, UDWT smoothed the spectrum. The restored signal becomes an estimate of the long distance smoothed signal.

Figure 4 shows the smoothed signal and the estimate of the chemical noise. To remove the chemical noise from the smoothed signal, we compare such an estimate of the chemical noise with the smoothed signal. Here a hard threshold is applied. If the net signal which is 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 $N_j s_j(t_k)$.

2.3 Peak identification

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

Normalization of the processed spectrum: The intensity of the processed spectrum is normalized by dividing it with the total ion current, which is defined as the mean intensity of the processed spectrum. This step is to correct the systematic differences in the total amount of ions desorbed from the sample plate represented by N_j. Now we got the estimate of the true signal $s_j(t_k)$.

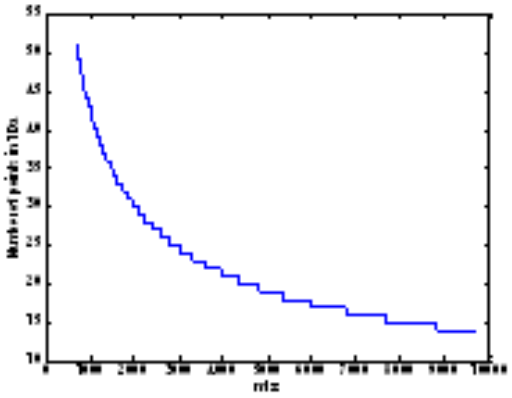

Figure 3: Approximate number of data points in 1Da region as a function of m/z. With the increasing of m/z, the region with same number of time point is becoming larger.

Definition of the signal-to-noise ratio(SNR): SNR is defined as the ratio of the true signal to the random noise after removal of the chemical noise. The true intensity of the signal is just $s_j(t_k)$. The noise level is defined as the mean intensity of the random noise in a certain region.

Peak identification: In our tested data, each protein or peptide corresponds to a cluster of peaks that includes all the isotopic peaks. So we need to combine all the peaks that correspond to the same protein or peptide together. To do this, we extract an envelope signal by fitting a curve to the local maximum points in

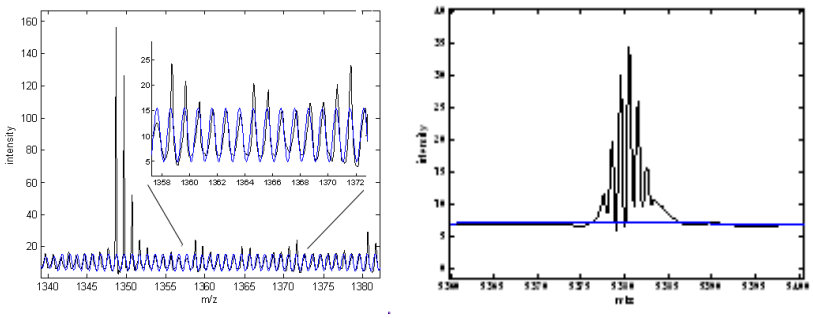

Figure 4: Examples of estimating the chemical noise with STDFT. These two figures show the smoothed signals of the two enlarged regions in Figure 2. The curves with same amplitude show the estimate of the chemical noise.

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

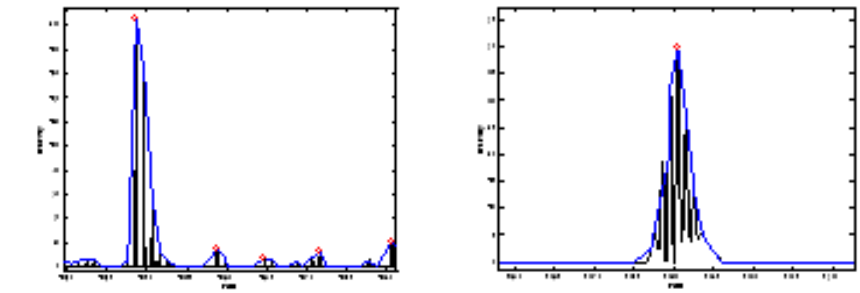

Figure 5: Examples of envelope extraction and peak detection. The curves above the de-noised spectrum show the envelopes extraction. Each concave part of the curve corresponds to the same protein. The small circles label the peaks detected by our method.

such concave curves is detected to determine which peak has the highest intensity among the cluster of isotope peaks. We take the m/z of all these local maximums as the locations of the proteins. This differs from some other peak detection methods [2, 5] that detect monoisotopic peak as the location of the protein. By selecting a proper SNR threshold, we can detect the desired peaks. Figure 5 shows the envelope of the normalized signal and the peaks we detected by setting the SNR to be 2.

3 Results

The process of finding the peaks has been shown in the method section. In this section, we show some results of peak detection with our proposed method.

The data got from the serum of a cohort of patients undergoing endarterctomy (EA) for occluded carotid arteries are applied for testing our method. Our final goal is to discern protein expression differences in serum. Here we use the data of a group of all the patients: the symptomatic EA patients (EAS). After the sample preparation and the data acquisition, we finally got 35 spectra.

In the wavelet transform, the threshold parameter is set to be 6, with which we found that the signal can be smoothed well. During the chemical noise removal, the threshold for determining the chemical noise is set to 80 percentiles of the value got 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 same 1Da region, 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 take them as noise peaks.

Our method can detect the peaks very well. For all the 35 samples in the EAS group, we remove one bad sample, and consider the rest 34 samples. Totally 12265 peaks are detected and after the alignment, finally we got 894 common peaks when SNR is 2. Figure 6 shows all the peaks in the 34 samples. For these 894 peaks, after

we combine the peaks that may correspond to the same protein together by choosing the peak with the maximum number in its neighborhood, we got 519 peaks finally.

