Chemometrics and Intelligent Laboratory Systems 142 (2015) 117–123

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

An efficient approach for compound identification based on the frequency features of mass spectra

Zhan-Li Sun [a,*], Kin-Man Lam [b], Jun Zhang [a]

[a] School of Electrical Engineering and Automation, Anhui University, Hefei, China
[b] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Received 1 August 2014
Received in revised form 28 January 2015
Accepted 30 January 2015
Available online 7 February 2015

Keywords:
Spectrum matching
Similarity measure
Discrete Fourier transform

ABSTRACT

Similarity-measure-based spectrum matching is an effective approach to chemical compound identification. When the sizes of both the query library and the reference library become increasingly large, most existing spectrum-matching methods encounter a seriously heavy computation burden. In this paper, an effective and efficient compound-identification approach is proposed based on the frequency features of mass spectra. Considering the sparsity of mass spectra, a nonzero feature-selection strategy is proposed to decrease the feature dimensionality of mass spectra. To further improve its efficiency, a correlation-based filtering strategy is presented to select the most correlated reference spectra in order to create a reduced reference library. Based on the decreased features and the reduced reference library, the frequency-feature-based composite similarity measures are computed to estimate the chemical abstracts service (CAS) registry numbers of the mass spectra blue in a query library. Due to the reduction in both the feature dimensionality and the reference library, the computation time of the proposed method is only about 6%–11% of that of the existing methods, while the identification performance remains sufficiently competitive. Experimental results demonstrate the feasibility and efficiency of the proposed method.

1. Introduction

High-throughput gas chromatography/mass spectrometry (GC/MS) has become an effective auxiliary measure to analyze chemical and biological samples in the petroleum industry, food sciences, biomedical sciences, and so on [1–3]. As one application, GC/MS data can be used to identify chemical compounds by means of statistical techniques. Similarity-based spectrum matching is a widely adopted approach for chemical-compound identification.

Chemical compounds are annotated with their chemical abstracts service (CAS) registry numbers. The mass spectra of those known chemical compounds are stored in a reference library. To identify an unknown chemical compound, spectrum matching is performed by computing the similarities between the query spectrum and all of the spectra in the reference library. Then, the reference spectrum corresponding to the largest similarity value is identified, and the CAS registry number of the query spectrum is considered to be that of the most similar reference spectrum in the reference library.

So far, various spectrum-matching algorithms have been developed for chemical-compound identification, e.g. composite similarity [4], cosine correlation [5], the dot-product algorithm [6], and so on. Usually, the peaks with large mass-to-charge (m/z) values have small peak intensities, but carry the most important characteristics for compound identification. Therefore, weighting the peak intensity on the basis of its m/z value can increase the relative significance of smaller peaks, as well as their contribution to compound identification [6]. In [4], a classical composite similarity measure was proposed by computing the cosine correlation with the weighted intensities. The information in the frequency domain is useful for pattern recognition [7,8]. Instead of using the ratios of the peak pairs [4], the dot product of the frequency features obtained via discrete Fourier transform (DFT) was adopted in [6] to measure the similarities of the query spectra to the reference spectra. Furthermore, the skewness and kurtosis of the similarity scores are considered in [9] to design the optimal weighting factors. In [10], the partial and semi-partial correlations, along with the various transformations of peak intensities, were proposed as the similarity measures for compound identification. In [11–13], a retention index is utilized to assist the determination of the CAS indices of unknown mass spectra. A partial set-covering model is developed in [14] for protein-mixture identification using mass spectral data. A statistical-relation-based search algorithm, namely X-Rank, was proposed in [15] to cope with the high variability of liquid chromatography tandem mass spectrometry (LC–MS/MS). Instead of taking

* Corresponding author. Tel.: +86 551 63861461.
E-mail address: zhlsun2006@126.com (Z.-L. Sun).

http://dx.doi.org/10.1016/j.chemolab.2015.01.019
0169-7439/

into account the absolute or the relative peak intensities, the X-Rank algorithm relies on a scoring model to differentiate between two fragmentation MS spectra. In [16], a very effective method was proposed to compress the hyperspectral data by combining random projection and principal component analysis (PCA). Then, the *k*-means clustering method is applied to the compressed data to segment the hyperspectral data.

The Fourier-transform-based spectrum-matching approach [6] is a novel and effective compound-identification algorithm. The simulations performed on the well-known standard mass spectral library of the National Institute of Standards and Technology (NIST) have demonstrated its prominent identification performance. Nevertheless, the heavy computation burden encountered in the algorithm is a significant and intractable problem, in particular when the sizes of both the query library and the reference library become increasingly large.

In this paper, an efficient compound-identification approach is proposed, which is based on the frequency features of mass spectra. Selecting effective features is important to the identification performance. Considering the sparsity of the peak intensities in mass spectra, a nonzero feature-selection strategy is proposed to decrease the feature dimensionality of mass spectra. Moreover, a two-stage similarity measure scheme is devised to reduce the computation burden of spectrum matching. In the first stage, for a query spectrum, a less accurate but more efficient similarity measure, Pearson's linear correlation coefficient, is used to select a few of the most correlated mass spectra from the reference library; these constitute a reduced reference library. In the second stage, a highly accurate but less efficient similarity measure, the frequency-feature-based composite similarity measure [6], is adopted to perform spectrum matching on the reduced reference library. Due to the reduction in both the feature dimensionality and the reference library, the computational complexity of the proposed method is significantly less than that of the existing methods. Experimental results on the NIST spectral library demonstrate the feasibility and efficiency of the proposed method.

