Analytica Chimica Acta 686 (2011) 57–63

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

# Principal component directed partial least squares analysis for combining nuclear magnetic resonance and mass spectrometry data in metabolomics: Application to the detection of breast cancer

Haiwei Gu[a], Zhengzheng Pan[b], Bowei Xi[c], Vincent Asiago[b], Brian Musselman[d], Daniel Raftery[b,*]

[a] Department of Physics, Purdue University, West Lafayette, IN 47907, United States
[b] Department of Chemistry, Purdue University, West Lafayette, IN 47907, United States
[c] Department of Statistics, Purdue University, West Lafayette, IN 47907, United States
[d] IonSense Inc., 999 Broadway, Suite 404, Saugus, MA 01906, United States

ARTICLE INFO

*Article history:*
Received 20 May 2010
Received in revised form
17 November 2010
Accepted 18 November 2010
Available online 26 November 2010

*Keywords:*
Metabolomics
Breast cancer
Nuclear magnetic resonance
Direct analysis in real time
Mass spectrometry
Partial least squares
Orthogonal signal correction
Human serum

ABSTRACT

Nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are the two most commonly used analytical tools in metabolomics, and their complementary nature makes the combination particularly attractive. A combined analytical approach can improve the potential for providing reliable methods to detect metabolic profile alterations in biofluids or tissues caused by disease, toxicity, etc. In this paper, $^1$H NMR spectroscopy and direct analysis in real time (DART)-MS were used for the metabolomics analysis of serum samples from breast cancer patients and healthy controls. Principal component analysis (PCA) of the NMR data showed that the first principal component (PC1) scores could be used to separate cancer from normal samples. However, no such obvious clustering could be observed in the PCA score plot of DART-MS data, even though DART-MS can provide a rich and informative metabolic profile. Using a modified multivariate statistical approach, the DART-MS data were then reevaluated by orthogonal signal correction (OSC) pretreated partial least squares (PLS), in which the Y matrix in the regression was set to the PC1 score values from the NMR data analysis. This approach, and a similar one using the first latent variable from PLS-DA of the NMR data resulted in a significant improvement of the separation between the disease samples and normals, and a metabolic profile related to breast cancer could be extracted from DART-MS. The new approach allows the disease classification to be expressed on a continuum as opposed to a binary scale and thus better represents the disease and healthy classifications. An improved metabolic profile obtained by combining MS and NMR by this approach may be useful to achieve more accurate disease detection and gain more insight regarding disease mechanisms and biology.



## 1. Introduction

Metabolomics, along with the related areas of metabonomics and metabolite profiling, is a powerful systems biology approach which combines data-rich analytical techniques with chemometrics for advanced investigations of metabolism in biological systems [1–5]. Among the many promising applications in the field of metabolomics, early detection of disease through the discovery of new biomarkers is an attractive driving force for research [6]. Metabolic profiling, in which quantitative information on a limited set of metabolites is measured, and fingerprinting, where the focus is on a broader pattern of metabolite signals, are two frequently used approaches in metabolomics studies [7]. These as well as other targeted or global approaches are being examined intensely to evaluate their success in detecting metabolic perturbations for a variety of fundamental studies and important applications.

Nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are the two most commonly used analytical tools in metabolomics [6–11]. $^1$H NMR spectroscopy is useful in metabolomics studies primarily because it is quantitative and highly reproducible, while MS provides much better sensitivity and is more selective than NMR. An increasing number of studies are taking advantage of the complementary nature of both methods [12–17]. While the NMR instrumentation used in the field of metabolomics is relatively standard, a variety of MS instruments and techniques are currently being applied in metabolomics. In addition to the widely used methods of gas chromatography (GC)–MS [18] and liquid chromatography (LC)–MS [19], atmospheric sample introduction methods are being applied in

* Corresponding author. Tel.: +1 765 494 6070; fax: +1 765 494 0239.
*E-mail address:* raftery@purdue.edu (D. Raftery).

0003-2670/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2010.11.040

metabolomics, including desorption electrospray ionization (DESI) [16,20] and extractive electrospray ionization (EESI) [15,21].

DART (direct analysis in real time) is a newly developed atmospheric ionization method that has extensive potential in applications such as analyzing chemical reagents, drugs, metabolites, and peptides [22–24]. The DART method requires no sample separation prior to analysis and sampling is completed by simply dipping the closed end of a glass melting point capillary tube into the serum. Another advantage of applying DART-MS in metabolomics is that despite the presence of sodium and potassium salts in serum the ionization method protonated metabolites without production of either sodium- or potassium-adducts of those same metabolites resulting in a simplified mass spectrum with fewer ions to quantitate. These features make it reasonable to anticipate that high throughput DART-MS analysis with inexpensive consumable samplers could be accomplished for numerous biological applications [25].

In terms of data analysis, several recent metabolomics studies have been reported that combine both NMR and MS techniques using advanced statistical methods. Statistical heterospectroscopy (SHY) and orthogonal partial least squares (O-PLS) algorithms have been used to integrate profiles from different analytical platforms [14,26]. Pan et al. applied Pearson correlation between NMR and DESI-MS data sets to obtain a list of molecules associated with different inborn errors of metabolism (IEMs) [16]. Chen et al. [12] improved the classification between healthy mice and mice with lung cancer using a combined 3D score plot, with two principal component (PC) scores obtained from the DESI-MS data and one PC score obtained from the NMR data. Since NMR and MS generate unique metabolic profiles, the combination of these two analytical tools using various statistical methods can provide new metabolic insights as well as avenues for inquiry and development in metabolomics.

