Chemometrics and Intelligent Laboratory Systems 132 (2014) 75–81

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

An effective two-stage spectral library search approach based on lifting wavelet decomposition for complicated mass spectra

Cuiping Li [a,b,*], Jiuqiang Han [a], Qibin Huang [b], Baoqiang Li [b], Zhongyao Zhang [b], Chuntao Guo [c]

[a] Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, XianYang West Road #28, Xi'an PO 710049, China
[b] Beijing Institute of Pharmaceutical Chemistry, PO BOX 1044-403, Beijing PO 102205, China
[c] Beijing Persee General Instrument Co., Ltd., No.3 Pingsan Road, Pinggu District, Beijing PO 101200, China

ARTICLE INFO

Article history:
Received 9 July 2013
Received in revised form 31 December 2013
Accepted 5 January 2014
Available online 16 January 2014

Keywords:
Spectral library search
Search similarity
Compound identification
Lifting wavelet decomposition
Mass spectrometer

ABSTRACT

For rapid and accurate matching of complicated spectra in a standard spectral library with different instruments in various environments, a two-stage spectral library search approach that involves preliminary and main searches based on lifting wavelet decomposition was proposed. In the preliminary search stage, similar spectra were effectively extracted and the number of reference spectra was greatly reduced by using the low frequency component (outlines) of lifting wavelet decomposition to calculate spectral similarity. In the main search stage, search accuracy was improved by using the high-frequency component (details) of lifting wavelet decomposition to calculate spectral similarity. In addition, a self-adaptive method used to confirm the level of lifting wavelet decomposition according to the resolution (half-peak width) of the mass spectrometer was also presented. More than 100 sets of mass spectra as well as measured data of octafluoronaphthalene (OFN) samples analyzed by our instrument were used to compare the proposed method with common methods such as the weighted dot product similarity (WC), weighted dot product composite similarity (WRstC), and wavelet transform (WT) methods in terms of similarity, accuracy, and search time. Relative to the WRstC method, the average improvement of similarity for the proposed method was 15.1%. The results indicate that the proposed method improves the similarity and accuracy of spectral recognition for complicated mass spectra, particularly under conditions of poor spectral quality, owing to the low signal-to-noise ratio at low sample concentrations.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Gas chromatography–mass spectrometry (GC-MS) is an important and powerful tool that facilitates work in a variety of fields concerned with complex mixtures of organic compounds. One of the most important procedures for analyzing GC-MS data is chemical compound recognition, in which an experimental mass spectrum is assigned by means of spectral library search algorithms to a compound recorded in a reference spectral library [1].

Two stages are usually required for the spectral library search procedure: a preliminary search stage, and a main search stage for spectral matching. In the former, spectra similar to the target in all standard reference libraries must be extracted to reduce the number involved in matching and hence shorten the search time. A number of preliminary search methods can remove unwanted compounds but are likely to overlook some useful compounds. In contrast, others may avoid missing useful compounds but have low search efficiency. In the latter stage, an important step involves accurately matching a desired compound in the pre-search results. To optimize search efficiency, investigators have proposed various spectral library matching algorithms including the probability-based matching (PBM) [2], dot product or cosine [3–6], Hertz similarity search [7], normalized Euclidean distance [8,9], Grotch consistency criterion [10], and Knoch difference measurement [11] methods. Stein and Scott [12] and Horai et al. [13] evaluated these methods and concluded that the highest matching precision was obtained with the dot product similarity method. In light of the effects of peak intensity and peak position on compound recognition, a weighted dot product similarity (WC) method has also been proposed [12]; in addition, to consider the importance of the ratio between adjacent mass spectral peaks common to the sample and reference spectra, the weighted dot product similarity composite (WRstC) [12] and integrated [14] methods were proposed at around the same time. These algorithms showed higher similarity and precision than other matching algorithms. A hybrid algorithm that integrates distance and similarity algorithms has also been tested [15]. An algorithm that employs a wavelet transform (WT) to calculate spectral similarity shows somewhat better recognition accuracy than the traditional WC and WRstC [1,16].

The recognition accuracies of the spectral library matching methods mentioned above are continually being optimized and improved. The recognition accuracy, however, is highly dependent on the purity of the

* Corresponding author.
E-mail addresses: cuipingli86@126.com, lcplcp86@163.com (C. Li).

0169-7439/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2014.01.004

mass spectra, signal-to-noise ratio (SNR) of the mass spectrum, and instrument conditions. This often leads to the following shortcomings. First, the occurrence of false positives and negatives when searching for complicated spectra (e.g., weak signals with low SNR, certain interference peaks, and certain weak molecule or isotope peaks that represent compounds) can lead to widely varying results obtained by different spectral search methods [17]. Second, investigators may have difficulty in accurately recognizing desired compounds because so many similar compounds are included in the list of search results. Third, the search algorithms with high recognition accuracy are usually time-consuming and generally slow.

To overcome the limitations of current methods and find a fast and accurate library search algorithm, we developed a two-stage approach involving preliminary and main spectral search procedures based on the outlines and details, respectively, of lifting wavelet decomposition (LWD). First, a preliminary search was performed by calculating the dot product similarity on the basis of the low-frequency component (outlines) of LWD to narrow the number of reference spectra. Second, in the preliminary search results, the similarity was optimized by using the high-frequency component (details) of lifting wavelet decomposition to accurately recognize complex features such as adjacent mass peaks and weak peaks. Subsequently, the relationship between the wavelet decomposition levels and the mass spectrometer resolution was deduced. Finally, to verify the effectiveness of the algorithm, the proposed method was compared with three conventional methods–WC, WRstC, and WT–in terms of similarity, accuracy, and search time by using more than 100 spectral data sets in NIST08 as well as measurements obtained by instruments.

