GABORLOCAL: PEAK DETECTION IN MASS SPECTRUM BY GABOR FILTERS AND GAUSSIAN LOCAL MAXIMA

Nha Nguyen

Department of Electrical Engineering, University of Texas at Arlington, TX, USA
Email: nhn3175@exchange.uta.edu

Heng Huang*

Department of Computer Science and Engineering, University of Texas at Arlington, TX, USA
**Email: heng@uta.edu*

Soontorn Oraintara

Department of Electrical Engineering, University of Texas at Arlington, TX, USA
Email: oraintar@uta.edu

An Vo

Department of Electrical Engineering, University of Texas at Arlington, TX, USA
Email: vpnan@msp.uta.edu

Mass Spectrometry (MS) is increasingly being used to discover disease related proteomic patterns. The peak detection step is one of most important steps in the typical analysis of MS data. Recently, many new algorithms have been proposed to increase true position rate with low false position rate in peak detection. Most of them follow two approaches: one is denoising approach and the other one is decomposing approach. In the previous studies, the decomposition of MS data method shows more potential than the first one. In this paper, we propose a new method named GaborLocal which can detect more true peaks with a very low false position rate. The Gaussian local maxima is employed for peak detection, because it is robust to noise in signals. Moreover, the maximum rank of peaks is defined at the first time to identify peaks instead of using the sinal-to-noise ratio and the Gabor filter is used to decompose the raw MS signal. We perform the proposed method on the real SELDI-TOF spectrum with known polypeptide positions. The experimental results demonstrate our method outperforms other common used methods in the receiver operating characteristic (ROC) curve.

1. INTRODUCTION

Mass Spectrometry (MS) is an analytical technique has been widely used to discover disease related proteomic patterns. From these proteomic patterns, researchers can identify bio-markers, make a early diagnosis, observe disease progression, response to treatment and so on. Peak detection is one of most important steps in the analysis of mass spectrum because its performance directly effects the other processing steps and final results such as profile alignment [1], bio-marker identification [2] and protein identification [3].

There are two types of peak detection approaches: denoising [4, 5] and non-denoising (or decomposing) [6, 7] approaches. There are several similar steps between these two approaches such as baseline correction, alignment of spectrograms and normalization. They also use local maxima to detect peak positions and use some rules to quantify peaks. Specially, both approaches use the signal to noise ratio (SNR) to remove some small energy peaks whose their SNR values are less than a threshold. However, in the denoising approach, before detecting peaks, a denoising step is added to reduce the noise of mass spectrum data. In the non-denoising approach, a decomposition step is used to analyze mass spectrum into different scales before the peak detection by local maxima. When the smoothing step is applied into the denoising approach, it possibly removes both noise and signal. If the real peaks are removed by

*Corresponding author.

smoothing step, they can never be recovered in the other processing steps. As a result, we lose some important information and introduce error into MS data analysis. Thus, the way we decompose a signal into many scales without denoising is a really better approach with great potentials.

The SNR is used to identify peaks in both denoising and non-denoising methods. In paper [6], P. Du *et al* estimated the SNR in the wavelet space and got much better results than the previous work. But they still failed to detect the peak at 147300 and some peaks with small SNR. This problem came from the SNR value estimation and all previous methods estimated the SNR value by using the relationship between the peak amplitude and the surrounding noise level. Since some sources of noise can also have high amplitudes, the high amplitude peak does not always guarantee to be real peak. On the other hand, some low amplitude peaks can also be real peaks. It is clearly that the way using SNR to quantify peaks is not efficient and accurate. More details of this problem will be discussed in section 3.4. In this paper, we propose a new robust decomposing based MS peak detection approach. We use the Gabor filters to create many scales from one signal without smoothing. The Gaussian local maxima is exploited to detect peaks instead of the local maxima because the Gaussian local maxima method is more robust to the noise of mass spectrum. Finally, we use the maximum rank (MR) of peaks to remove some false peaks instead of the SNR. The real SELDI-TOF spectrum with known polypeptide composition and position is used to evaluate our method. The experimental results show that our new approach can detect both high amplitude and small amplitude peaks with a low false position rate and is much better than the previous methods.

2. METHODS

In this section, we first introduce the basic knowledge of Gabor filters. After that, our proposed method which is a combination of the Gabor filters and the Gaussian local maxima will be detailed. At last, we will use one example to show how our method works.

2.1. Gabor Filters

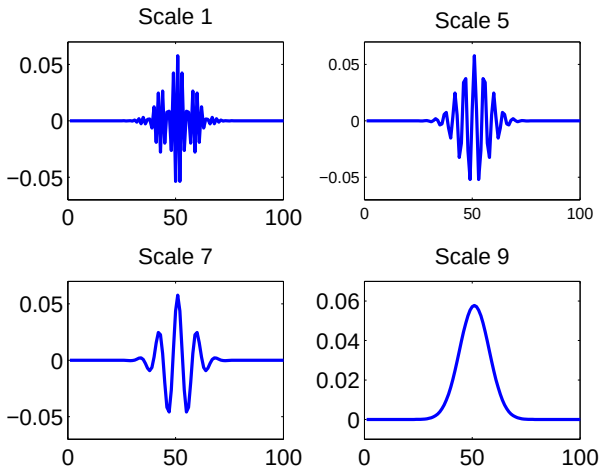


Fig. 1. The real parts of the uniform Gabor filters.

The Gabor filters [8] were developed to create Gaussian transfer functions in the frequency domain. Thus, taking the inverse Fourier transform of this transfer function, we get a filter closely resembling to the Gabor filters. The Gabor filters have been shown to have optimal combined localization in both the spatial and the spatial-frequency domain [9, 10]. In certain applications, this filtering technique has been demonstrated to be robust and fast [11] and the recursive implementation of 1D Gabor filtering has been shown in paper [12]. This recursive algorithm for the Gabor filter achieves the fastest possible implementation. For a signal consisting of N samples, this implementation requires $O(N)$ multiply-and-add (MADD) operations. A generic one dimensional Gabor function and its Fourier transform are given by:

$$h(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{t^2}{2\sigma^2}) \exp(j2\pi F_i t), \qquad (1)$$

$$H(f) = \exp(-\frac{(f - F_i)^2}{2\sigma_f^2}), \qquad (2)$$

where $\sigma_f = 1/(2\pi\sigma)$ represents the bandwidth of the filter and F_i is the central frequency.

