BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 17 2006, pages 2059–2065
doi:10.1093/bioinformatics/btl355

Genome analysis

Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching

Pan Du[1], Warren A. Kibbe[1] and Simon M. Lin[1,*]

[1]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL 60611, USA

Received on April 24, 2006; revised on June 22, 2006; accepted on June 23, 2006
Advance Access publication July 4, 2006
Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: A major problem for current peak detection algorithms is that noise in mass spectrometry (MS) spectra gives rise to a high rate of false positives. The false positive rate is especially problematic in detecting peaks with low amplitudes. Usually, various baseline correction algorithms and smoothing methods are applied before attempting peak detection. This approach is very sensitive to the amount of smoothing and aggressiveness of the baseline correction, which contribute to making peak detection results inconsistent between runs, instrumentation and analysis methods.

Results: Most peak detection algorithms simply identify peaks based on amplitude, ignoring the additional information present in the shape of the peaks in a spectrum. In our experience, 'true' peaks have characteristic shapes, and providing a shape-matching function that provides a 'goodness of fit' coefficient should provide a more robust peak identification method. Based on these observations, a continuous wavelet transform (CWT)-based peak detection algorithm has been devised that identifies peaks with different scales and amplitudes. By transforming the spectrum into wavelet space, the pattern-matching problem is simplified and in addition provides a powerful technique for identifying and separating the signal from the spike noise and colored noise. This transformation, with the additional information provided by the 2D CWT coefficients can greatly enhance the effective signal-to-noise ratio. Furthermore, with this technique no baseline removal or peak smoothing preprocessing steps are required before peak detection, and this improves the robustness of peak detection under a variety of conditions. The algorithm was evaluated with SELDI-TOF spectra with known polypeptide positions. Comparisons with two other popular algorithms were performed. The results show the CWT-based algorithm can identify both strong and weak peaks while keeping false positive rate low.

Availability: The algorithm is implemented in R and will be included as an open source module in the Bioconductor project.

Contact: s-lin2@northwestern.edu

Supplementary material: http://basic.northwestern.edu/publications/peakdetection/. Colour versions of the figures in this article can be found at *Bioinformatics Online*.

1 INTRODUCTION

Peak detection is one of the important preprocessing steps in Mass Spectrometry (MS)-based proteomic data analysis. The performance of peak detection directly affects the subsequent process, such as profile alignment (Jeffries, 2005), biomarker identification (Li *et al*., 2005) and protein identification (Rejtar *et al*., 2004). However, owing to the complexity of the signals and multiple sources of noise in MS spectrum, high false positive peak identification rate is a major problem, especially in detecting peaks with low amplitudes (Hilario *et al*., 2006).

We are particularly interested in surface enhanced laser desorption ionization-time of flight (SELDI-TOF) spectroscopy, which is utilized in clinical and cancer proteomics (Petricoin *et al*. 2004). In contrast to MS/MS identification of proteins, SELDI-TOF is usually used to detect the differential expressions of intact proteins in different samples. Peak detection is a first step to identify the regions of interest. Currently, most of the peak detection algorithms identify peaks by searching local maxima with a local signal-to-noise ratio (SNR) over a certain threshold. The estimation of SNR is usually dependent on the peak amplitude relative to the surrounding noise level. However, high amplitudes do not always guarantee real peaks: some sources of noise can result in high amplitude spikes. Conversely, low amplitude peaks can still be real. In order to reduce the false positive rate, peak detection algorithms impose different constraints. Although the application of these constraints decreases the false positive rate of the algorithm, it also decreases the sensitivity of the method, resulting in undetected peaks.

The baseline removal and smoothing are two preprocessing steps of MS data. Usually they are performed before peak detection. There are quite a few baseline removal and smoothing algorithms available (Hilario *et al*., 2006). However, the results of these algorithms are not consistent. Baggerly *et al*. (2004) showed that different preprocessing algorithms could severely affect downstream analysis. Moreover, the baseline removal and smoothing are unrecoverable, i.e. if a real peak is removed during these preprocessing steps, it can never be recovered in the subsequent analysis. By adopting the Continuous wavelet transform (CWT)-based pattern-matching algorithm, the baseline can be implicitly removed and no smoothing is required. That means the algorithm can be directly applied to the raw data and the results will be more consistent for different spectra.

For MS data, 'true' peaks have characteristic shapes and patterns, some of which are determined by the geometric construction of the instrument (Gentzel *et al*., 2003). Some algorithms have tried to take advantage of the peak width (Gras *et al*., 1999) by setting a peak width range to reduce the false positive rate. This will be helpful in the simple cases, but for peaks with complex patterns and noise, the peak width estimation itself is difficult and the results are highly variable and dependent on sample composition. In this work, in order to take advantage of the additional information

[*]To whom correspondence should be addressed.

encoded in the shape of peaks, we perform peak detection by pattern matching in the wavelet space. Transforming into the wavelet space and making use of the additional shape information present in the wavelet coefficients can greatly enhance the effective SNR. As a result, the CWT-based method can detect weak peaks but maintain a low overall false positive rate.

