Chemometrics and Intelligent Laboratory Systems 108 (2011) 33–48

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Exploring liquid chromatography–mass spectrometry fingerprints of urine samples from patients with prostate or urinary bladder cancer

Rolf Danielsson, Erik Allard, Per Johan Ragnar Sjöberg, Jonas Bergquist *

Department of Analytical Chemistry, Biomedical Centre, Uppsala University, P.O. Box 599, SE-75124 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 30 September 2010
Received in revised form 16 March 2011
Accepted 17 March 2011
Available online 23 March 2011

Keywords:
Urine profile
LC MS
Metabolic fingerprinting

ABSTRACT

Data processing and analysis have become true rate and success limiting factors for molecular research where a large number of samples of high complexity are included in the data set. In general rather complicated methodologies are needed for the combination and comparison of information as obtained from selected analytical platforms. Although commercial as well as freely accessible software for high-throughput data processing are available for most platforms, tailored in-house solutions for data management and analysis can provide the versatility and transparency eligible for e.g. method development and pilot studies.

This paper describes a procedure for exploring metabolic fingerprints in urine samples from prostate and bladder cancer patients with a set of in-house developed Matlab tools. In spite of the immense amount of data produced by the LC–MS platform, in this study more than 10^{10} data points, it is shown that the data processing tasks can be handled with reasonable computer resources. The preprocessing steps include baseline subtraction and noise reduction, followed by an initial time alignment. In the data analysis the fingerprints are treated as 2-D images, i.e. pixel by pixel, in contrast to the more common list-based approach after peak or feature detection. Although the latter approach greatly reduces the data complexity, it also involves a critical step that may obscure essential information due to undetected or misaligned peaks. The effects of remaining time shifts after the initial alignment are reduced by a binning and 'blurring' procedure prior to the comparative multivariate and univariate data analyses. Other factors than cancer assignment were taken into account by ANOVA applied to the PCA scores as well as to the individual variables (pixels). It was found that the analytical day-to-day variations in our study had a large confounding effect on the cancer related differences, which emphasizes the role of proper normalization and/or experimental design. While PCA could not establish significant cancer related patterns, the pixel-wise univariate analysis could provide a list of about a hundred 'hotspots' indicating possible biomarkers. This was also the limited goal for this study, with focus on the exploration of a really huge and complex data set. True biomarker identification, however, needs thorough validation and verification in separate patient sets.

1. Introduction

1.1. General

Metabolic fingerprinting is a strategy for investigating systematic changes in the metabolome of a living organism due to diseases or external influences, such as exposure to pharmacologically active substances. Neither quantitation nor a priori knowledge of the measured compounds is needed. Metabolic fingerprinting is therefore a promising strategy for finding new potential biomarkers of various diseases [1].

Mass spectrometry has emerged as one of the major analytical platforms in metabolomics, either with direct infusion or coupled to a separation technique [2]. The development and use of hyphenated techniques for metabolite analysis have been reviewed for LC–MS [3,4], GC–MS [5] and CE–MS [6–8]. As a non-targeted approach, metabolic fingerprinting implies that as much as possible of the metabolome should be covered by the analytical procedure. In separation-based MS the fingerprint for each sample is obtained as a series of consecutive mass spectra, and the number of data points may reach the order of 10^8 or even higher. It is not unusual that the task to handle and explore the vast amount of data is the bottleneck in a metabolic study. To enable high-throughput properties of metabolic/proteomic assays efficient software have been developed, both commercially and freely accessible (a list of free software for processing MS data is available at http://www.ms-utils.org). Several reviews of computational methods and available software for processing separation-based MS data in metabolomics/proteomics have been published [9–13].

Our group has previously developed Matlab tools for exploration of two-way fingerprints of complex samples obtained from CE/LC/GC separation with MS detection [14–17]. These tools are mainly

* Corresponding author.
E-mail address: jonas.bergquist@kemi.uu.se (J. Bergquist).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.03.008

intended for use in method development, pilot studies or initial data exploration. Much effort has been spent to enable visually guided and interactively controlled procedures in the data evaluation, with less focus on optimized high-throughput data processing.

1.2. Related work

The data processing workflow for non-targeted metabolic/proteomic studies involves data preprocessing followed by data analysis. Usually the preprocessing phase results in lists of peaks or features corresponding to separate, although unknown, chemical substances in the samples. The lists may then be subject to data analysis by standard programs for univariate or multivariate statistics. Preprocessing packages perform a series of preprocessing steps including raw data extraction, m/z binning into m/z channels, noise removal and baseline subtraction, peak or feature detection, and alignment of peaks or features between samples.