The Gabor filter can be viewed as a Gaussian modulated by a complex sinusoid (with centre frequencies F_i). This filter responds to some frequency, but only in a localized part of the signal. The coefficients of Gabor filters are complex. Therefore,

the Gabor filters have one-side frequency support as shown in Fig. 2 and Fig. 4. We also illustrate the real parts of the Gabor filters in Fig. 1 and Fig.3.

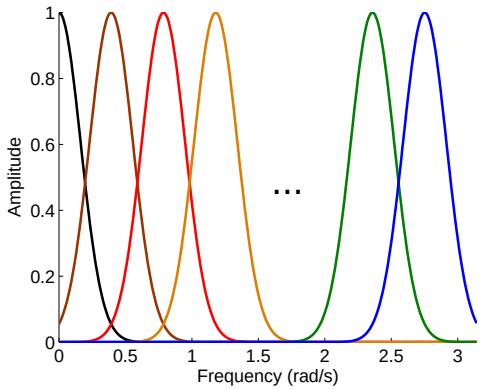


Fig. 2. Frequency supports of the uniform Gabor filters.

Given a certain number of subbands, in order to obtain a Gabor filter bank, the central frequencies F_i and bandwidths σ_f of these filters are chosen to ensure that the half-peak magnitude supports of the frequency responses touch each other as shown in Fig. 2 and Fig.4. The Gabor filter bank can be designed to be uniform (in Fig. 2) or non-uniform (in Fig. 4). In our experiments, we use the Gabor filter bank with nine non-uniform subbands.

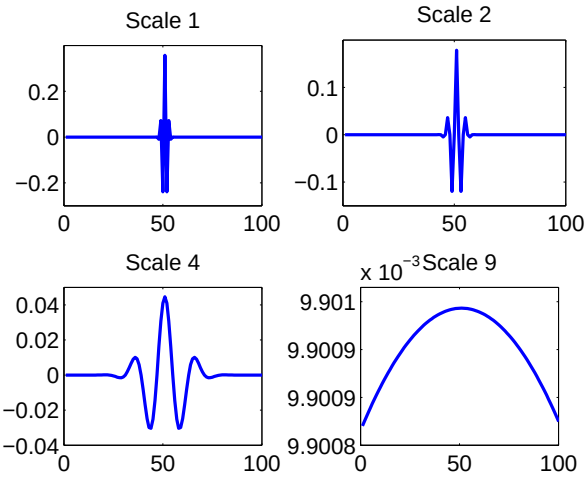


Fig. 3. The real parts of the non-uniform Gabor filters.

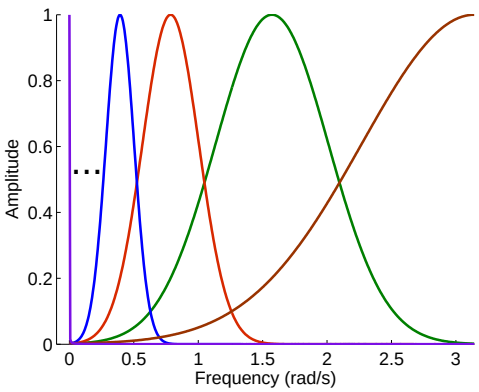


Fig. 4. Frequency supports of the non-uniform Gabor filters.

After decomposing a MS signal, nine subbands are created as follows:

$$y_i(t) = h_i(t) * x(t), \tag{3}$$

where $x(t)$ is the input signal, i = 1,2,...,9, and $*$ is the 1D convolution. This is an over-complete representation with the redundant ratio of 9.

2.2. GaborLocal Method

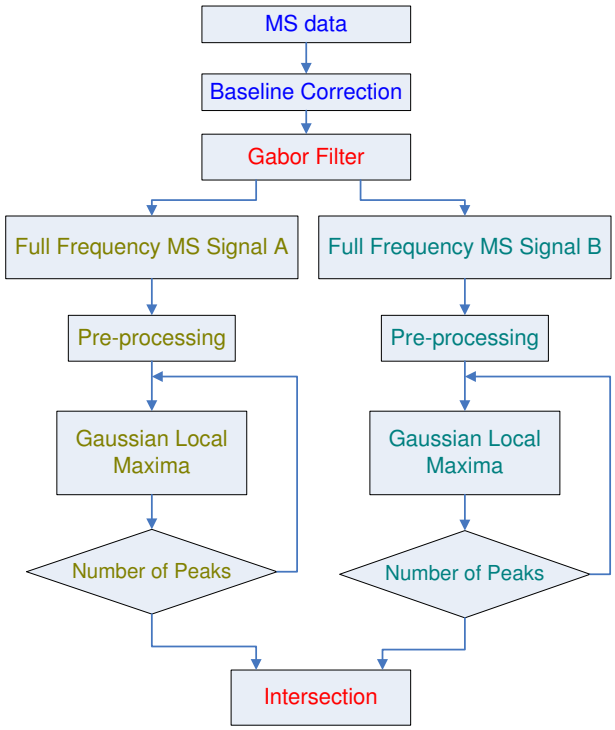


Fig. 5. Flowchart of Gabor-Gaussian local maxima method for peak detection in the MS data.

Our main idea is to amplify the true signal and compress the noise of mass spectrum by using the Gabor filter bank. After that, we use the Gaussian local maxima to detect peaks and the maximum rank of peaks which will be defined later to quantify peaks. This method is named as Gabor filter - Gaussian local maxima (GaborLocal). Fig. 5 is the flowchart of our GaborLocal method. The GaborLocal can be detailed into the four steps including the full frequency MS signal generation, the peak detection, the peak quantification, and the intersection.