The second difficulty of peak detection comes from the following observation: the width and height of 'true' MS peaks can vary a great deal in the same spectrum, for instance, peaks at high *m/z* value regions are usually wider and have lower amplitude. In addition, the shape of a peak can be altered because of the overlap of multiple peaks and noise. Thus, fixed pattern matching, like some matched filtering (Andreev *et al*., 2003) and deconvolution algorithms (Vivo-Truyols *et al*., 2005), will usually fail. The wavelet transformation provides a method for resolving these problems and has been widely used in signal processing and bioinformatics for multi-scale analysis of DNA sequence (Dasgupta *et al*.), protein sequence (Lio and Vannucci, 2000) and microarray temporal profile (Klevecz and Murray, 2001). In proteomics research, the wavelet transformation has been used for denoising (Coombes *et al*., 2005) and feature extraction (Qu *et al*., 2003; Randolph, 2005). Lange *et al*. (2006) recently proposed using the CWT in peak detection and peak parameter estimation. Their idea is to first decompose the spectrum into small segments, and then use the CWT at a certain scale to detect peaks and estimate peak parameters. However, it is not easy to select the right CWT scales for different segments of a spectrum before analyzing the data.

In order to build a robust pattern matching method, we also applied the CWT in MS peak detection. In contrast to the algorithm proposed by Lange *et al*. (2006), we directly apply the CWT over the raw spectrum and utilize the information over the 2D CWT coefficients matrix, which provides additional information on how the CWT coefficients change over scales. By visualizing the 2D CWT coefficients as a false color image, the ridges in the image can be correlated with the peaks in the MS spectrum, and this provides an easy visualization technique for assessing the quality of the data and the ability of the method to resolve peaks. Therefore, instead of directly detecting peaks in the MS spectrum, the algorithm identifies ridges in the 2D CWT coefficient matrix and utilizes these coefficients to determine the effective SNR and identify peaks. By identifying peaks and assigning SNR in the wavelet space, the issues surrounding the baseline correction and data are both removed, since these preprocessing steps are not required.

As presented below, the algorithm was evaluated with MS spectra with known polypeptide compositions and positions. Comparisons with other two peak detection algorithms are also presented. The results show that for these spectra the CWT-based peak detection algorithm provides lower false negative identification rates and is more robust to noise than other algorithms.

2 METHODS

In this section, we will briefly introduce the CWT, describe the algorithm for identifying the ridges over the 2D CWT coefficients and define the SNR in the wavelet space. Finally, we specify a robust rule set for peak identification.

2.1 Continuous wavelet transform

Wavelet transformation methods can be categorized as the discrete wavelet transform (DWT) or the CWT. The DWT operates over scales and positions based on the power of two. It is non-redundant, more efficient and is sufficient for exact reconstruction. As a result, the DWT is widely used in data compression and feature extraction. The CWT allows wavelet transforms at every scale with continuous translation. The redundancy of the CWT makes the information available in peak shape and peak composition of MS data more visible and easier to interpret. The change in the CWT coefficients over different scales provides additional information for pattern matching. In addition, there is no requirement for an orthogonal wavelet in the CWT. The CWT is widely used in pattern matching, such as discontinuity and chirp signal detection (Carmona *et al*., 1998). Mathematically, the CWT can be represented as (Daubechies, 1992):

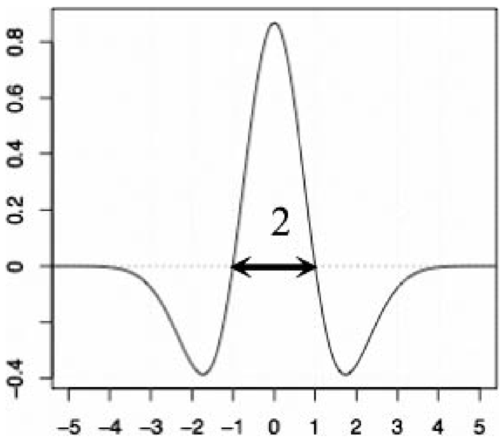

Fig. 1. Mexican Hat wavelet.

$$C(a,b)=\int_R s(t)\psi_{a,b}(t)\mathrm{d}t,\ \psi_{a,b}(t)=\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right),\ a\in R^+-\{0\},b\in R,\quad (1)$$

where $s(t)$ is the signal, a is the scale, b is the translation, $\psi(t)$ is the mother wavelet, $\psi_{a,b}(t)$ is the scaled and translated wavelet and C is the 2D matrix of wavelet coefficients.

Intuitively, the wavelet coefficients reflect the pattern matching between the signal s and $\psi_{a,b}(t)$. Higher coefficients indicate better matching. By changing the scale a, $\psi_{a,b}(t)$ can match the patterns at different scales without invoking more complicated non-linear curve fitting.