The remainder of the paper is organized as follows. In Section 2, we present our proposed algorithm. Experimental results and discussions are given in Section 3, and concluding remarks are presented in Section 4.

2. Methodology

Let an $m \times p$ matrix $\mathbf{X}^0$ denote the spectra in a reference library, which includes m spectra with a dimensionality of p. The element x_{ij}^0 of $\mathbf{X}^0$ is the peak intensity corresponding to the jth mass-to-charge ratio (m/z) in the ith spectrum. Each row $\mathbf{x}_i^0(\mathbf{x}_i^0 = [x_{i1}^0, \cdots, x_{ip}^0],\ i = 1, \cdots, m)$ of $\mathbf{X}^0$ denotes a reference spectrum. Referring to notations in pattern recognition, here, the peak intensity x_{ij}^0 is considered as a one-dimensional feature of the spectrum $\mathbf{x}_i^0$. Therefore, the spectrum $\mathbf{x}_i^0$ is a p-dimensional feature vector in the spectrum-matching algorithm. Similarly, denote an $n \times p$ matrix $\mathbf{Y}^0$ as the query spectra in the query library. In the reference library, each reference spectrum is associated with a CAS registry number, which is used to denote the compound category. In contrast, the spectra without knowing the compound category are stored in the query library. The task of compound identification is to assign the correct CAS registry numbers to the query spectra using a spectrum-matching algorithm.

Fig. 1 shows the flowchart of our proposed method. There are three main parts in the proposed method: reduction of feature dimensionality, selection of the most correlated reference spectra and constitution of a reduced reference library, and extraction of frequency feature and computation of the similarity measure. A detailed description of these three parts is presented in the following subsections.

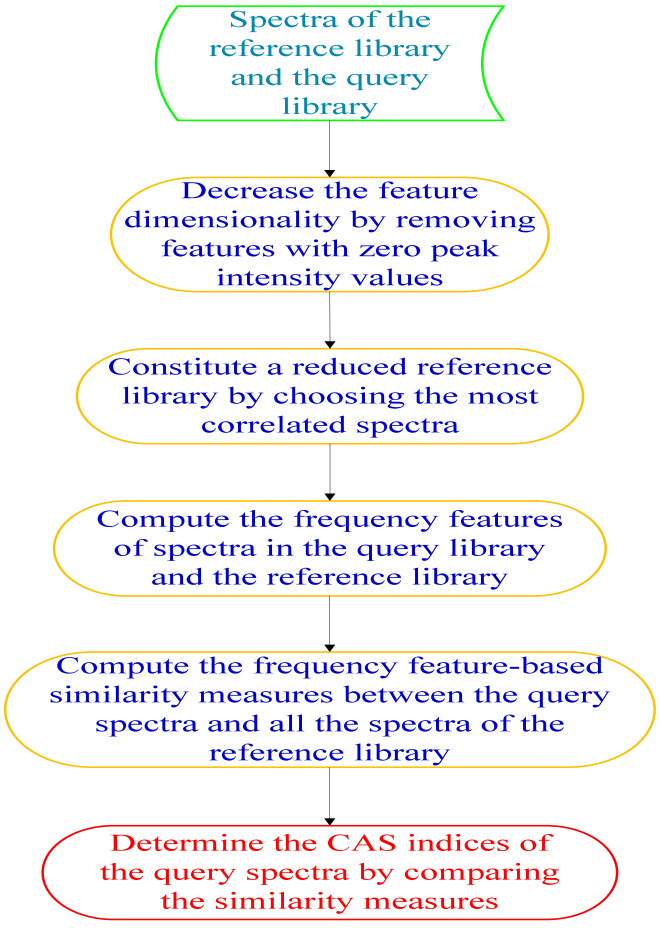

Fig. 1. Flowchart of the proposed frequency-feature-based compound identification method.

2.1. Nonzero feature selection

By observation, we found that the mass spectra are very sparse, i.e. most peak-intensity values are zero. In order to investigate the sparsity of a spectrum $\mathbf{x}(\mathbf{x} = (x_1, \cdots, x_p))$, we first compute the ℓ^0 norm of $\mathbf{x}$, i.e.

$$\|\mathbf{x}\|_0 = \sum_{j=1}^{p} I\left(x_j\right), \tag{1}$$

where $I(\cdot)$ denotes a nonzero indicator function such that

$$I(x) = \begin{cases} 1, & x \neq 0, \\ 0, & x = 0. \end{cases} \tag{2}$$

Furthermore, the sparse ratio (R_s), i.e. the ratio of the ℓ^0 norm and the dimensionality p of the spectrum, is defined to evaluate the spectrum sparsity, as follows:

$$R_s = \frac{\|\mathbf{x}\|_0}{p}. \tag{3}$$

Table 1 shows the sparse ratios of five mass spectra randomly selected from the NIST library. It can be seen that these mass spectra are very

Table 1
The sparse ratios (R_s) of five mass spectra randomly selected from the NIST library.

Sample	1	2	3	4	5
R_s	0.0488	0.0410	0.0193	0.0874	0.0259