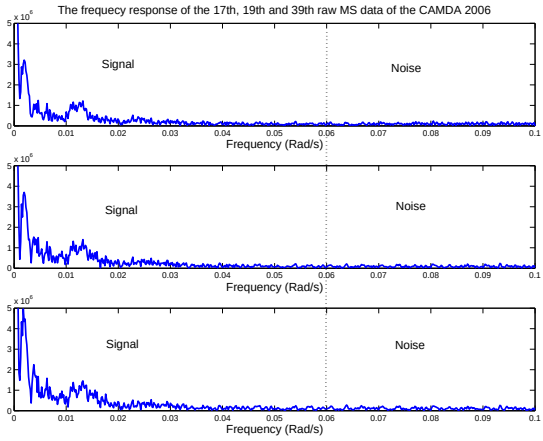


Fig. 6. The frequency response of three raw MS signals - the 17^{th}, 19^{th} and 39^{th} MS data of the CAMDA 2006.

2.2.1. *Full frequency MS signal generation*

Mass spectrum is decomposed to many scales by using the Gabor filters after the baseline correction. Our purpose is to emphasize some hidden peaks buried by noise. When we analyze 60 MS signals of the CAMDA 2006 in the frequency domain, we notice that the valuable information of these signals locate from zero to around 0.06 (rad/s), and the noises locate from 0.06 to π (rad/s). The frequency responses of three raw MS data (the 17^{th}, 19^{th} and 39^{th} MS data of the CAMDA 2006) are shown in Fig. 6 as an example. Therefore, the bandwidth σ_f of the Gabor filters which enhances peaks must be less than 0.06. In our experiments, we use $\sigma_f = 0.01$. If the uniform Gabor filter is used, the number of scales must be

$$N = \frac{\pi}{0.01} \approx 314 \; scales. \tag{4}$$

With 314 scales in (4), we know that the uniform Gabor filter is not efficient. If the non-uniform Gabor filter is used, the number of scales should be calculated as follows:

$$\sigma_f = \frac{\pi}{2^N},$$

$$N = log_2(\frac{\pi}{\sigma_f}),$$

$$N \approx 8.3 \; scales \; with \; \sigma_f = 0.01. \tag{5}$$

Based on the Eq. (5), we use the non-uniform Gabor filters with 9 scales to decompose the MS data (we use CAMDA 2006 data [13] for experiments). If we transform $y_i(t)$, $h_i(t)$ and $x(t)$ in Eq. (3) into the frequency domain, we get

$$Y_i(f) = X(f).H_i(f), \tag{6}$$

where $X(f)$ is the frequency response of the raw MS signal, $H_i(f)$ is the frequency response of the i^{th} Gabor filter, and $Y_i(f)$ is the frequency response of the i^{th} scale. After getting 9 signals according to 9 frequency sub-bands in complex values, the full frequency signal A will be created by summing above signals in complex values first and taking their absolute values at the final. To create the full frequency signal B, we take the absolute values for each sub-band and then sum all these sub-bands. After this step, we have two full frequency signals A and B. Let's denote $y(t)$ and $Y(f)$ as the full frequency signal in time domain and frequency domain, respectively.

$$Y(f) = \sum_{i=N_i} Y_i(f), \tag{7}$$

where N_i are the scales which are used to create the full frequency signal. From Eq. (6) and (7), we get

$$Y(f) = \sum_{i=N_i} X(f)H_i(f)$$

$$= X(f) \sum_{i=N_i} H_i(f) = X(f)H_s(f), \tag{8}$$

where $H_s(f) = \sum_{i=N_i} H_i(f)$ is called the summary filter. From Eq. (2), the summary filter can formulated as follows

$$H_s(f) = \sum_{i=N_i} \exp(-\frac{(f-F_i)^2}{2\sigma_f^2}). \tag{9}$$

Our purpose in this step is to amplify the true signal and compress the noise. The black line in the Fig. 7 is $H_s(w)$ which can amplify the true signal from 0 to 0.06 $\frac{rad}{s}$ and compress noise from 0.06 to π. In this

case, if we use $N_i = [1\ 2\ ...\ 9]$ we can get the summarized filter represented by the blue line in Fig. 7. The Fig. 9 shows the frequency response of the 19^{th} raw MS signal (blue line) and that of full frequency signal (red line). We can see that the signal from 0 to 0.06 is amplified and the noise from 0.06 to π is compressed. Therefore, in both full frequency MS signal A and B, all peaks have been emphasized to help the next peak detection step. In this step, baseline correction is also used and is detailed as follows

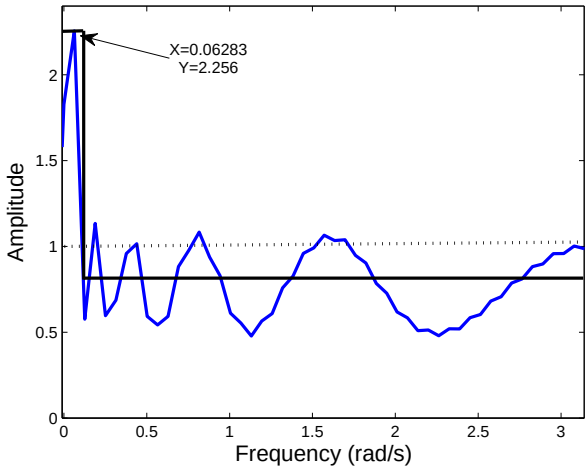


Fig. 7. The frequency response of the summary filter.

Baseline correction The chemical noise or the ion overloading is the main reason causing a varying baseline in mass spectrometry data. Baseline correction is an important step before using Gabor filter to get the full frequency MS signals. The raw MS signal x_{raw} includes some real peaks x_p, the baseline x_b, and the noise x_n.

$$x_{raw} = x_p + x_b + x_n. \tag{10}$$

The baseline correction is used to remove the artifact x_b. In this paper, we use 'msbackadj' function of MATLAB to remove baseline. The msbackadj function estimates a low-frequency baseline first which is hidden among the high-frequency noise and the signal peaks and then subtracts the baseline from the spectrogram. This function follows the algorithms in Andrade *et al.*'s paper [14].

Illustration In order to understand this step easier, one example of the way to create full frequency MS signal is shown in Fig. 8.

