Analytica Chimica Acta 978 (2017) 10–23

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data

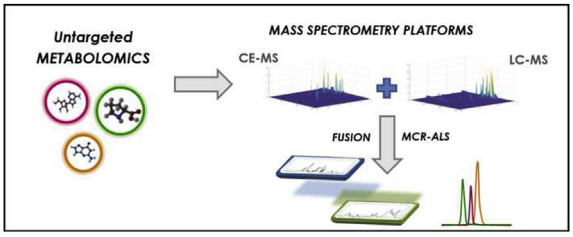
Elena Ortiz-Villanueva [a], Fernando Benavente [b], Benjamín Piña [a], Victoria Sanz-Nebot [b], Romà Tauler [a], Joaquim Jaumot [a, *]

[a] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain
[b] Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 645, 08028 Barcelona, Spain

HIGHLIGHTS

- Two data fusion strategies were proposed for untargeted metabolomics studies.
- Data fusion and results integration approaches were based on MCR-ALS.
- Goodness of proposed strategies was proven in a metabolomic study of yeast growth.
- Proposed chemometric approaches allowed the joint analysis of CE-MS and LC-MS data.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 September 2016
Received in revised form
7 March 2017
Accepted 25 April 2017
Available online 11 May 2017

Keywords:
Knowledge integration
Data fusion
Capillary electrophoresis-mass spectrometry
Liquid chromatography-mass spectrometry
MCR-ALS
Untargeted metabolomics

ABSTRACT

In this work, two knowledge integration strategies based on multivariate curve resolution alternating least squares (MCR-ALS) were used for the simultaneous analysis of data from two metabolomic platforms. The benefits and the suitability of these integration strategies were demonstrated in a comparative study of the metabolite profiles from yeast (*Saccharomyces cerevisiae*) samples grown in non-fermentable (acetate) and fermentable (glucose) carbon source. Untargeted metabolomics data acquired by capillary electrophoresis-mass spectrometry (CE-MS) and liquid chromatography-mass spectrometry (LC-MS) were jointly analysed. On the one hand, features obtained by independent MCR-ALS analysis of each dataset were joined to obtain a biological interpretation based on the combined metabolic network visualization. On the other hand, taking advantage of the common spectral mode, a low-level data fusion strategy was proposed merging CE-MS and LC-MS data before the MCR-ALS analysis to extract the most relevant features for further biological interpretation. Then, results obtained by the two presented methods were compared. Overall, the study highlights the ability of MCR-ALS to be used in any of both knowledge integration strategies for untargeted metabolomics. Furthermore, enhanced metabolite identification and differential carbon source response detection were achieved when considering a combination of LC-MS and CE-MS based platforms.

Abbreviations: BGE, Background electrolyte; BSA, Bovine serum albumin; CMTF, Coupled matrix and tensor factorization; COW, Correlation optimized warping; DISCO-SCA, Distinctive and common components with simultaneous component analysis; GSVD, Generalized singular value decomposition; JIVE, Joint and individual variation explained; MCR-ALS, Multivariate curve resolution alternating least squares; MWCO, Molecular weight cut-off; O2PLS, Two-way orthogonal projections to latent structures; OnPLS, Multiblock orthogonal projections to latent structures; PBS, Phosphate buffered saline; PIPES, 2,2'-(1,4-Piperazinediyl)diethanesulfonic acid; PLS-DA, Partial least squares discriminant analysis; YPD, Yeast extract peptone dextrose; YEPA, Yeast extract peptone acetate.

* Corresponding author.
E-mail address: joaquim.jaumot@idaea.csic.es (J. Jaumot).

http://dx.doi.org/10.1016/j.aca.2017.04.049
0003-2670/

1. Introduction

In the framework of *omic* sciences, there is a need towards the application of chemometric methods to analyse the large amount of data generated in chemical and biological studies. In the last few years, advanced data analysis tools and novel approaches have been developed to enhance the acquisition of new knowledge and improve the understanding of biological processes [1,2]. Data fusion and integration methods are nowadays the subject of active research in computational statistics and chemometrics [3–7]. These new methods allow the simultaneous study of datasets considering diverse viewpoints, such as datasets coming from different analytical platforms, *omic* levels, organisms or sample types [8,9]. For instance, in an inter-platform data fusion procedure for metabolomics, the same samples are jointly analysed using diverse methods or techniques [5,10]. In this way, the strengths of a particular analytical platform can be used to compensate the weaknesses of the rest, in an attempt to gather better metabolic information with greater accuracy and lower uncertainty. So, these strategies combining different sources of information are promising tools for fundamental *omic* studies as they allow an improved biomarker detection, hence a better characterization of biological responses.