With the extraction of the chemical noise, our method can improve the SNR of the peaks with small amplitude greatly. Figure 7 shows the comparison of the SNR of the same spectrum with and without chemical extraction. In the left figure, the SNR is got with the same method as described above but without the chemical noise extraction. And the right figure shows the SNR of the spectrum when there is chemical noise extraction. It is clear that with the chemical noise extraction, SNR of some small peaks is enhanced greatly.

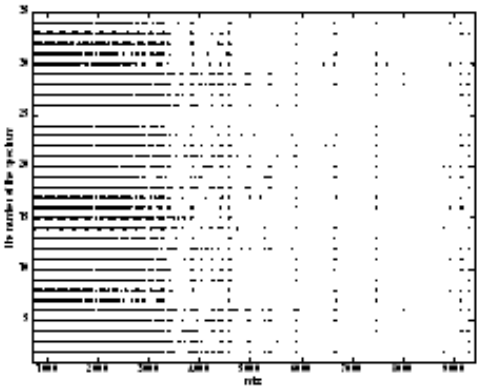

Figure 6: All peaks detected in EAS group when SNR is 2.

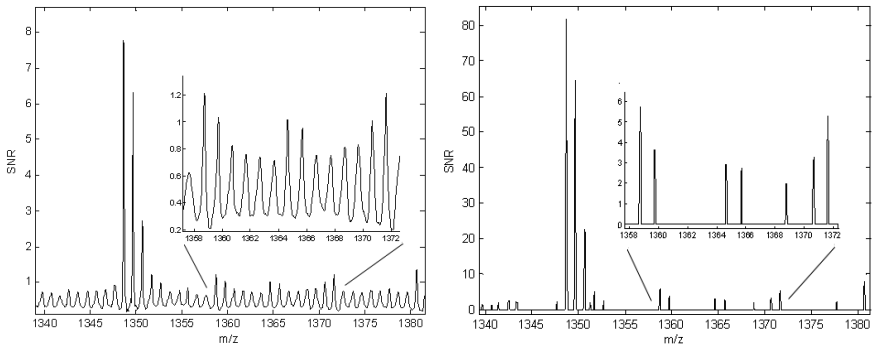

Figure 7: Comparison of SNR between the same spectrum without and with removal of chemical noise. Left: the SNR got without chemical noise removal. Right: SNR got with chemical noise removal.

4 Concluding Remarks

Peak detection plays an important part in the biomarker extraction from the MS data. The design of the peak detection method depends on the property of the data greatly. An automatic peak detection method is proposed in this paper for a kind of MALDI data with regular low frequency noise called chemical noise except the random noise. We design an adaptive short time discrete Fourier transform to remove the chemical noise, which increases the signal-to-noise ratio of the true peaks compared to the results without chemical noise removal. The peaks are located at the positions where they have the highest intensity among all their isotopic peaks. Differing from most existed methods for identifying peaks, this method is independent of the knowledge on the mass of the peptides or proteins, which can be applied for a wide range of mass since with the mass increasing, the structure of the proteins is more complicated, and the existed methods may not be applied any more. The method is easily implemented and very effective.

For all the spectra, the peaks with maximum intensity among their neighborhood may shift 1Da. Next step, we need to align all the peaks that correspond to the same protein or peptide but with shifting. The ultimate goal of this study is to discover the potential biomarkers that can differentiate the disease samples from the normal ones so that they can be used clinically.

Acknowledgements

The authors will thank Dr. Lisa Sapp in PerkinElmer (Boston MA) for acquiring all the MS data.

References

[1] Berndt, P., Hobohm, U., Langen, H., Reliable automatic protein identification from matrix-assisted laser desorption ionization mass spectrometric peptide finger prints, *Electrophoresis*, 1999, 20, 3521-3526.

[2] Breen, E. J, Hopwood, F.G., Williams, K.L., Wilkins, M. R., Automatic poisson peak harvesting for high throughput protein identification, *Electrophoresis*, 2000, 21, 2243-2251.

[3] Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M. C., H. M. Kuerer, PImproved 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.

[4] Du, P., Kibbe, W. A., Lin, S. M., Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, *Bioinformatics*, 2006, 22, 2059-2065.

[5] Gras, R., Muller, M., Gasteiger, E., Gay, S., Binz, P. A., Bienvenut, W., Hoogland, C., Sanchez, J. C., Bairoch, A., Hochstrasser, D. R., Appel, R. D., Improv-

ing protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection**,** *Electrophoresis*, 1999, 20, 3535-3550.

[6] 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.

[7] Krutchinsky, A. N., Chait, B. T., On the nature of the chemical noise in MALDI mass spectra, *Journal American Society for Mass Spectrometry*, 2002,13, 129-134.

[8] Morris, J. S., Coombes, K. R., Koomen , J., Baggerly, K. A., Kobayashi, R., Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum*, Bioinformatics*, 2005, 21, 1764-1775.

[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.

[10] Yasui, Y., McLerran, D., Adam, B., Winget, M., Thornquist, M., Feng, Z., An automated peak-identification / calibration procedure for high-dimensional protein measures from mass spectrometers, *Journal of Biomedicine and Biotechnology*, 2003, 4, 242-248.

[11] Yu, W., Wu, B., Lin, N., Stone, K., Williams, K., Zhao, H., Detecting and Aligning Peaks in Mass Spectrometry Data with Applications to MALDI, *Computational Biology and Chemistry,* 2006, 30, 27-38.