Analytica Chimica Acta 853 (2015) 402–414

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Novel algorithm for simultaneous component detection and pseudo-molecular ion characterization in liquid chromatography–mass spectrometry

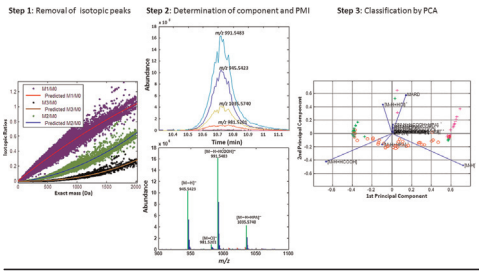
Yufeng Zhang [a], Xiaoan Wang [a], Siukwan Wo [a], Hingman Ho [b], Quanbin Han [b], Xiaohui Fan [c], Zhong Zuo [a,*]

[a] School of Pharmacy, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China
[b] School of Chinese Medicine, Hong Kong Baptist University, 7 Baptist University Road, Kowloon Tong, Hong Kong, China
[c] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, PR China

HIGHLIGHTS

- Novel stepwise component detection algorithm (SCDA) for LC–MS datasets.
- New isotopic distribution and adduct-ion models for mass spectra.
- Automatic component classification based on adduct-ion and isotopic distributions.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 29 May 2014
Received in revised form 25 September 2014
Accepted 2 October 2014
Available online 8 October 2014

Keywords:
Component detection
Pseudo-molecular ion
Mono-isotopic mass
Isotopic distribution
Natural products

ABSTRACT

Resolving components and determining their pseudo-molecular ions (PMIs) are crucial steps in identifying complex herbal mixtures by liquid chromatography–mass spectrometry. To tackle such labor-intensive steps, we present here a novel algorithm for simultaneous detection of components and their PMIs. Our method consists of three steps: (1) obtaining a simplified dataset containing only mono-isotopic masses by removal of background noise and isotopic cluster ions based on the isotopic distribution model derived from all the reported natural compounds in dictionary of natural products; (2) stepwise resolving and removing all features of the highest abundant component from current simplified dataset and calculating PMI of each component according to an adduct-ion model, in which all non-fragment ions in a mass spectrum are considered as PMI plus one or several neutral species; (3) visual classification of detected components by principal component analysis (PCA) to exclude possible non-natural compounds (such as pharmaceutical excipients). This algorithm has been successfully applied to a standard mixture and three herbal extract/preparations. It indicated that our algorithm could detect components' features as a whole and report their PMI with an accuracy of more than 98%. Furthermore, components originated from excipients/contaminants could be easily separated from those natural components in the bi-plots of PCA.

1. Introduction

Traditional Chinese medicine (TCM) formula is considered as a complex system: usually consists of thousands of components, either active or inactive, with unpredictable synergistic or

* Corresponding author. Tel.: +852 3943 6832; fax: +852 2603 5295.
E-mail address: joanzuo@cuhk.edu.hk (Z. Zuo).

http://dx.doi.org/10.1016/j.aca.2014.10.002
0003-2670/

antagonistic effects [1–4]. To better understand its therapeutic mechanism, a comprehensive identification of these components is essential. In the past 20 years, with the advance of liquid chromatography–mass spectrometry (LC–MS) [5], it is more and more popular to directly identify natural compounds from herbal extracts on line, thus avoiding the labor-intensive separation and purification steps in a traditional phytochemical study. On the other hand, handling and interpreting such huge amount of multi-dimensional and massive datasets, especially those from high-resolution mass spectrometers (such as TOF-MS and Orbitrap), is tedious and challenging [6]. Whilst the development of LC–MS based metabolomics/metabonomics greatly enhance the efficiency of several tedious data processing steps, such as determination of charge states, mono-isotopic masses, and features of the compound(s) of interest [7], the characterization and identification of these compounds (e.g., whether it is detected as pseudo-molecular ion, dimer, adduct-ion or any other cluster ions; whether the detected features are from one compound; whether there exists any relationship between all the detected compounds) is still challenging. Fortunately, it is not necessary to identify all the components in most omics-studies, and only the disease or exogenous intervention correlated metabolites (only less than a dozen among thousands of potential endogenous metabolites in biological fluids such as blood, plasma and urine) need to be further analyzed [8,9]. In contrast, for a typical chemical-base study of TCM preparations, it is expected to identify the (active) components as many as possible. However, most of the active components in TCM preparations are secondary metabolites, whose structural diversity and potential amount is far beyond those endogenous metabolites from biological fluids [10]. For samples analyzed by different LC–MS instruments and/or experimental conditions, the features arisen from the same component (e.g., its ionized form, its fragment ions and the relative abundance of these ions) would be different and so a comprehensive standard-metabolite MS library generated from LC–MS data would not be as successful as those in GC–MS based studies [11]. On top of the data processing issues, the co-elution of compounds from complex herbal samples still cannot be completely resolved. Thus, the task for enhancing the efficiency of component identification from LC–MS datasets is an urgent problem to be conquered.

