International Journal of Mass Spectrometry 409 (2016) 53–58

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

journal homepage: www.elsevier.com/locate/ijms

An improved algorithm for peak detection in mass spectra based on continuous wavelet transform

Ying Zheng[a], Runlong Fan[b,*], Chunling Qiu[a], Zhen Liu[c], Di Tian[a,*]

[a] College of Instrumentation & Electrical Engineering, Jilin University, Changchun 130021, China
[b] SHRIMP Center, Institute of Geology Chinese Academy of Geological Sciences, Beijing 100037, China
[c] National Institute of Metrology, Beijing 100013, China

ARTICLE INFO

Article history:
Received 26 July 2016
Received in revised form
29 September 2016
Accepted 29 September 2016
Available online 4 October 2016

Keywords:
Mass spectrometry
Peak detection
Continuous wavelet transform
Overlapping peaks
Mass spectrum

ABSTRACT

Peak detection is an important preprocessing step in the analysis of mass spectrometry data. Currently, most peak detection methods have limited identification ability when there are overlapping or low-amplitude peaks. In this paper, we present an improved algorithm that combines the continuous wavelet transform (CWT) with the crazy climber algorithm. Particles move on the CWT coefficient matrix according to certain rules, and gradually gather at local maximal points. Peaks are identified by drawing ridges based on the weighted occupation measure of the grid points. The proposed method is evaluated using simulated noisy spectra and real spectra of benzene in nitrogen. The results show that the proposed approach is better at identifying overlapping peaks than existing methods. Receiver operating characteristic curves show that the proposed method can detect more true peaks while maintaining a low false discovery rate.

1. Introduction

Mass spectrometry (MS) is a powerful analytical technique that has been applied in scientific fields such as analytical chemistry [1], environmental sciences [2], and geology [3]. Peak detection is an important preprocessing step in the analysis of MS data. The performance of peak detection directly affects the analysis results. However, real experimental signals contain random noise, alternating baselines, differing peak shapes, and overlapping peaks. The adverse effects of these problems make the design of an automatic and accurate peak detection method complex [4].

Various peak detection methods have been developed, including the direct peak-located algorithm [5], first- and second-derivative methods [6,7], and curve fitting [8]. However, the simple direct peak-located algorithm and derivative method give poor results when the signal contains noise. Fixed-size curve fitting methods usually fail when the height and width of the MS peaks vary significantly in one signal. In recent years, the continuous wavelet transform (CWT)-based algorithm has been widely studied because of its accuracy, performance and multiscale nature. By transforming the spectrum into the wavelet space, the algorithm can take advantage of the additional information encoded in the shape of the peaks to reduce the false positive rate.

Du et al. applied the CWT in MS peak detection and directly carried out the algorithm over the raw spectrum to avoid the influence of baseline removal and smoothing [9]. By linking the local maxima in the CWT coefficient matrix, the locations of the peaks can be obtained. The CWT-based algorithm possesses the advantages such as robustness and consistency. However, it also presents some limitations in the identification of peaks with low amplitudes and overlapping peaks. Gregoire et al. proposed the concept of a mother ridge and descendant ridge to improve the algorithm's resolution of overlapping peaks [10]. Zhang et al. introduced a peak tree to represent the peak information, and presented an improved ant colony optimization biomarker selection method to provide higher sensitivity and fewer false detections [11]. These algorithms all use ridges to identify peaks, and ridge lines are determined based on the local maxima found by a sample sliding window. However, noise can produce similar local maxima, which will influence the method's true positive rate. In the case of overlapping peaks, the

Abbreviations: CWT, continuous wavelet transform; CWTC, continuous wavelet transform with crazy climber; MS, mass spectrometry; MALDI-TOF-MS, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry; TOF-MS, time-of-flight mass spectrometry; SNR, signal-to-noise ratio; EI, electron impact; FDR, false discovery rate; TPR, true positive rate; ROC, receiver operating characteristic.

* Corresponding authors.
E-mail addresses: fanrunlong2003@126.com (R. Fan), tiandi@jlu.edu.cn (D. Tian).

http://dx.doi.org/10.1016/j.ijms.2016.09.020
1387-3806/

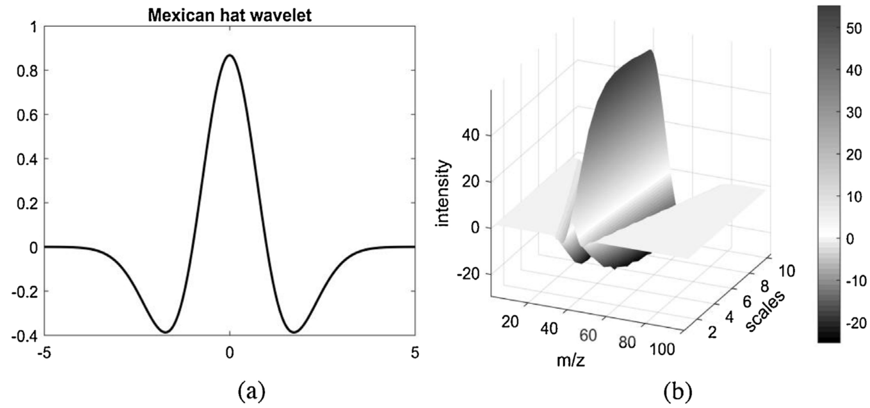

Fig. 1. (a) The Mexican hat wavelet. (b) The CWT coefficient matrix of a signal mass spectra peak simulated by Gaussian function.