For peak detection, we examine the effect of changes in the width and height of peaks by the scaled and translated wavelet $\psi_{a,b}(t)$. In order to get better performance, the wavelet should have the basic features of a peak, which includes approximate symmetry and one major positive peak. In this work, we selected the Mexican Hat wavelet as the mother wavelet in the analysis (Daubechies, 1992). The Mexican Hat wavelet (Fig. 1) is proportional to the second derivative of the Gaussian probability density function. The effective support range of Mexican Hat wavelet is $[-5, 5]$. The Mexican Hat wavelet without scaling ($a = 1$) provides the best matches for the peaks with the width of about two sample intervals, as shown in Figure 1. With the wavelet scale increased to a_1, the peaks with bigger width, $2a_1$, provide the best matches. For the peaks in a MS spectrum, the corresponding CWT coefficients at each scale have a local maximum around the peak center. Starting from scale $a = 1$, the amplitude of the local maximum gradually increases as the CWT scale increases, reaches a maximum when the scale best matches the peak width, and gradually decreases later. In the 3D space, this is just like a ridge if we visualize the 2D CWT coefficients with the amplitude of the CWT coefficients as the third dimension. It transforms the peak detection problem into finding ridges over the 2D CWT coefficient matrix, a problem that is less susceptible to local minima and more robust to changes in coefficients in the search space. The peak width can be estimated based on the CWT scale corresponding to the maximum point on the ridge. And the maximum CWT coefficient on the ridge is approximately proportional to the area under curve (AUC) of the peak within the wavelet support region. AUC is the canonical way to identify the strength of a peak in spectral analysis. By looking at the ridge in the wavelet space, additional information about the shape and distribution of a putative peak can be obtained.

2.2 Removal of the baseline

With the adoption of the CWT-based pattern matching, peak detection can be directly applied over the raw data without preprocessing steps, including the baseline removal. Suppose each peak in the raw data, $P_{\text{raw}}(t)$, can be represented as follows:

$$P_{\text{raw}}(t) = P(t) + B(t) + C, \;\; t \in [t_1, t_2], \tag{2}$$

where $P(t)$ is the real peak, $B(t)$ is the baseline function with 0 mean, C is a constant and $[t_1, t_2]$ is the support region of the peak.

Based on Equation (1), we can calculate the CWT coefficients of the peak:

$$C(a,b) = \int_R P(t)\psi_{a,b}(t)\mathrm{d}t + \int_R B(t)\psi_{a,b}(t)\mathrm{d}t + \int_R C\psi_{a,b}(t)\mathrm{d}t, \tag{3}$$

where $\psi_{a,b}$ is the scaled and translated wavelet function.

As we assume the baseline is slow changing and monotonic in the peak support region, the baseline of the peak can be locally approximated as a constant C plus an odd function $B(t)$ defined in the peak support region and with the peak center as the original point. Because the wavelet function $\psi_{a,b}$ has a zero mean, the third term in Equation (3) will be zero. For symmetric wavelet function, like Mexican Hat wavelet, the second term will also approximately be zero. Thus, only the term with real peak $P(t)$ is left in Equation (3). That is to say, as long as the baseline is slowly changing and locally monotonic in the peak support region, it will be automatically removed in calculating the CWT coefficients.

2.3 Peak identification process

Figure 2 shows an example of the peak identification process. In order to provide a better visual image, we performed the CWT at 33 scale levels (from 1 to 64 at an interval of 2) directly over the raw MS spectrum. A segment of the computed 2D CWT coefficients are shown in false color in Figure 2b. The yellow color represents the high amplitude and green represents low. We can clearly identify the ridges in the 2D CWT coefficients matrix corresponding to the peaks in the raw spectrum (Figure 2a). The major peaks correspond to long and high ridges, while the small peaks correspond to short and low ridges. This provides a visual indication of peaks using ridges with different heights and lengths.

Identify the ridges by linking the local maxima The ridges can be identified by linking the local maxima of CWT coefficients at each scale level. First, the local maxima at each scale are detected. The identification of local maxima is similar to the method used in the PROcess R package in Bioconductor (www.bioconductor.org) (Gentleman, 2005). A sliding window is used, whose size is proportional to the wavelet support region at the scale. The next step is to link these local maxima as lines, which represent the ridges we are trying to identify.

Suppose the 2D CWT coefficient matrix is $N \times M$, where N is the number of CWT scales, and M is the length of the MS spectrum. The procedure of ridge identification is as follows:

(1) Initialize the ridge lines based on the local maxima points identified at the largest scale, i.e. row n $(n = N)$ in the CWT coefficient matrix, and set the initial gap number of ridge lines as 0.

(2) For each ridge line with its gap number less than a certain threshold, search the nearest maximum point at the next adjacent scale, row $n - 1$ in the coefficient matrix. The maximum allowed distance between the nearest points should be less than the sliding window size at that scale level. If there is no closest point found, the gap number of the ridge line is increased by one, or else the gap number is set to zero.

(3) Save the ridge lines having gap number larger than the threshold and remove them from the searching list.

(4) For the maxima points not linked to the points at the upper level, they will be initiated as new ridge lines.