Recently, we have developed a Matlab-based tool to classify TCM components according to their adduct-ion pattern obtained from full scan mass spectra [12]. The key idea lies in that compounds of different classes or functional groups have characteristic adduct-ion formation (e.g., ginsenosides prefer to form acetate adduct-ions and flavonoids prefer chloride adduct-ions when compared to their corresponding pseudo-molecular ions, PMI). Based on these characteristic ions and their relative abundance (named as adduct-ion patterns), the original TIC can be divided into several sub-datasets of which each sub-dataset contains only one to a few classes of herbal components and hence simplify the process of data interpretation. As a further development of the above tool, a novel stepwise component detection algorithm (SCDA) is developed aiming to extract all features of the compound of interest as a whole, to determine the identity (e.g., $[M-H]^-$, $[M+Cl]^-$, $[M+HCOO]^-$, $[2M-H]^-$, etc.) of each feature and to obtain the "purified" mass spectrum (i.e., noise-free and free from interference/co-eluent components' ions) of each component at the same time. We here present the principle and implementation of SCDA using a standard mixture of TCM reference compounds. In addition, the SCDA approach has been applied to three TCM preparations (an oral granule, an intravenous injection and a single herbal extract) to illustrate the specificity and merits of the developed algorithm.

2. Methodology

The ion distribution of a compound (component) obtained from LC–MS is dimensional dependent: on the retention time dimension (axis), it is continuous and approximately complies with a Gaussian distribution; whilst it is discrete on the mass-to-charge (m/z) dimension. The formation of these ions, (e.g., PMI, dimer and/or its adduct-ions), as presented in the full scan mass spectrum, depends on the physicochemical properties of the component, the composition of mobile phase, the ionization source and the type of the mass spectrometer used and is hard to be presented by a simple fixed function. Since they are originated from the same compound, they should have same chromatographic distribution along the time dimension, not considering the distortion by random noise. Thus, for each normal scan mass spectrum and at any particular time point (i.e., at one particular scan), we regard these discrete ions as correlated ions of one component. By evaluating these correlated ions at a specified time interval or scan range would screen out those 'unrelated' signals and hence result in a 'purified' reconstructed mass spectrum of that particular component, which is then utilized to determine the PMI automatically. Based on this concept, a stepwise component detection algorithm (SCDA) has been developed for processing LC–MS data (in mzXML or netCDF files) (Fig. 1): (1) determination of mono-isotopic mass and removal of isotopic peaks from each mass spectrum; (2) resolving components from the remaining data matrix; (3) classification of detected components. The first step identifies and removes isotopic ions so as to reduce data. The second step is the main body of SCDA, in which components of interest and their PMIs would be determined in a one-by-one highest abundant component elimination approach. The third step classifies the detected components by their adduct-ion and isotopic distribution characteristics so as to facilitate their identification. The three key steps of SCDA are explained in detail as follows:

2.1. Determination of mono-isotopic masses and removal of isotopic peaks

In general, isotopic clusters in the mass spectrum of a small molecule compounds consists of a mono-isotopic ion of the high abundance and its associated isotopic ions, from which the abundance of these isotopic peaks provides valuable information on the charge state of the molecule and its elemental composition. However, these isotopic ions would complicate the feature detection and result in a list of thousands of features, which have limited assistance for identification purposes. Thus, only the mono-isotopic ions, always the highest abundant in an isotopic cluster of small molecules, are utilized for resolving components in SCDA.

To identify the isotopic clusters, the following steps are performed: (i) search the most intense peak in each mass spectrum, (ii) locate the potential isotopic peaks having a m/z difference of $1.00235/c$ (c represents the charge number) within a pre-defined m/z tolerance (*tol*) from both forward and backward directions, (iii) if the length (the number of ions) of the detected isotopic cluster is longer than a pre-defined *min_cluster (i.e.,* 3), store it and remove the corresponding ion peaks from the current mass spectrum, and (iv) repeat steps (i) to (iii) until the most intense ion is less than the pre-set noise threshold level (*NL*), which is estimated by removing the top 5% of ions (those with the high abundance) from the original LC–MS dataset (PEAKS) and then averaging the intensities of the remaining ions.

Then, mono-isotopic ions in the above detected isotopic clusters are determined according to the following criteria: (a) if the intensity of the leftmost ion in a cluster is highest, then the leftmost ion is postulated as the mono-isotopic mass; (b) if not, for the ions from the leftmost to the most intense ion in a cluster, by