amplitudes of local maxima produced by lower peaks will be weakened or even negative because of the influence of higher peaks. In this situation, the above algorithms cannot identify the lower peaks. There is another direction of the improvement of the CWT-based algorithm. Other feather points are introduced to improve the robustness and consistency. Nguyen applied zero-crossing lines in muti-scales of a Gaussian derivative wavelet of peak detection [12]. Zhang et al. located each peak with its ridge, valley and aero-crossing information in muti-scale wavelet space [4]. However, these algorithms are too complex to carry out, especially for overlapping peaks.

In the present study, we developed an algorithm named CWTC that combines CWT with a method called the "crazy climber" algorithm [13]. CWTC is designed in such a way that particles move on the CWT coefficient matrix according to certain rules, and weighted occupation densities form the ridges. The proposed method improves the particles' moving rules and occupation measure based on the characteristics of the mass spectrum, and combines a new search strategy for the ridges to accomplish peak detection. CWTC detects ridges according to the characteristics of the surface curve of the coefficient matrix rather than a simple numerical comparison, which improves its robustness to noise. Moreover, the algorithm can solve the problem of overlapping peaks to provide higher resolution.

This paper is organized as follows. First, the principle and procedures of our detection algorithm are presented in Section 2. CWTC is applied to both simulated matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) spectra and real time-of-flight mass spectrometry (TOF-MS) data to demonstrate its accuracy and effectiveness in Section 3. The results and a discussion of the applications are presented in Section 4, and we conclude the paper in Section 5.

2. Methods

2.1. Continuous wavelet transform

The CWT is an important time-frequency analytical tool which is broadly used in signal processing, image compression, and mathematical modeling. As indicated in Eq. (1), the CWT divides a continuous-time signal into scaled and translated wavelets to generate a 2-D coefficient matrix. By transforming to the time-frequency space, additional information is obtained and peak detection becomes simpler and more robust.

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

Where $s(t)$ is the signal, a is the scale, b is the translation, $\psi_{a,b}(t)$ is the scaled and translated mother wavelet and $C(a, b)$ is the 2-D coefficient matrix.

The Mexican hat wavelet (Fig. 1(a)) is usually selected as the mother wavelet because of the features of Gaussian MS peaks, such as approximate symmetry and one major positive peak. Fig. 1(b) is the CWT coefficient matrix of a signal MS peak simulated by a Gaussian function. Higher coefficients indicate better matching between the signal and the mother wavelet. The surface of the coefficient matrix generates a ridge perpendicular to the m/z axis around the location of the peak. By linking the ridge lines, the positions of spectral peaks can be located.

2.2. The crazy climber algorithm

2.2.1. Theory

The main idea of the crazy climber algorithm is as follows. The 2-D CWT coefficient matrix is considered as a grid plane. A large number of particles are initially and randomly seeded on the plane. Particles then move according to certain rules, and gradually gather at local maxima. Finally, ridges are drawn based on the grid points' weighted occupation measures.

The design goal of the moving rules is to position the particles on the ridges. As shown in Fig. 1(b), the ridge of the MS peak is perpendicular to the m/z axis. If we assume that the direction of the m/z axis is horizontal and the scale axis is vertical, the vertical motion of particles is the standard symmetric random walk, with reflection at the boundary. As a result, the particles are uniformly distributed along the direction of the scale axis. Horizontally, the particles are encouraged to "climb on the ridges". We introduce a mechanism similar to simulated annealing to prevent particles from becoming trapped around the local maxima produced by noise. Simulated annealing is a probabilistic technique for approximating the global optimum of a given function [14]. We simplify the method to reduce the computational complexity. First, a temperature value was initialized. The difference between the CWT coefficients is calculated when the particle moves downward, and the move is allowed if the difference is less than the temperature. Otherwise, the particle remains at its present location. The temperature is then gradually decreased to steady particles on the ridges.

After each combined horizontal and vertical movement, the particle occupancy of each grid point is measured. The superposition of all measures forms the initial measurement matrix. Considering that the peak amplitudes are usually higher than the background and noise, we use the original spectral signal intensity weighted by the initial measurement matrix. This heightens the measures of the peak positions, which improves the sensitivity of the algorithm.