(5) Repeat steps 2–4, until it reaches the smallest scale row $n = 1$ in the CWT coefficient matrix.

The identified ridge lines are shown in Figure 2c. The colors of the dots represent the relative strength of the coefficients. Blue represents high (the maximum CWT coefficients in the picture) and yellow represents close to zero. Comparing Figure 2b and c, we can see a high degree of correlation.

Definition of the signal to noise ratio Before defining the SNR in the wavelet space, signal and noise must be defined first. Based on the assumption that the real MS peaks have an instrument-specific characteristic shape, the signal strength of a peak is defined as the maximum CWT coefficient on the ridge line within a certain scale range. As for the noise, we assume that it is composed of positive or negative peaks with very narrow width. Since the baseline drift has been removed by the transformation into the wavelet space, the CTW coefficients at the smallest scale $(a = 1)$ are a good estimate of the noise level. The local noise level of a peak is defined as the 95-percentage quantile of the absolute CWT coefficient values $(a = 1)$ within a local window surrounding the peak. A minimum noise level can be provided to avoid the noise level close to zero, which could happen when some region is very smooth. Thus, the SNR is defined as the ratio of the estimated peak signal strength and the local noise level of the peak.

Identify the peaks based on the ridge lines Three rules are defined to identify the major peaks:

(1) The scale corresponding to the maximum amplitude on the ridge line, which is proportional to the width of the peak, should be within a certain range;

(2) The SNR should be larger than a certain threshold;

(3) The length of ridge lines should be larger than a certain threshold;

Usually, there are small peaks around the major peaks, which are commonly assigned as polypeptide adducts with the matrix molecules. These peaks have shorter ridge lines than the major strong peaks. By reducing the threshold of rule 3 in the surrounding area of major peaks, the small surrounding peaks can be easily identified. The proposed algorithm provides an option to do this. The algorithm also allows changing the threshold in rule 1 and 3 over the *m*/*z* value, which can better reflect the peak width changing over the *m*/*z* value.

The proposed algorithm provides two options to estimate the peak position. One way is following the ridge line from high scale to the small scale, the position at some small scale estimates the position of the peak maximum; another way is to estimate the peak centroid position based on the maximum CWT coefficient within certain scale range on the ridge line. The first method provides the results similar to other conventional peak detection algorithms. While the peak center estimation usually is more consistent over multiple spectra. In Figure 2a, the identified peaks with the default setting are marked as red circles at the peak maxima position, which also include the nearby small peaks of the strong peaks.

Refine the peak parameter estimation For the computational efficiency, peak identification is based on the CWT coefficients at selected scales. As a result, only the approximate peak strength, peak width and peak center position can be estimated. If a better estimate of peak parameters is required, the CWT needs to be performed over refined scales. Since we have already estimated the approximate peak parameters, for each identified peak, additional calculations of CWT only need to be performed over defined segments of MS spectrum (twice the support range of the CWT wavelet at the largest scale) surrounding the estimated peak center, and over the refined CWT scales. Other steps in peak identification are performed as previously.

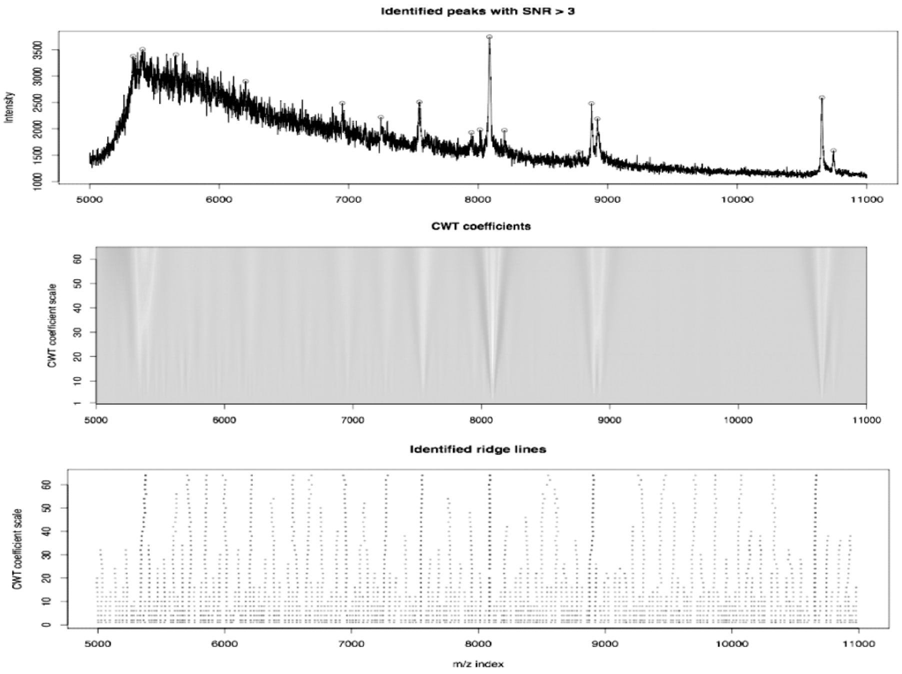

Fig. 2. Peak identification process based on CWT. (**a**) The raw MS spectrum. (**b**) The CWT coefficient image (yellow represents high amplitude, green represents low.) (**c**) The identified ridge lines based on CWT coefficient image. The colors of the dots represent the relative strength of the CWT coefficients. Blue represents high (the maximum CWT coefficients in the picture) and yellow represents close to zero.

