Journal of Chromatography A, 1429 (2016) 134–141

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Algorithm for comprehensive analysis of datasets from hyphenated high resolution mass spectrometric techniques using single ion profiles and cluster analysis

Guillaume L. Erny[a,*], Tanize Acunha[b], Carolina Simó[b], Alejandro Cifuentes[b], Arminda Alves[a]

[a] LEPABE – Laboratory for Process Engineering, Environment, Biotechnology and Energy, Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[b] Laboratory of Foodomics, CIAL, CSIC, Nicolas Cabrera 9, 28049 Madrid, Spain

ARTICLE INFO

Article history:
Received 9 October 2015
Received in revised form
27 November 2015
Accepted 1 December 2015
Available online 11 December 2015

Keywords:
Chemometrics
Hyphenated techniques
Data mining
Representation

ABSTRACT

Various algorithms have been developed to improve the quantity and quality of information that can be extracted from complex datasets obtained using hyphenated mass spectrometric techniques. While different approaches are possible, the key step often consists in arranging the data into a large series of profiles known as extracted ion profiles. Those profiles, similar to mono-dimensional separation profiles, are then processed to detect potential chromatographic peaks. This allows extracting from the dataset a large number of peaks that are characteristics of the compounds that have been separated. However, with mass spectrometry (MS) detection, the response is usually a complex signal whose pattern depends on the analyte, the MS instrument and the ionization method. When converted to ionic profiles, a single separated analyte will have multiple images at different m/z range. In this manuscript we present a hierarchical agglomerative clustering algorithm to group profiles with very similar feature. Each group aims to contain all profiles that are due to the transport and monitoring of a single analyte. Clustering results are then used to generate a 2 dimensional representation, called clusters plot, which allows an in-depth analysis of the MS dataset including the visualization of poorly separated compounds even when their intensity differs by more than two orders of magnitude. The usefulness of this new approach has been validated with data from capillary electrophoresis time of flight mass spectrometry hyphenated via an electrospray ionization. Using a mixture of 17 low molecular endogenous compounds it was verified that ionic profiles belonging to each compounds were correctly clustered even with very low degree of separation (R below 0.03). The approach was also validated using a urine sample. While with the total ion profile 15 peaks could be distinguished, 70 clusters were obtained allowing a much thorough analysis. In this particular example, the total computing took less than 10 min.

1. Introduction

Separation techniques coupled with high resolution mass spectrometry (X-HRMS, where X stands for GC, LC, CE...) are key analytical hyphenated instruments in many areas of research [1–8]. In these configurations, the mass spectrometer is placed at the outlet of the separation device, allowing recording full mass spectra at regular intervals. Initially aimed as a technique to identify the main peaks in a profile (following the recommendation of the IUPAC, profile can be an electropherogram, chromatogram or any continuous trace recorded by an analytical instrument [9]), X-HRMS is now essential for analysing complex matrices. MS instruments are considered almost universal detectors while extremely selective and sensitive. As a result, a typical non-targeted analysis by X-HRMS comprises a huge amount of information in a dataset practically impossible to handle in a manual way. A typical approximation during any X-HRMS analysis is to narrow the mass range of the MS detector to only detect those ions whose m/z are within the selected range. While this can be done via the instruments setting, the easiest way is often done by filtering the resulting dataset to obtain the extracted ion profile (EIP). Interestingly, filtering can be used multiple times using the same dataset, thereby, obtaining profiles that are selective to potentially represent every analyte present in the sample. However, knowledge of the target mass interval is a

* Corresponding author.
E-mail address: guillaume@fe.up.pt (G.L. Erny).

http://dx.doi.org/10.1016/j.chroma.2015.12.005
0021-9673/

key requirement for minor components as they will only be visible if the dataset is adequately filtered. This is the main limitation in non-targeted X-HRMS analysis. Because it is important in many fields to obtain a complete analysis of the sample, especially of unknown or unexpected components, many algorithms have been proposed to improve the quality and quantity of information that can be obtained after a non-targeted X-HRMS analysis [10].

Automatic processing of hyphenated dataset often starts with the transformation of the raw MS data into centroid data. This can be done by various algorithms that detect in each MS scan every peak and calculate for each of them the centroid mass (also called the accurate mass [9]). The centroid data only records the centroid m/z value and the peak amplitude. This allows a drastic reduction of the data size, and as a consequence, faster computational speed. Centroid data can be sometimes obtained by the proprietary acquisition or associated software, such as, compassXport (Bruker) or Masslynx (Waters), or secondary software [11–13]. Those software also allow to export the data to formats such as mzML [14], mzXML [15] or JCAMP [16] that can easily be read by users [16]. It should be emphasized that the transformation to centroid data is an important transformation of the original data that can influence a subsequent quantitative analysis [17]. The information in the resulting file can be seen as a collection of 3 coordinates data point: the accurate mass m/z value (referred from now on as mass coordinate), the scan number and the peak intensity. Such dataset can be processed in two ways, either each MS scan is analyzed to detect peaks that are probably related to the same compounds (isotopic ions, common adducts, possible fragmentations) [18–20] or profiles characteristic of the transport and separation of the formed ions can be reconstituted and processed as chromatographic profiles. The latter is often the method of choice in hyphenated MS datasets [21–23].