2. Theory and method

2.1. Measurement of weighted similarity of multi-variable characteristic vector

Suppose that the spectrum of a compound is expressed by the vector $X = (X_1, X_2, \ldots, X_n)$, of which X_n is the characteristic peak corresponding to $n = 1, 2, \ldots, N$ among the spectral signals. Here, we define m_n, i_n, and w_n as the mass unit (m/z), peak intensity, and half-peak width, respectively, which indicate features of the spectral peak. Some studies in the literature indicate that these features have different degrees of influence on the spectral search accuracy [17].

If $X_n = (i_n, m_n, w_n)$, it is expressed as follows after each feature is weighted:

$$X_n = (m_n)^a (i_n)^b (w_n)^c \tag{1}$$

where the weight coefficient is given by $p = (a,b,c)$, with a, b, and c being the weighting factors for the mass unit, peak intensity, and half-peak width, respectively.

Suppose U and L are the spectral vectors of the unknown compound and the reference spectra in the standard library, i.e., $U_n = (U_1, U_2, \ldots, U_n)$ $(n = 1, 2, \ldots, N)$ and $L_m = (L_1, L_2, \ldots, L_m)$ $(m = 1, 2, \ldots, M)$, respectively. Suppose $U_n^w = (m_{U_n})^a (i_{U_n})^b (w_{U_n})^c$ and $L_m^w = (m_{L_m})^a (i_{L_m})^b (w_{L_m})^c$. The weighted dot product similarity is then expressed as follows:

$$F_d^w = dot(U_n^w, L_m^w) = \langle \infty U_n^w, L_m^w \rangle / (|U_n^w| \cdot |L_m^w|). \tag{2}$$

The larger the values of F_d^w, the greater is the similarity between the unknown compound spectra and the reference spectra. In this paper, the half-peak width is not considered (i.e., $c = 0$), where $p = (a, b, 0)$. As for the values of a and b, Sokolow et al. [18] reported that the optimal value should be $p = (0.5, 1)$; thereafter, Stein and Scott [12] reported that the optimal value should be $p = (0.6, 3)$, and Horai et al. [13] suggested that the value should be $p = (0.5, 2)$. Recently, Kim et al. [19] pointed out that the weighting factor depends on the standard spectral database, and suggested that the value should be $p = (0.53, 1.3)$ in the latest database NIST11. The value of a ranges from about 0.5 to 0.6 and that of b is between 1 and 3.

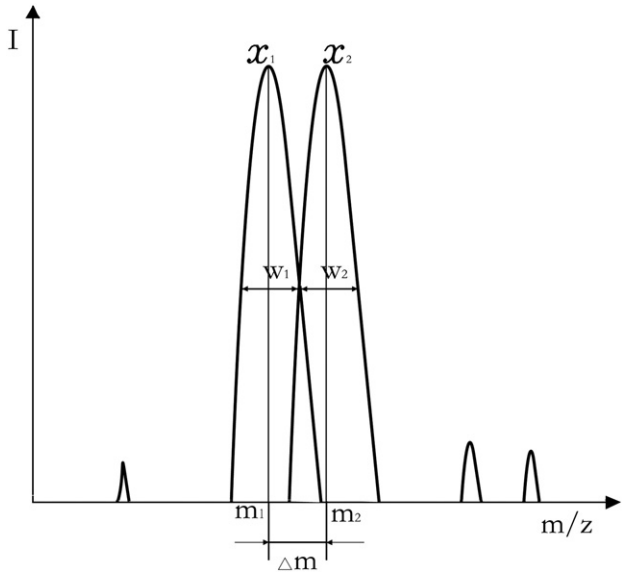

Fig. 1. Peak shapes of two adjacent spectral signals.

Considering the influence of the ratio of adjacent mass spectral peaks that are common between the unknown compound spectra and reference spectra, the optimized dot product similarity is given as follows [12]:

$$F_r = \frac{1}{K_{U\wedge L}} \sum_{k=1}^{K_{U\wedge L}} \left(\frac{L_k}{L_{k-1}} \frac{U_{k-1}}{U_k} \right)^{\alpha} \tag{3}$$

where F_r is the peak intensity ratio in the spectra, used to compare the consistency between the relative peak intensities in the unknown compound spectra and those in the reference spectra. Here, $K_{U\wedge L}$ is the number of common peaks between the unknown compound and reference spectra, when $\frac{L_k}{L_{k-1}} < \frac{U_{k-1}}{U_k}$, $\alpha = 1$, and when $\frac{L_k}{L_{k-1}} > \frac{U_{k-1}}{U_k}$, $\alpha = -1$.

On the basis of Eqs. (2) and (3), the matching factor *MF* of the composite similarity is calculated as follows:

$$MF = \beta \times \frac{K_U F_d^w + K_{U\wedge L} F_r}{K_U + K_{U\wedge L}} \tag{4}$$

where K_U is the number of peaks involved in calculating the unknown compound spectra and β (=100) is the normalization coefficient. The search results are arranged into a hit list used for analysis and confirmation by the analyst, ordered in terms of the *MF* from largest to smallest.

Table 1
Statistical results for each parameter after secondary pre-search (for a similarity threshold value of 0.6).

Level of wavelet decomposition L	The number of remaining reference spectra after pre-search	Time of spectral library search (s)
1	10	2.98 (3.8%)
2	13	3.12 (8.7%)
3	13	3.13 (9.0%)
4	14	3.19 (11.1%)
5	17	3.21 (11.8%)

Download English Version:

https://daneshyari.com/en/article/1179455

Download Persian Version:

https://daneshyari.com/article/1179455

Daneshyari.com