Finally, the refined peak parameter estimation can be obtained using the additional calculations.

3 RESULTS

A reference MS dataset of known polypeptide compositions and positions was used to evaluate the algorithm, since it provides the opportunity to determine the false positive and false negative peak detection rates. The CAMDA 2006 dataset (CAMDA, 2006, http://camda.duke.edu) of All-in-1 Protein Standard II (Ciphergen Cat. # C100-0007) was the reference dataset. The MS spectra were measured on Ciphergen NP20 chips. There are seven polypeptides in the sample with the *m*/*z* values of 7034, 12 230, 16 951, 29 023, 46 671, 66 433 and 147 300. Figure 3 shows the result of one MS spectrum. In Figure 3a, the identified peaks are marked with red circles; the vertical lines represent the known positions of the seven polypeptides with multiple changes. By comparing the vertical lines and the identified peaks, we can see the algorithm identified six of the seven polypeptides with both one and two charges, except for the one at the very high end of the spectrum (*m*/*z* = 147 300) which is undetectable because of the low laser energy used in data acquisition. Also detected are three polypeptides with three charges and two polypeptides with four charges. Meanwhile, there is only one isolated false positive identification at 5184, which has relatively low SNR and could be a decomposition product or a contaminant in the sample. Figure 3a also shows an enlarged spectrum from 21 500 to 24 000. We can identify two peaks buried in the noise, which indicates that the algorithm, by utilizing the information available in the shape of the peak, can detect weak peaks without increasing the false discovery rate (FDR). Figure 3b shows the SNR values corresponding to the peaks in Figure 3a, with the SNR of the identified peaks shown as red color. We can see most of the identified peaks have much higher SNR than their surrounding peaks.

3.1 Comparison with other peak detection algorithms

Two algorithms were selected as benchmarks. One is the peak detection algorithm in the Bioconductor PROcess package. Another is based on the wavelet denoising method (Coombes *et al*., 2005). Both algorithms require preprocessing steps to remove the baseline prior to peak detection.

The PROcess peak detection algorithm incorporates a moving average method to smooth the spectrum, and estimates the median absolute deviation (MAD) within the sliding window at the time of smoothing; then it detects local maxima of the smoothed spectrum within the sliding window. The SNR is defined as the ratio of the smoothed value to the estimated MAD. The estimated SNR is shown in Figure 4b. These results show that the SNR estimation in the high *m*/*z* region is unreliable as there are no peaks in this region but the method resulted in high SNRs. In order to control the high false positive rate, other constraints are added, which include a filter with the minimum amplitude of a peak, and a threshold value of the peak area ratio which is defined as the peak AUC (within 0.3% range surrounding the peak) divided by the maximum peak AUC of the entire spectrum. The identified peaks by PROcess method based on the default settings are shown in Figure 4a. Weak peaks with *m*/*z* values in the range of 20 000–50 000 cannot be detected. By adjusting the peak area ratio threshold to 0.1, some of the peaks will be detected (see Supplementary Material)—however, the false positive rate will be increased in the *m*/*z* range of 3000–10 000. While it would be possible to incorporate a variable threshold that is responsive to the *m*/*z* value, these values