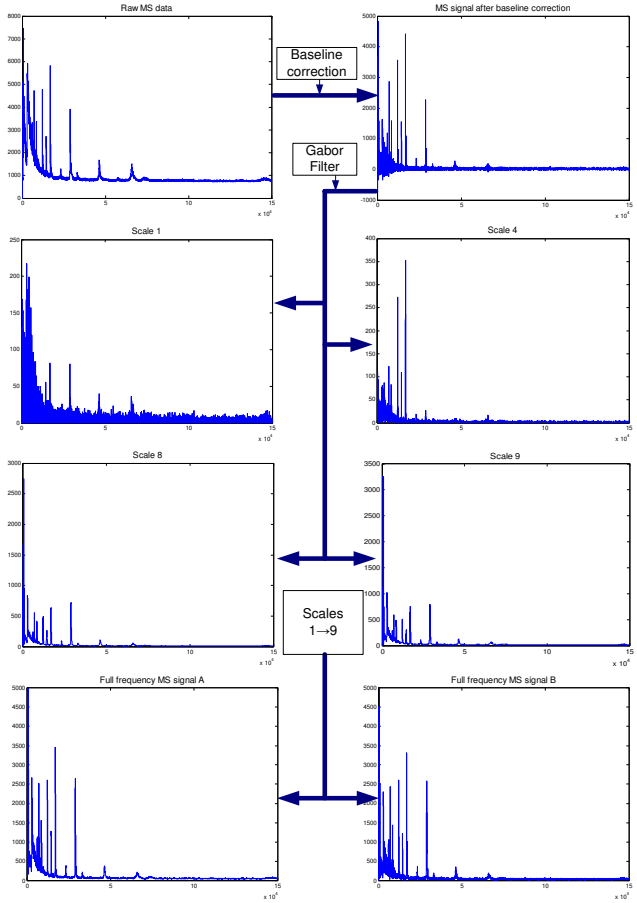


Fig. 8. One example of the step named full frequency MS signal generation. Raw MS data is the 19^{th} MS signal of CAMDA 2006.

In this example, the 19^{th} MS signal of CAMDA 2006 is chosen as raw MS data. After the baseline correction, MS signal is used as the input of the Gabor filters. A Gabor filter bank with 9 non-uniform sub-bands is employed to create 9 MS signals with 9 different frequency sub-bands. In Fig. 8, the signals of scale 1, 4, 8 and 9 are visualized. Some noises in high frequency are separated from the MS signal of the scale 1, 2, ..., 5. In the MS signal under the scales 6, ..., 9, all high intensity peaks are still kept. After combining the MS signals of all scales in two ways, the full frequency MS signal A and B are created. The comparison between the raw MS and full frequency signal in frequency domain is shown in Fig. 9. This figure shows our purpose which amplifies the important signal and compresses the noise

has been achieved. We should remember that this is just a compression of noise instead of removing noise. As the outputs, two full frequency MS signal A and B will be used to detect peaks in the next step instead of raw MS data.

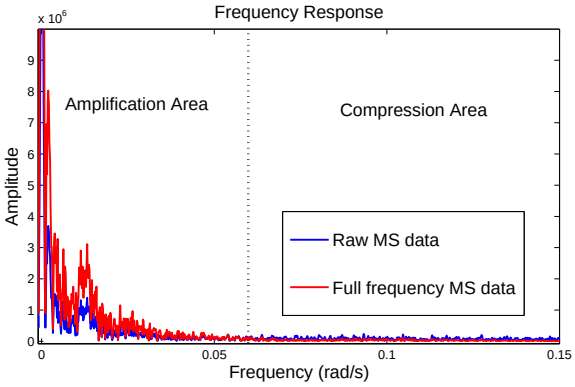


Fig. 9. The frequency response of the 19^{th} MS signal of CAMDA 2006 before and after using the summary filter.

2.2.2. *Peak detection by Gaussian local maxima*

All peaks are detected as many as possible by using Gaussian local maxima with the full frequency MS signal A as well as the full frequency MS signal B. The Gaussian local maxima is used instead of local maxima because Gaussian local maxima is robust with noise in peak detection. Before detecting peaks, pre-processing step is also applied such as elimination peaks in the low-mass region. Now, the Gaussian local maxima will be introduced as follows

Gaussian local maxima We assume that we want to find local maxima of $y(t)$. We should follow two steps: computing derivative of $y(t)$ and finding zero crossing. The derivative of $y(t)$ is approximated by the finite difference as follows:

$$\frac{d(y(t))}{dt} = \lim_{h->0} \frac{y(t+h) - y(t)}{h} \approx y(t+1) - y(t). \quad (11)$$

At $t = t_0$, if the derivative of $y(t)$ equals to zero and has a change from positive to negative or from negative to positive, we have zero-crossing. If the derivative of $y(t)$ changes from positive to negative at t_0, we have local maxima at t_0. With discrete signal, (11) can be rewritten as follows

$$\frac{d(y(n))}{dn} = y(n+1) - y(n) = y(n) * [1 \quad -1]. \quad (12)$$

Unfortunately, MS data always have noise. Thus, we assume that Gaussian filter $g(t, \sigma)$ is used to handle the denoise in MS data (this is not a denoising step). Finally, derivative of $y(t) * g(t, \sigma)$ will replace the derivative of $y(t)$ as follows

$$\frac{d(y(t) * g(t,\sigma))}{dt} = \frac{d(\int (y(\tau).g(t-\tau,\sigma)d\tau))}{dt}$$

$$= \int (y(\tau).\frac{d(g(t-\tau,\sigma))}{dt}d\tau) = y(t) * \frac{d(g(t,\sigma))}{dt}, \quad (13)$$

where

$$g(t,\sigma) = \exp(-\frac{t^2}{2\sigma^2}). \quad (14)$$

Taking the derivative of $g(t, \sigma)$ in (14), we have

$$\frac{d(g(t,\sigma))}{dt} = \frac{-t}{\sigma^2}\exp(-\frac{t^2}{2\sigma^2}). \quad (15)$$

From (13) and (15), we have

$$\frac{d(y(t) * g(t,\sigma))}{dt} = y(t) * (\frac{-t}{\sigma^2}\exp(-\frac{t^2}{2.\sigma^2})). \quad (16)$$

Instead of finding zero crossing of $\frac{d(y(t))}{dt}$, we find zero-crossing of $\frac{d(y(t)*g(t,\sigma))}{dt}$ by (16). With discrete signal, (16) can be rewritten as follows

$$\frac{d(y(n) * g(n,\sigma))}{dn} = y(n) * v(n), \quad (17)$$

where $v(n)$ is listed in the table 1. Using Gaussian filters makes the Gaussian local maxima method more robust with noise.