Two approaches are used to obtain chromatographic profiles, i.e., the binning approach and the pure ion profile. In the binning approach, a series of EIPs are defined in such a way that each EIP will only contain data within a very narrow mass range, so, all the information is distributed along the EIPs [11]. The mass range is the key parameter and should be carefully optimized to avoid peak splitting. Algorithms have also been proposed to correct for this default [24]. The EIPs are then processed to detect for the presence of chromatographic peaks. Those are then measured and figure of merits (peak, center, amplitude, m/z, area...) recorded in a table [22,24–26]. Another recent alternative is the use of pure ion profile (PIP), which consists of scanning the dataset to find sequences of points that follow each other in their scan coordinate and does not differ in their mass coordinate by more than a certain threshold that corresponds to the variation in the accurate mass value determined for the same peak in the mass profile (original data) at different scan [27–29]. The advantage of PIP over EIP (or binning approach) is that in the PIP only the data points related to the target ions are obtained. This makes the use of peak picking algorithms redundant. However, while EIP can be obtained from centroid or profile dataset, PIP can only be calculated using centroid dataset. Automatic processing of EIP or PIP allows obtaining a large number of chromatographic peaks that are characteristic of the analytes of interest present in the sample. Most algorithms are able to retrieve peaks at very low intensity that would have never been detected manually without information on their mass.

Nevertheless, this is not sufficient for a comprehensive analysis of the whole X-HRMS dataset due to its intrinsic complexity. When processing the dataset, multiple EIE or PIP will be found for every analyte that has been separated. Thus, the remaining problem is to determine whether a minor peak is due to a new analyte or an image of an existing peak. Usually, this problem is bypassed by working in differential analysis [12,30,31]. However, PIP related to the same compounds are highly related with each other. All PIP due to isotopic ions, adducts or fragments are an image of the main profile where the only variation is the intensity of the peak. We have already used this concept in a previous work to propose a new representation [21]. The present work is a significant improvement from that previous approach. Here we use for the first time a hierarchical agglomerative cluster approach (HACA) [32] to classify the different profiles, based on similarity as measured by the correlation coefficient into clusters. Ideally, selected clusters should contain all profiles related to the same analyte. While peak aggregation as already been discussed [33,34], HACA has never been tested in this context. While previously published approaches [33,34] used classical figures of merits such at time at peak apex and peak width, the HACA designed in this work is based on the valor of the correlation coefficients between profiles. It is therefore less dependent on noise that can prevent the accurate determination of peak apex and peak width. In the present work, we demonstrate, using test and real complex samples, that HACA allows to rationally and systematically organize the various profiles in a limited amount of clusters. In this work, the clusters are used to provide a new two dimensional representation of the dataset, called clusters plot, which allows a comprehensive analysis of the results, making possible to distinguish analytes that are not well separated.

2. Materials and methods

2.1. Programing

Programing was done using Matlab 2013b and functions were run on a personal computer (Intel Pentium CPU G3320@3.00 GHz, 4.00 GB RAM, 64-bit operating system) with Windows 7. All the functions programmed and used during this work can be obtain by contacting the corresponding author.

2.2. Chemicals and samples

All reagents employed in the preparation of the CE buffer and sheath liquid (isopropanol, formic acid and ammonium hydroxide, all of MS grade) were from Sigma–Aldrich (St. Louis, MO, USA). A mixture of 17 small molecular mass compounds from Sigma–Aldrich was employed as a test mixture: 0.3 mM adenosine triphosphate (ATP), 7.8 mM nicotinic acid, 0.5 mM glutamic acid, 1.0 mM aspartic acid, 0.3 mM glutathione oxidized, 0.6 mM glutathione reduced, 2.7 mM iminodiacetic acid, 0.2 mM adenosine monophosphate (AMP), 0.2 mM panthotenic acid, 1.4 mM succinic acid, 0.1 mM gluconic acid, 0.4 mM hippuric acid, 0.3 mM malic acid, 0.2 mM citric acid, 0.2 mM tartaric acid, and 2.8 mM 1,4 piperazinediethanesulfonic acid (PIPES) and 156.5 mM malonic acid. The urine was filtered through 0.2 μm polyethersulfone filter before CE-MS analysis.

2.3. CE-TOF MS analysis

The capillary electrophoresis (CE) apparatus used was a P/ACE 5010 from Beckman (Fullerton, CA, USA). The CE instrument was controlled by a PC running System GOLD software from Beckman. The CE equipment was coupled to a time-of-flight (TOF) instrument "microTOF" from Bruker Daltonik. CE-TOF coupling was carried out via an ESI interface model G1607A from Agilent Technologies. Electrical contact at the electrospray needle tip was established via a flow of sheath liquid composed of 2-propanol-water (50:50, v/v) delivered by a 74900-00-05 Cole Palmer syringe pump (Vernon Hills, IL, USA) at a flow rate of 4 μL/min. Bare fused-silica capillary with 50 μm i.d. and 85 cm of total length was from Composite Metal Services (Worcester, England). The inner capillary wall was coated with a cationic TEDETAMA-co-HPMA copolymer [35]. CE separation was performed at −20 kV in an acidic BGE (1 M formic acid