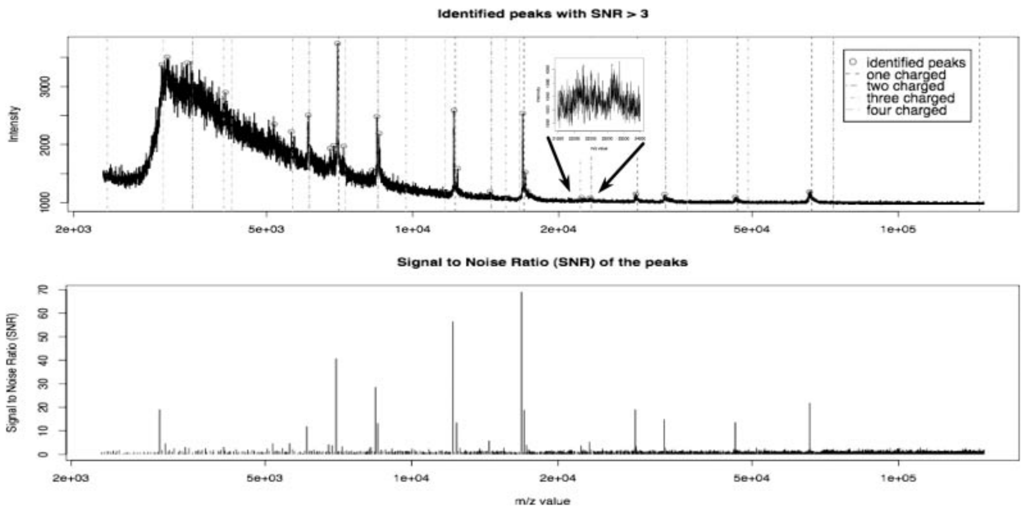

Fig. 3. Evaluate the identified peaks of CWT-based algorithm with known polypeptide positions. (**a**) Raw MS spectrum with peaks identified by CWT-based algorithm marked with red circles. Vertical dotted lines indicate the location of known polypeptides with different charge number; (**b**) SNR of the peaks calculated in CWT-based algorithm with identified peaks plotted as red.

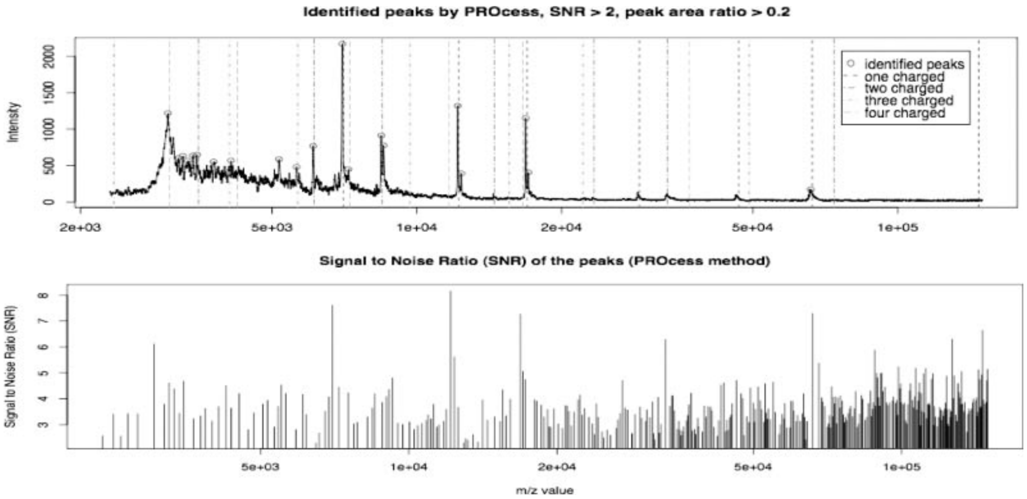

Fig. 4. Evaluate the identified peaks by PROcess method with known polypeptide positions. (**a**) Baseline removed and smoothed MS spectrum with peaks identified peaks; (**b**) Estimated SNR; other annotations are the same as Figure 3.

cannot be easily calculated for a specific spectrum and will not resolve all the issues of reproducibly calling peaks throughout the spectrum.

The wavelet denoising peak detection method, which removes noise based on the undecimated DWT decomposition, smoothes the spectrum. Thus the noise is the difference between the smoothed and the raw spectrum, and the noise level is defined as the MAD of the noise within a sliding window. The estimated SNR and identified peaks by wavelet denoising method are shown in Figure 5. The baseline removal algorithm in Coombes *et al*. (2005) assumes that the baseline is a monotonic, decreasing function. The baseline corrected spectrum starts at 3054 *m*/*z* value. Using the same visualization methods, we can see this method has good estimation performance in the high *m*/*z* region, although it has an increased false positive rate in the low *m*/*z* region. The algorithm uses a global threshold in the denoising procedure (Coombes *et al*., 2005). As a result, the noise with high amplitude, as in Figure 5b, was not removed. A possible improvement for this method is to adopt an adaptive threshold instead of a global one during wavelet denoising process.

The results in Figures 3–5 indicate the CWT-based peak detection algorithm has much better performance in detecting both strong and weak peaks throughout the spectrum, while keeping the false positive rate very low. Comparing Figures 4a and 5a, we can also see the significant variability between different baseline removal methods and smoothing algorithms.

In order to more broadly assess the performance of the peak detection methods, we applied all three algorithms over 60 spectra (CAMDA, 2006), and estimated the sensitivity and FDR for each algorithm on each spectrum at multiple SNR thresholds. The sensitivity of the algorithm was defined as the number of identified true positive peaks divided by the total number of real peaks.