2.2.3. *Peak quantification by maximum rank*

After detecting many peaks in full frequency MS signals, a new signal is obtained from these peaks. This new signal will be the input of the next peak detection loop where the Gaussian local maxima method is also applied. Then, many loops are repeated until the number of peaks obtained is less than a threshold. Now, we define the maximum rank of peaks as follows:

Maximum rank We assume n loops are used and get m_1 peaks at the loop 1, m_2 peaks at loop 2,...and m_n peaks at the loop n. We have $m_1 >$

Table 1. The value of vector v(n) with different lengths.

length	n = 1	2	3	4	5	6	7	8	9	10
5	0.0007	0.2824	0	-0.2824	-0.0007					
6	0.0007	0.1259	0.7478	-0.7478	-0.1259	-0.0007				
7	0.0007	0.0654	0.6572	0	-0.6572	-0.0654	-0.0007			
8	0.0007	0.0388	0.4398	0.6372	-0.6372	-0.4398	-0.0388	-0.0007		
9	0.0007	0.0254	0.2824	0.7634	0	-0.7634	-0.2824	-0.0254	-0.0007	
10	0.0007	0.0180	0.1851	0.6572	0.5329	-0.5329	-0.6572	-0.1851	-0.0180	-0.0007

Table 2. Definition of maximum rank of peaks. Y means that the peak can be detected at that loop. N means that the peak can not be detected at that loop. The peak with the maximum rank equaling to 1 is able to be detected at all of the loops. The peak with the maximum rank equaling to n only appeared at the first loop.

Maximum Rank	*Loop* 1	*Loop* 2	*Loop* 3	*Loop* 4	... *Loop* $(n-1)$	*Loop* n
1	Y	Y	Y	Y	... Y	Y
2	Y	Y	Y	Y	...Y	N
...	...	...	...	...	...	...
n	Y	N	N	N	...N	N

$m_2 > ... > m_n$. Maximum peak (MR) is defined as the table 2.

We have m_n peaks with $MR = 1$, $m_{n-1} - m_n$ peaks with $MR = 2$,...and $m_1 - m_2$ peaks with $MR = n$. In our algorithm, the probability of the true peaks with $MR = i$ is higher than with $MR > i$.

Demonstration Fig. 10 shows an example of the step named the peak quantification by using the maximum rank. First, the full frequency MS signal A is used to detect peaks by using Gaussian local maxima. At the loop 1, we can detect 1789 peaks. From these 1789 peaks, we create a new signal with 1789 positions. At the next loops 2, 3, 4, we can detect 509, 143, 39 peaks, respectively. At the loop 5, 15 peaks can be detected. Because we choose a threshold of 16 and *number of peaks* $= 15 < 16$, we stop at the loop 5. Actually, we can select the threshold from 38 to 16 and also get 15 peaks at the final loop. Now, we get 15 peaks with $MR = 1$, $39 - 15 = 24$ peaks with $MR = 2$, $143 - 39 = 104$ peaks with $MR = 3$, $509 - 143 = 366$ peaks with $MR = 4$ and $1789 - 509 = 1280$ peaks with $MR = 5$. In this case, we only keep 15 peaks with $MR = 1$. We also do the same on the full frequency MS signal B and can get 12 peaks with $MR = 1$ at the last loop.

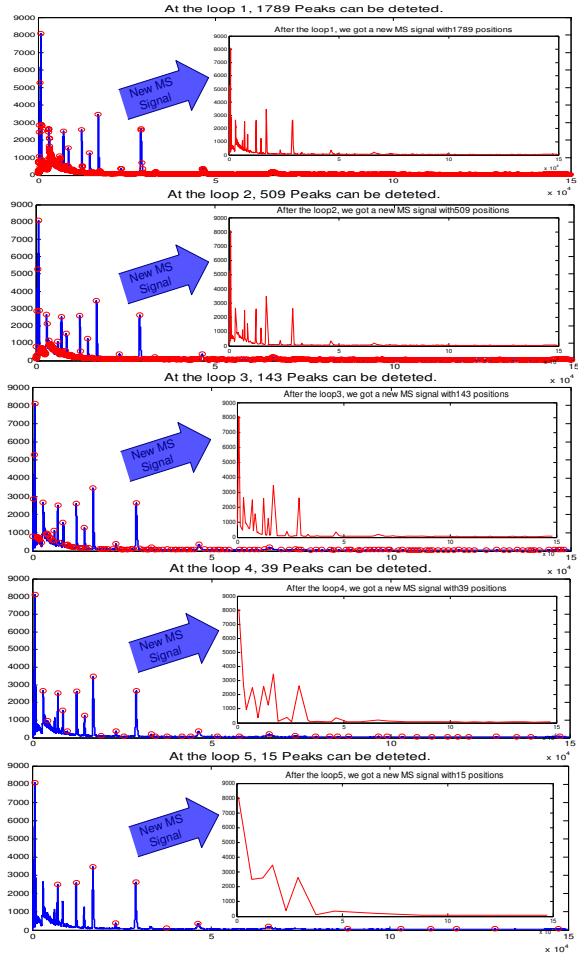


Fig. 10. One example of the step named peak detection and quantification.