A variety of multivariate statistical methods are currently in use in the metabolomics field. Principal component analysis (PCA) is a dimension reduction method based on identifying variance and is probably the most widely used multivariate approach [27,28]. Consensus PCA (CPCA) performs PCA analysis on multiple blocks of data measured on the same objects [29,30]. The bilinear statistical approach of partial least squares discriminant analysis (PLS-DA) is one of the most popular supervised methods used in metabolomics. In PLS-DA, the X matrix contains the data variables, while the Y matrix contains the class variable for which values are chosen to be the class descriptor [31–33]. Orthogonal signal correction (OSC) is a PLS-based data filtering technique that removes the information in X matrix which is uncorrelated to the Y matrix, and consequently a PLS model based on the now corrected X matrix may focus the analysis more exclusively on the variable(s) of interest [34–36]. Orthogonal projection to latent structures [37] is an alternative model. OSC-PLS and O-PLS have the same objective but achieve the goal through different means. OSC-PLS uses an internal iterative method to find orthogonal components and O-PLS is a modification of non-linear iterative partial least squares (NIPALS) [38]. Cross-model validation is recommended to accurately estimate the classification error rates of PLS models [30,39,40]. An extra layer of validation is provided by cross-model validation. Hence the result is a conservative estimate of the robustness of the model and its expected performance from a new dataset.

In the present study, we propose an alternative to PLS-DA in which we combine NMR and DART-MS data to discover potential serum biomarkers for breast cancer. Instead of using a dummy Y matrix, we select a more meaningful Y vector in the PLS regression, using the first principal component from the PCA of the NMR data. This proposed approach provides a continuous variable for the Y matrix, instead of the binary dummy variable. To avoid uninteresting noise in generating the metabolic profile, an OSC-PLS model was generated based on the DART-MS data regression against PC1 scores from the NMR data, which is believed to carry the most variation related to breast cancer (*vida infra*). Samples in each class (control or breast cancer) no longer shared the same Y values. Instead, the Y vector reflects both the variation between the two classes and that within each class. The combination of these two analytical techniques will likely have powerful capabilities in the areas such as disease detection and biomarker discovery.

## 2. Methods

### 2.1. Sample collection

Commercial human serum samples from 30 healthy controls and 27 breast cancer patients were purchased from Asterand (Asterand, plc. Detroit, MI). All the serum samples were obtained from female volunteers with ages ranging from 40 to 75 years old, and were approximately age matched. A table summarizing the clinical characteristics of the cancer patient is shown in Supplemental Information Table S1. Samples from cancer patients were obtained prior to therapy. Samples were de-identified at Asterand. Samples were transported over dry ice to Purdue University and stored at −80 °C until measurements were conducted.

### 2.2. $^1$H NMR spectroscopy

Samples were prepared by mixing 400 μL serum with 300 μL of a 1.5 mM 3-(trimethylsilyl) propionic-(2,2,3,3-$d_4$) acid sodium salt (TSP) solution (in $D_2O$), in which TSP was used as the frequency standard ($\delta = 0.00$ ppm). Sample solutions were vortexed for 60 s and centrifuged for 10 min at 7000 rpm. Aliquots of 580 μL were transferred into standard 5 mm NMR tubes for NMR measurements. A Bruker DRX 500 MHz spectrometer equipped with a room temperature HCN probe was used to acquire 1D $^1$H spectra. Samples were measured using a standard 1D CPMG (Carr-Purcell-Meiboom-Gill) pulse sequence coupled with water presaturation. For each spectrum, 32 transients were collected resulting in 32k data points using a spectral width of 6000 Hz. An exponential weighting function corresponding to 0.3 Hz line broadening was applied to the free induction decay (FID) before applying Fourier transformation. After phasing and baseline correction using Bruker's XWINNMR software, the processed data were saved in ASCII format for further multivariate statistical analysis.

### 2.3. DART-MS spectroscopy

DART-MS experiments were carried out using a Finnegan LCQ Classic quadrupole ion trap coupled with a DART ion source (IonSense, Boston, MA). For the DART ion source, helium gas was introduced into the corona discharge chamber at 2.0 L min$^{-1}$. The needle electrode was held at −3000 V. The first DC-biased electrode was held at 300 V and the exit electrode at 150 V. The DART ion source was located 20 mm away from the mass spectrometer inlet, which was held at a potential of 54 V. Samples were positioned and held on a mechanized sliding arm, which assured reproducible sample position within the ionization stream. 100-fold diluted serum samples were examined without any further sample pretreatment and each sample was deposited directly to the bottom of a 1.5 mm OD × 90 mm long capillary tube. The nitrogen gas in the DART ion source was heated to 350 °C. Data were acquired for 1 min to establish the background signal. The capillary, with the sample on its surface, was then quickly moved into and through the desorption ionization region immediately in front of the exit of the DART source and between that exit and the atmospheric pressure inlet of the mass