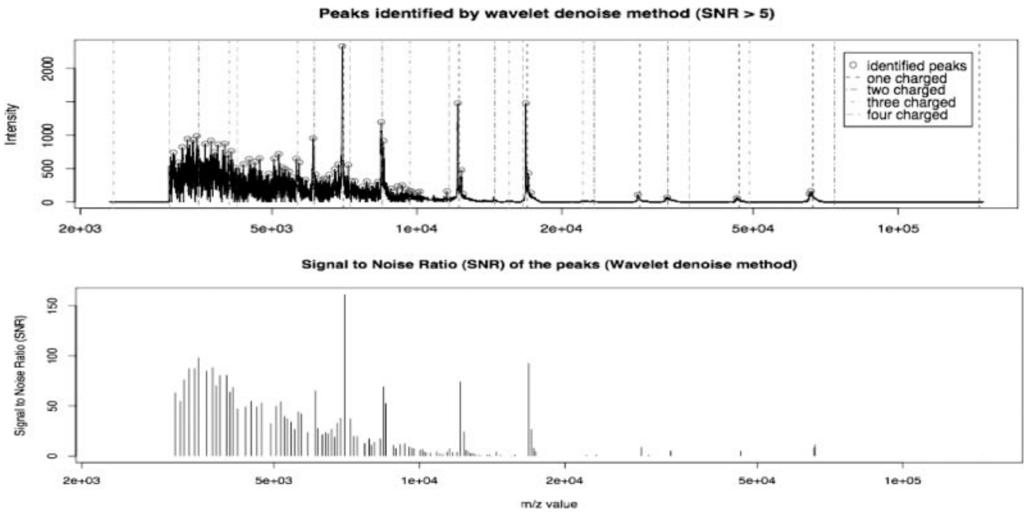

Fig. 5. Evaluate the identified peaks by wavelet denoising method with known polypeptide positions. (**a**) Baseline removed and smoothed MS spectrum. (**b**) Estimated SNR. Other annotations are the same as Figure 3.

(Each spectrum sample contained 7 polypeptides that resulted in 21 real peaks assuming up to 3 charges.) The FDR of each algorithm was defined as the number of falsely identified peaks divided by the total number of identified peaks. We call an identified peak as false peak if the identified peak is not located within the error range of ±1% of the known *m*/*z* values of real peaks. Finally a curve showing the FDR-sensitivity trade-off over the 60 spectra for the 3 methods is shown in Figure 6. (See Supplementary Material for the FDR-sensitivity relations before fitting the curve.) This curve is similar to the receiver operating characteristic curve. The point that represents the ideal performance is located at the upper left corner of the figure, i.e. it has 100% sensitivity and 0% FDR. For the PROcess peak detection algorithm, the peak area ratio constraint directly affects the sensitivity of the algorithm. In order to evaluate the algorithm over a wide range of sensitivity, we removed the peak area ratio constraint.

For these 60 spectra, Figure 6 demonstrates that the performance of the CWT-based algorithm is much better than the other benchmark methods. The CWT-based algorithm provides consistent high sensitivity at different FDRs. For the wavelet denoising based algorithm, because of the high false positive rate in spectra or regions of the spectra with high noise levels, it has lower sensitivity than the CWT method at a given FDR. Using these evaluation criteria, the performance of the PROcess peak detection algorithm was the worst among three for these spectra. The major reason is that the estimation of SNR is not robust, resulting in many false positives, as shown in Figure 4b.

Note that the real performance of all these algorithms should be better than that depicted in Figure 6. In reality, not all charge states (up to three) of the polypeptides exist, which result in the underestimation of the algorithm sensitivity. On the other hand, even in this controlled dataset, some peaks located near the major peaks may be polypeptide adducts with the matrix molecules, and some identified peaks may correspond to peaks of higher than three-charge states, e.g. four-charge state peaks shown in Figure 3a. So the FDR of the methods may also be overestimated in our scoring method.

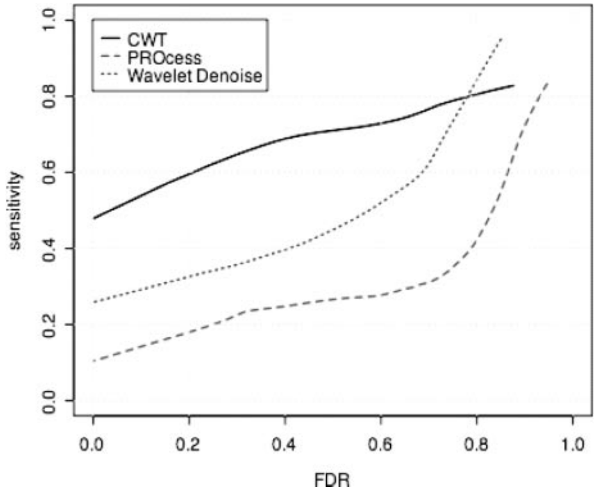

Fig. 6. Comparison of algorithm performance based on FDR-sensitivity relationship.

DISCUSSION

Peak detection is a critical step in MS data processing. The accurate detection of both strong and weak peaks throughout the spectrum is crucial for the success of the downstream analysis. Current peak detection algorithms cannot simultaneously identify strong and weak peaks without adversely affecting the false positive rate. The proposed algorithm identifies the peaks by applying CWT-based pattern matching, which takes advantage of the additional information present in the shape of the peaks, can greatly enhance the estimation of the SNR and robustly identify peaks across scales and with varying intensities in a single spectrum. Using this technique it is possible to detect both strong and weak peaks while still maintaining a high sensitivity and low FDR, as shown in the benchmarking results. Although, we only evaluated the algorithm by SELDI-TOF data, we believe the algorithm is also applicable to other types of mass spectrum.