2.2.4. *Intersection*

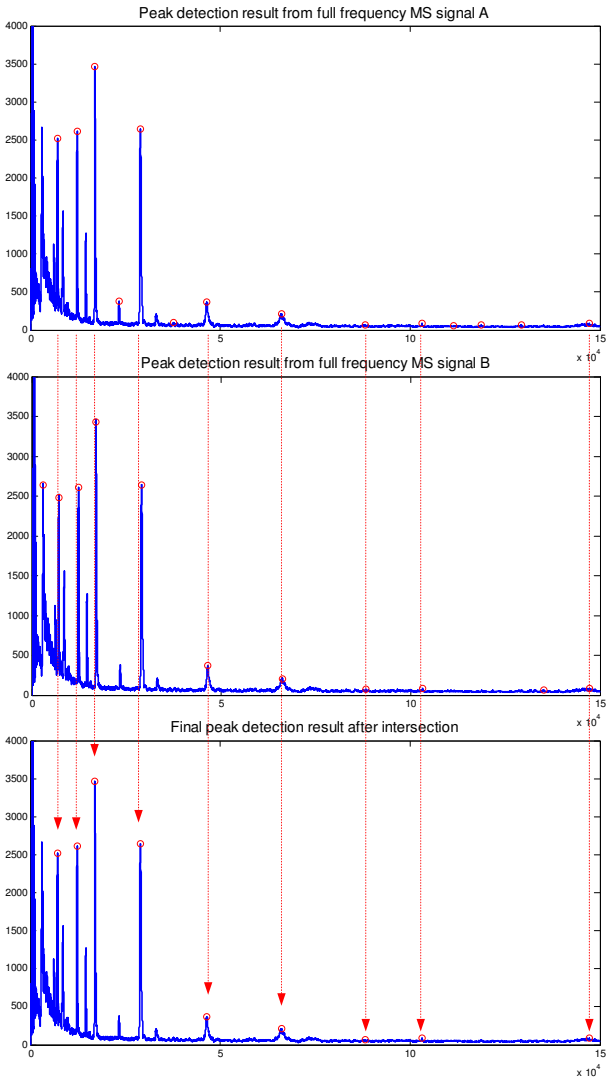


Fig. 11. One example of the step named intersection.

Now, we have two results of peak detection from two full frequency MS signals. The intersection of two above results will be the final result. For example, Fig. 11 shows how to do the intersection of two results. We have 15 peaks in the signal A and 12 peaks in the signal B but we just get 9 peaks as the final result. With this result, we get 7 true peaks and 2 false peaks. This result shows that the true position rate equal to $\frac{7}{7} = 1$ and the false position rate equal to $\frac{2}{9} \approx 0.22$.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, our GaborLocal method will be compared to two other most common used methods: the Cromwell [4, 5] and the CWT [6]. We will evaluate the performance of three methods by using the ROC curve that is the standard criterion in this area.

3.1. Cromwell Method

Cromwell method is implemented as a set of MATLAB scripts which can be downloaded from [15]. The algorithms and the performance of the Cromwell were described in [5, 4]. The main idea of the Cromwell method can be summarized as follows

(a) Denoise the individual spectrum using the undecimated discrete wavelet transform. The hard thresholding method was used to reset small wavelet coefficients to zero. In these papers, the authors used the median absolute deviation (MAD) to estimate the thresholding.
(b) Estimate and remove the baseline artifact by using a monotone local minimum curve on the smoothed signal.
(c) Normalize the spectrum by dividing the total ion current, defined to be the mean intensity of the denoised and baseline corrected spectrum.
(d) Identify peaks by using local maxima and signal to noise ratio (SNR).
(e) Match peaks across spectrum and quantify peaks using either the intensity of the local maximum or computing the area under the curve for the region defined to be the peaks.

3.2. CWT Method

The algorithm of CWT method has been implemented in R (called as 'MassSpecWavelet') and the Version 1.4 can be downloaded from [16]. This method was proposed by Pan Du *et al.* [6] in 2006 and can be summarized as follows:

(a) Identify the ridges by linking the local maxima. Continuous wavelet transform (CWT) is used to create many scales from one mass spectrum. The local maxima at each scale is detected. The next step is to link these local maxima as lines.

(b) Identify the peaks based on the ridge lines. There were three rules to identify the major peaks. They are the scale with the maximum amplitude on the ridge line, the SNR being larger than a threshold and the length of ridge being larger than a threshold. We should notice that the SNR is estimated in the wavelet space. This is a nice motivation of this method.
(c) Refine the peak parameter estimation.

3.3. Evaluation Using ROC Curve

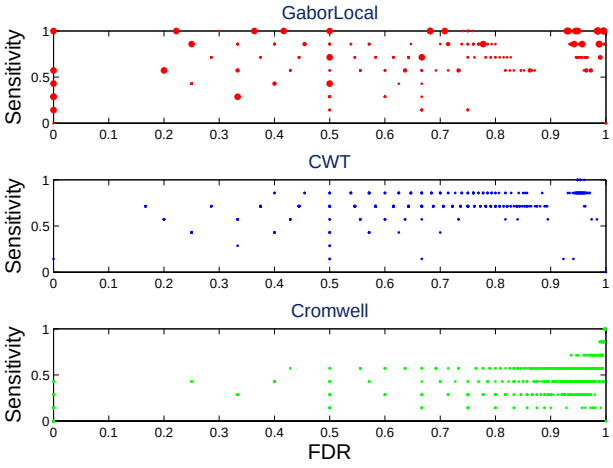


Fig. 12. Detailed receiver operating characteristic (ROC) curves obtained from 60 MS signals using Cromwell, CWT, and our GaborLocal method. The sensitivity is the true position rate.

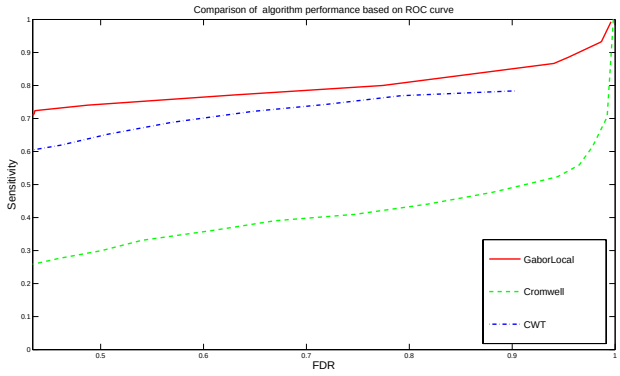


Fig. 13. Average receiver operating characteristic (ROC) curves obtained from 60 MS signals using Cromwell, CWT, and our GaborLocal method. The sensitivity is the true position rate.

The CAMDA 2006 dataset [13] of all-in-1 Protein Standard II (Ciphergen Cat. # $C100-007$) is used to evaluate three algorithms: the Cromwell, the CWT, and our method. Because we know polypeptide composition and position, we can estimate the true position rate (TPR) and the false position rate (FPR). Another advantage of this dataset is that they are real data and better than the simulated data in evaluation.