From a chemometric point of view, extensive work has been done in the field of data fusion (that can also be known as multiset analysis). More recently, these data fusion strategies have been applied in diverse research fields and a classification of these possible different approaches has been proposed distinguishing high-level, mid-level and low-level data fusion [5,11,12]. High-level fusion implies optimal preprocessing and modelling procedures for each data block separately. Different models outputs are then jointly evaluated to provide a global overview, which is often hardly interpretable. Similarly to this high-level fusion, integration of the results obtained in these individual analyses of the blocks can yield an enhanced biological interpretation. For instance, identified features can be visualized at the same time in a metabolic or gene network to gain a deeper knowledge of the underlying biological processes, *i.e.* pathway based integration [10,13]. In contrast, low-level and mid-level fusion strategies aim to combine first data blocks to obtain later a model with an improved joint interpretation. Low-level fusion generates vast size fused data with a large amount of variables; whereas mid-level fusion is based on a previous dimensionality compression of data blocks where only a reduced number of pre-selected variables (usually the most relevant or combinations of them) from each data block are fused and jointly interpreted. Regarding the chemometric methods used for these data fusion studies, several proposals have been done considering these data fusion levels. In the case of high-level data fusion, each dataset is analysed individually using traditional chemometric methods for sample classification and feature detection [9,14–17]. More interestingly, additional efforts have been done for the development of mid-level data fusion methods. The methods found in the literature try to identify the common and specific variance coming from each one of the analysed blocks after a feature selection to reduce the size of the dataset. Various methods should be highlighted such as GSVD (generalized singular value decomposition) [18], O2PLS (two-way orthogonal projections to latent structures) [19], OnPLS (multiblock orthogonal projections to latent structures) [20], DISCO-SCA (distinctive and common components with simultaneous component analysis) [9], JIVE (joint and individual variation explained) [21], and CMTF (coupled matrix and tensor factorization) [14]. Some of these mid-level data fusion methods can also be used for low-level data fusion depending on the raw data characteristics (usually an appropriate block scaling is required). However, other methods could also be an option for this purpose, such as multivariate curve resolution alternating least squares (MCR-ALS). MCR-ALS allows the joint analysis of multiple datasets from different samples (experiments) or techniques or from different samples and techniques simultaneously. The benefits of MCR-ALS in diverse research fields have been extensively discussed in previous literature not related to the *omic* sciences for the join analysis of different data sources (*i.e.* spectroscopies, physical or chemical parameters) or different samples (*i.e.* biological processes, chromatographic runs) [22–26]. In addition, to the best of our knowledge, this is the first study investigating this low-level data fusion strategy in CE-MS and LC-MS inter-platforms for untargeted metabolomics.

There is a broad variety of instrumental techniques that can be used for *omic* studies. Nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are the most widely used techniques in metabolomics [27,28]. NMR is accepted as the most advisable platform to obtain reproducible results with negligible coefficients of variation, but its sensitivity is low. In contrast, hyphenated MS-based platforms are highly recommended for the great selectivity and sensitivity. Gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS) and capillary electrophoresis-mass spectrometry (CE-MS) provide easy and reliable metabolite separation, detection, identification and quantification. These platforms often produce data multisets containing partly complementary information (in terms of selectivity of the separations, polarity of the compounds detected and concentration ranges) that jointly analysed may reveal underlying pathways in highly complex samples, difficult to extract otherwise. Particularly, CE-MS presents some properties that complement the more commonly used techniques (reversed-phase LC-MS and GC-MS). Thus, CE-MS provides information about charged and highly polar compounds in a fast and simple way (without the need of chemical derivatization) and using an extremely small volume of sample. On particular occasions, these multiple analytical platforms can be useful to cover all the changes induced in an organism by an external stimulus. However, the datasets from these diverse analytical sources can be heterogeneous, and the analysis is still challenging. Several articles have proved the importance of data fusion models in untargeted *omic* studies, using an inter-platform fusion of ^{1}H NMR and LC-MS [29], ^{1}H NMR and GC-MS [30] or LC-MS and GC-MS [31,32]. However, to the best of our knowledge, only a few of these studies have been previously reported using CE (*e.g.* CE and NMR) [33].

The main aim of this work is the proposal of two different inter-platform knowledge integration strategies based on the application of MCR-ALS for the joint analysis of untargeted metabolomics data. The suitability of these two data analysis strategies presented in this work is demonstrated in a comparative study of the metabolic changes induced in yeast (*Saccharomyces cerevisiae*) growth by a non-fermentable carbon source (acetate) with regard to the typical fermentable carbon source (glucose). The selection of this biological system is due to the advantages of yeast as an optimal model organism. First, yeast metabolism is well known and it is essentially as complex as any other eukaryotic organisms. This knowledge facilitates the identification of the detected metabolites, the interpretation of the observed metabolic changes and the extrapolation of the biological interpretation to any other eukaryote. The changes in the yeast metabolome derived from growing on fermentable or non-fermentable carbon sources are of major interest for biotechnology and food industries. These industries rely on the unique yeast metabolic properties for a vast number of applications, from bakery, brewery, and wine-making to the production of different recombinant proteins or its use as single-cell catalysts in fine chemistry [34–36].