The baseline removal and smoothing are preprocessing steps performed by current MS peak identification methods. Different baseline removal and smooth algorithms can produce different results, are sensitive to parameter settings and overall have a negative affect on the performance of further analysis. For the CWT-based peak detection algorithm, preprocessing steps are unnecessary. The ability to perform peak detection directly on the raw data increases the reliability of the detected peaks and simplifies the identification process. It reduces the potential variations in baseline removal and smoothing at the first step of large-scale analysis. This is important for the high-throughput proteomic analysis.

Because the CWT algorithm comparatively lacks of performance tuning parameters, it is easier to use and automate, and is comparatively more robust. Examples of applying the algorithm to other datasets with default parameters are available in the Supplemental Material.

Apart from identifying the peaks, the CWT-based algorithm also provides the estimation of the peak width and peak center position. These estimations are robust to noise. The estimation may not be accurate in the case of multiple peak overlapping or severely asymmetric peaks. The estimation of peak center of weak peaks may also be skewed when they are close to some strong peaks, which are shown as biased weak ridge lines near major peak lines in Figure 2. However, they still can be used as the initial guess when fitting some models to evaluate more precise peak parameter estimation, as was done in Lange *et al*. (2006).

One disadvantage of the algorithm is that the computational load is relatively high. It took ∼30 s to process one spectrum on a 1.67 GHz PowerPC G4 computer. This can be further improved by optimizing the algorithm and codes, and selecting several optimized CWT scales instead of tens of scales. The computational capabilities of newer 64-bit dual-core processors will also greatly reduce the processing time.

Although it was not our intention to use the wavelet transform to quantify peak intensity, we noticed that the estimated peak strength (as determined by the maximum CWT coefficient on the ridge line within a scale range) is proportional to the AUC of the peak in simple situations (neither severe overlap nor severe asymmetric). The AUC is widely used in spectroscopy, and for MS data it is believed to be a better quantification of signal than the amplitude of the peaks (Hilario *et al*., 2006). A natural next step in the utilization of this CWT method will be the incorporation of CWT-based estimates for quantization and analysis. Study of AUC estimation in spectra with multiple overlapping peaks should also be further pursued.

Conflict of Interest: none declared.

REFERENCES

Andreev,V.P. *et al*. (2003) A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. *Anal. Chem*., **75**, 6314–6326.

CAMDA (2006) CAMDA 2006 Competition Data Set.

Carmona,R., Hwang,W.-L. and Torrâesani,B. (1998) *Practical Time–Frequency Analysis: Gabor and Wavelet Transforms with an Implementation in S*. Academic Press, San Diego, CA.

Coombes,K.R. *et al*. (2005) 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*, **5**, 4107–4117.

Dasgupta,N. *et al*. (2004) Sequential modeling for Identifying CpG Island Locations in Human Genome. *IEEE Signal Proc. Lett*., **9**, 407–409.

Daubechies,I. (1992) *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Gentleman,R. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

Gentzel,M. *et al*. (2003) Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics*, **3**, 1597–1610.

Gras,R. *et al*. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis*, **20**, 3535–3550.

Hilario,M. *et al*. (2006) Processing and classification of protein mass spectra. *Mass Spectrom. Rev*., **25**, 409–449.

Jeffries,N. (2005) Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, **21**, 3066–3073.

Klevecz,R.R. and Murray,D.B. (2001) Genome wide oscillations in expression. Wavelet analysis of time series data from yeast expression arrays uncovers the dynamic architecture of phenotype. *Mol. Biol. Rep*., **28**, 73–82.

Lange,E., Gropl,C., Reinert,K., Kohlbacher,O. and Hildebrandt,A. (2006) High-accuracy peak picking of proteomics data using wavelet techniques. In *Proceedings of Pacific Symposium on Biocomputing 2006*, Maui, Hawaii, USA, pp. 243–254.

Li,J. *et al*. (2005) Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. *Clin. Chem*., **51**, 2229–2235.

Lio,P. and Vannucci,M. (2000) Wavelet change-point prediction of transmembrane proteins. *Bioinformatics*, **16**, 376–382.

Petricoin,E.F., Fishman,D.A., Conrads,T.P., Veenstra,T.D. and Liotta,L.A. (2004) Lessons from Kitty Hawk: from feasibility to routine clinical use for the field of proteomic pattern diagnostics. *Proteomics*, **4**, 2357–2360.

Qu,Y. *et al*. (2003) Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics*, **59**, 143–151.

Randolph,T.W. and Yasui,Y. (2006) Multiscale processing of mass spectrometry data. *Biometrics*, **62**, 589–597.

Rejtar,T. *et al*. (2004) Increased identification of peptides by enhanced data processing of high-resolution MALDI TOF/TOF mass spectra prior to database searching. *Anal. Chem*., **76**, 6017–6028.

Vivo-Truyols,G. *et al*. (2005) Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part I: peak detection. *J. Chromatogr. A*, **1096**, 133–145.