The TPR is defined as the number of identified true peaks divided by the total number of true peaks. The FPR is defined as the number of falsely identified peaks divided by the total number of identified peaks. We call an identified peak as true peak if it is located within the error range of 1% of the known m/z value of true peaks. There are seven polypeptides which create seven true peaks at 7034, 12230, 16951, 29023, 46671, 66433 and 147300 of the m/z values. Fig. 12 shows the TPR and the FPR of three above methods with an assumption that there is only one charge. To calculate the ROC curve of Cromwell and CWT methods, the SNR thresholding values are changed. The SNR thresholding values are chosen from 0 to 20 for Cromwell method, from 0 to 65 for CWT method. In our GaborLocal method, the threshold of number of peaks is changed from 2000 to 10 to create the ROC curve. In the Fig. 12, the performance of Cromwell method is much worse than CWT and our GaborLocal methods. Most of ROC points of Cromwell method locate at the bottom of right corner and most of ROC points of CWT and GaborLocal methods are well placed on the top regions. In our method, some ROC points appear at the top line with $TPR = 1$ and some ROC points go with $TPR = 1$ and $FPR = 0$. However, it does not happen to the CWT. Therefore, GaborLocal is the best one.

If we take the average of those detailed ROC results of Fig. 12, we get the average ROC curve as the Fig. 13. We should notice that we take average of all ROC points with the same SNR threshold (for Cromwell and CWT) and with the same peak threshold (for our method). From the Fig. 13, the results of our method and CWT are much better than the Cromwell's one. Therefore, the decomposing approach without smoothing (both SWT and

GaborLocal) is more efficient than the denoising approach (like Cromwell). At the same FPR, the TPR of our method is consistently higher than the TPR of CWT. Because the maximum rank was used to identify peaks in the GaborLocal method instead of the SNR. It is clear that the utilizing maximum rank to identify peak gives out valuable results. This method has a significant contribution to detect both high energy and small energy peaks. The other advantage of this method is that the threshold of number of peaks can be created easier than the SNR. Therefore, the GaborLocal method is an more efficient and accurate method for real MS data peak detection.

3.4. Examples

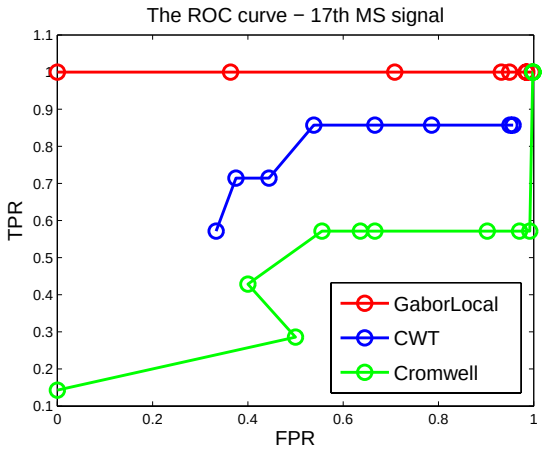


Fig. 14. The ROC curve of three methods such as Cromwell, CWT, and our method with the 17^{th} mass spectrum signal

Now, we study one example shown in Fig. 15 in which the 17^{th} spectrum signal of CAMDA 2006 dataset is picked and tested with three above methods. Fig. 15 (a) includes four sub-figures. The first sub-figure describes the real peak positions and raw data. The second and third sub-figure show the full frequency MS signal A&B with identified peaks. The last sub-figure is the final result after doing intersection.

We get 12 peak candidates from the full frequency MS signal A and 10 peak candidates from the full frequency MS signal B. Finally, we get 7 peak candidates after intersection of 12 and 10 peaks. In the result, our method can detect exactly 7 peaks over 7 true peaks. Fig. 15 (b) shows 9 detectable peaks from CWT method. Among 9 above peaks, there are only 5 true peaks. The CWT loses two peaks at 7034 and 147300 of the m/z values. Fig. 15 (c) shows the result of Cromwell's method. There are three true peaks being detectable by this method. Some peaks with low SNRs can not be detected. Of course, if we decrease the SNR threshold, more peaks can be detected. However, we also get more false peaks and the FPR will be increased dramatically. In general, if the thresholding values of three above methods are changed, we can get the ROC curve in Fig. 14. From this figure, the performance of our method keeps $TPR = 1$ with any value of the FPR (from 1 to 0). However, the TPR's values of Cromwell and CWT methods decrease very quickly when the FPR's value decreases. At the $FPR = 0$, the TPR of Cromwell method equals 0.1429. In CWT method, even the $FPR \approx 1$, the TPR only equals to 0.8571. The CWT and Cromwell methods are limited in peak detection performance because of the way using the SNR to identify peaks. Fig. 14 and Fig. 15, we can prove that

(1) Decomposition of MS data makes peak detection easier.
(2) Using SNR to identify peaks can not detect low SNR peaks.
(3) Using the MR can detect more true peaks than using the SNR.

