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
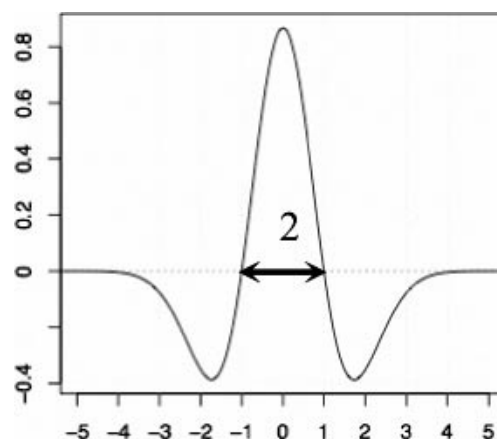

Fig. 1. Mexican Hat wavelet.

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):

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

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