1.2.1. m/z binning

The data for each run (sample) is usually arranged in a 2-D data array, with time as one dimension and m/z as the other. While there is a natural grid for the time dimension (i.e. the scan number), the distribution of the m/z values depends on the MS technique utilized. With signal acquisition at a constant speed, the difference between two adjacent m/z values is proportional to $(m/z)^{1/2}$ for a time-of-flight instrument and to $(m/z)^2$ for a Fourier transform based instrument (FT-ICR or Orbitrap). The m/z values actually reported by the MS instrumentation software will also depend on the built-in calibration procedure; in some cases frequent recalibration even during a run distorts the intrinsic discrete m/z distribution. The same holds true if data points below an intensity threshold are discarded or if only centroided peak values are reported. With most preprocessing packages, as with XCMS [18] and MZmine [19], m/z binning is made in equally sized steps. The subsequent processing (noise reduction, background subtraction, and peak detection) is then performed on the separate chromatograms for each m/z channel (rows or columns in the data matrix). However, the width of the mass spectral peaks, as well as the intrinsic m/z distribution, will vary over the m/z range. With a constant m/z bin size it is therefore difficult to avoid overlapping chromatographic peaks (too large bins) or splitting chromatographic peaks over several m/z channels, even with empty gaps between them (too small bins). The bin size will also affect the noise and background conditions; with large bins the m/z channels include more noise and with small bins the elevated background signal may shift between m/z channels during the run. These circumstances have led to development of peak/feature detection methods without regular m/z binning, mostly utilizing centroided m/z values [20–24].

1.2.2. Noise and baseline, peak list vs. 2-D image

Noise filtering and baseline removal may take place prior to peak/feature detection or be included in the latter procedure. In both cases the outcome of the peak/feature detection depends on the parameters selected for the processing steps. It has been demonstrated that even with optimized parameters the number of detected features for the same dataset depends on the software utilized [22,24]. To circumvent the inherent problems with peak/feature detection, the 2-D data matrix could be retained as an image characterizing the sample without reduction to peaks or features. Further data analysis, involving comparison between samples, is then performed on a datapoint-by-datapoint basis, or 'pixel-wise' with the pixels defined by the binning in time and m/z dimensions. This image-based approach has been taken by several groups [25–30]. It was also the basis for our Matlab tools, using a constant m/z bin size that is comparable with the average peak width over the m/z range (in the present study $\Delta m/z$ 0.1). For each m/z channel the baseline is found by iterative asymmetric least-squares estimation [31,32] of a spline function and then subtracted. All data points below a noise level derived from the variations around the fitted baseline are then discarded as noise. Noise reduction and baseline removal applied to the individual runs will considerably reduce the amount of data and also facilitate further data analysis.

1.2.3. Alignment

Image-based as well as peak or feature-based comparison between data from different runs or samples requires alignment with respect to retention times as well as m/z, and the large amount of work in this area has been extensively reviewed [33–36]. Especially the inevitable shifts in retention times between runs have been considered, usually with a non-linear time warping function for each run as a remedy. However, such a global alignment (i.e. common for all m/z channels) cannot handle local shifts; even reversed order of peaks have been reported [35,37] and also found in our study.

With our approach the comparative analysis is performed in a way that is fairly tolerant to shifts in time and m/z (to be described below in *Statistical analysis*). The influence of time shifts could be reduced by time binning, although bins that are large enough to account for possible time shifts would probably result in merging peaks of different origin. Therefore an initial alignment is performed of all runs vs. a master run, with visually guided selection of time points with high spectral correlation as knot points of a piecewise linear warping function. Prior to data analysis the data may be subject to further time binning. Choosing the bin sizes to be comparable with the peak widths in time and m/z, respectively, will reduce the amount of data with little loss of information. However, the risk of peak splitting between adjacent bins must be considered during the following analysis. The remaining time shifts could also correspond to several time bins.

1.2.4. Normalization, scaling, transformation

Three related issues when comparing pairs or groups of individual 2-D fingerprints are normalization, scaling and transformation of the data. These are all dealing with the quantitative measure obtained for each pixel, i.e. the summed intensities for all data points within each combination of time and m/z bins. A common strategy for normalization is to apply a single scale factor for all intensities in a run, while more elaborate normalization schemes also have been described [38]. Especially in urine samples the metabolite concentration may show strong fluctuations, and possible normalization strategies based on specific urine sample properties have been discussed [39]. A simple but more general approach is to normalize the intensities within a run to unit sum, which could at least partly compensate for variations in experimental factors in sample preparation and sample analysis. This is also the normalization method utilized in our tools.

The impact of scaling and transformation of metabolomics data was discussed by van den Berg et al. [40]. Common scaling methods may be appropriate for integrated peak or feature representation, while with image-based data also pixels with very low intensities are retained. For such positions autoscaling and similar methods could easily inflate even small random variations. The intention to reduce the predominant influence of high-intensity pixel positions can be fulfilled by operations like log transformation ($\log x$) or power transformation ($x^{1/k}$). In this work we do without scaling but apply square-rooting ($k=2$) as a compromise between original data and log transformed data. Square-rooting implies a level-independent change in relative differences while absolute differences change as the square-root of the level. Thus, with square-rooting the relative differences are enhanced in comparison with absolute differences, but not as much as with log transformation. Another feature with log transformation is that multiplicative effects turn into additive effects, which sometimes is favorable in ANOVA modeling of influential factors. In reality the factors may exert additive as well as