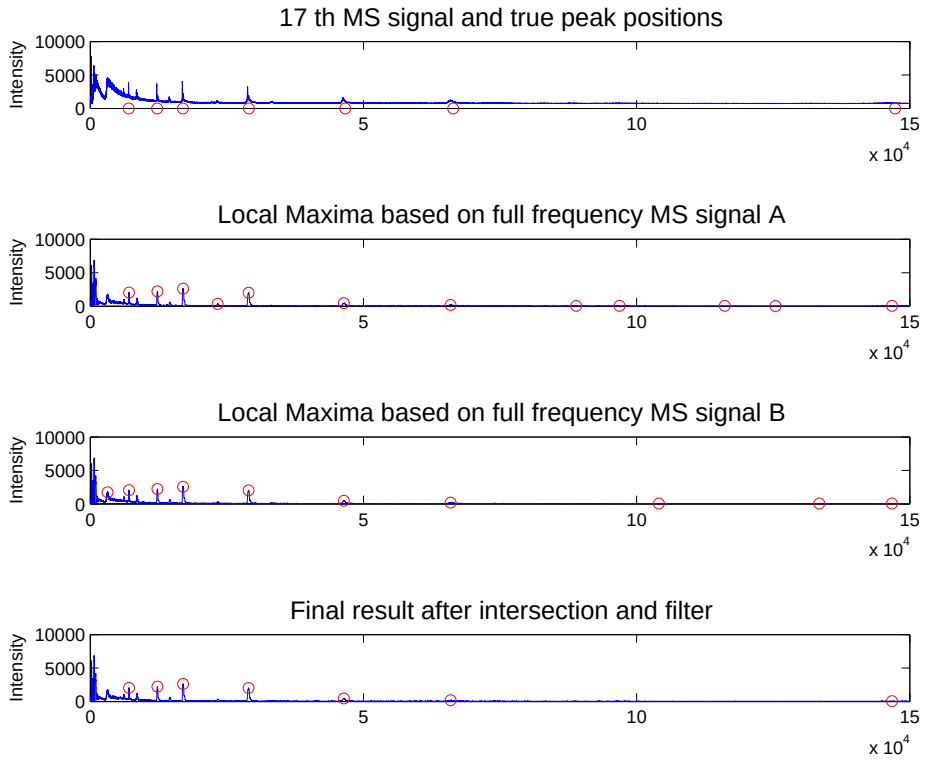
4. CONCLUSION

In this paper, we proposed a new approach to solve peak detection problem in MS data with promising results. Our GaborLocal method combines the Gabor filter with Gaussian local maxima approach. The maximum rank method is presented and used at the first time to replace the previous SNR method to identify true peaks. With real MS dataset, our method gave out a much better performance in the ROC curve compared to two other most common used peak detection methods. In our future work, we will develop new protein identification method based on our GaborLocal approach.

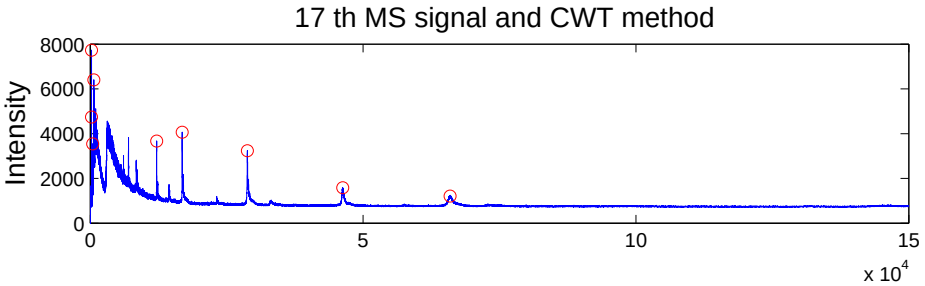
References

1. N. Jeffries, "Algorithms for alignment of mass spectrometry proteomic data," *Bioinformatics*, vol. 21, pp. 3066–3073, 2005.

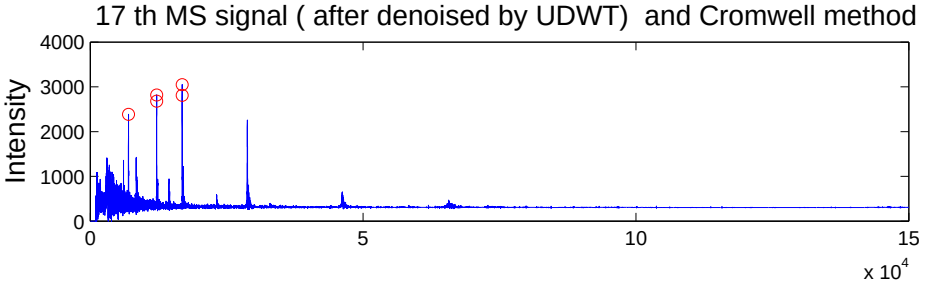
2. J. e. Li, "Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry," *Clin Chem*, vol. 51, pp. 2229–2235, 2005.
3. T. e. Rejtar, "Increased identification of peptides by enhanced data processing of high-resolution maldi tof/tof mass spectra prior to database searching," *Anal Chem*, vol. 76, pp. 6017–6028, 2004.
4. J. Morris, K. Coombes, J. Koomen, K. Baggerly, and R. Kobayashi, "Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum," *Bioinformatics*, vol. 21, no. 9, pp. 1764–1775, 2005.
5. K. Coombes and *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*, vol. 5, no. 16, pp. 4107–4117, 2005.
6. P. Du, W. Kibble, and S. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, 2006.
7. E. Lange and *et al.*, "High-accuracy peak picking of proteomics data using wavelet techniques," in *Processdings of Pacific Symposium on Biocomputing*, 2006, pp. 243–254.
8. D. Gabor, "Theory of communication," *J. Inst. Elec. Engr*, vol. 93, no. 26, pp. 429–457, Nov 1946.
9. J. Kamarainen, V. Kyrki, and H. Kalviainen, "Invariance properties of gabor filter-based features-overview and applications," *IEEE Transactions on Image Processing,*, vol. 15, no. 5, pp. 1088–1099, May 2006.
10. J. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A*, vol. 2, pp. 1160–1169, 1985.
11. C. L. D. Tsai, "Fast defect detection in textured surfaces using 1d gabor filters," *The International Journal of Advanced Manufacturing*, vol. 20, no. 9, pp. 664–675, Oct. 2002.
12. I. Young and M. G. L. Vliet, "Recursive gabor filtering," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2798–2805, Nov 2002.
13. C. C. F. S. R. Group, "Camda 2006 conference contest datasets." [Online]. Available: http://camda.duke.edu/camda06/datasets/index.html
14. L. Andrade and L. Manolakos, "Signal background estimation and baseline correction algorithms for accurate dna sequencing," *Journal of VLSI, special issue on Bioinformatics*, vol. 35, pp. 229–243, 2003.
15. U. M. A. C. Center, "The new model processor for mass spectrometry data." [Online]. Available: http://bioinformatics.mdanderson.org/cromwell.html
16. P. Du, "Mass spectrum processing by wavelet-based algorithms." [Online]. Available: http://bioconductor.org/packages/2.1/bioc/html/MassSpecWavelet.html



(a)



(b)



(c)

Fig. 15. Example of peak detection of the 17^{th} mass spectrum signal using Cromwell, CWT and our GaborLocal method.