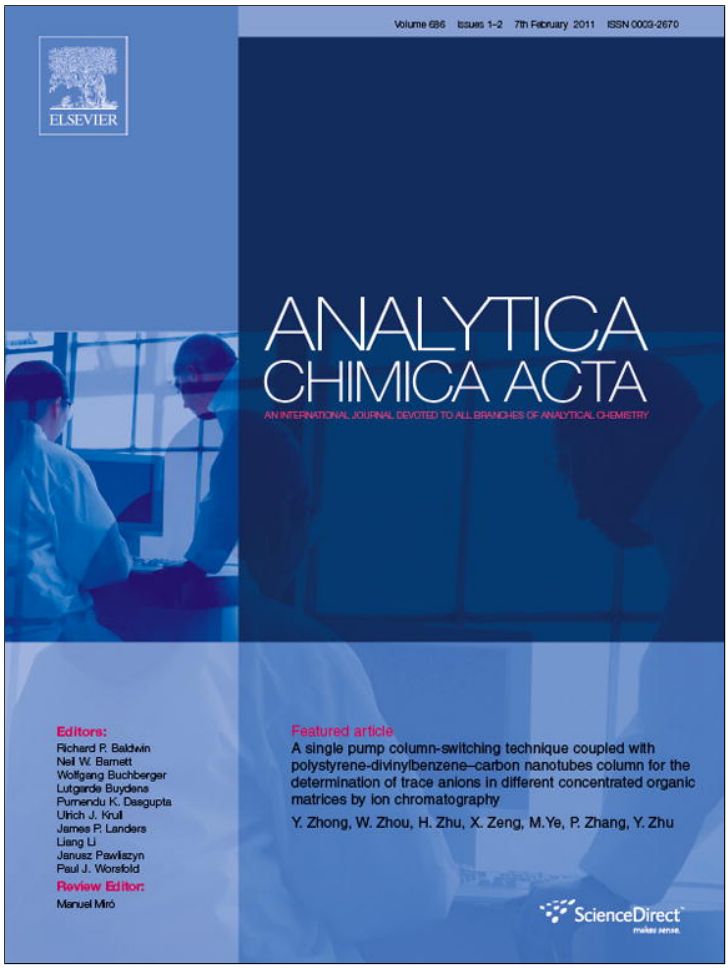

Analytica Chimica Acta 686 (2011) 57–63

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 (δ = 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

spectrometer. Spectra were acquired over a mass range of m/z 100–1000.

2.4. Data analysis

Each NMR spectrum was reduced to 800 frequency bins (0.02 ppm bin size) using the R statistical package (version 2.2.1). Spectral regions within the range of 0.94 ppm to 10 ppm were analyzed after deleting the region between 4.5 and 6.0 ppm that contained the residual water peak and urea signal. DART-MS spectra were analyzed with full (unit m/z) resolution. The data were then imported into Matlab software (Mathworks, MA) installed with the PLS toolbox (Eigenvector Research, Inc., version 4.0) for PCA and PLS modeling and open source software R (version 2.2.1) for k-nearest neighbor and binning.

Both NMR and DART-MS spectra were normalized (such that the total intensity of each spectrum is equal to 1) and mean-centered prior to PCA. PCA scores of NMR spectra alone did not show separate clusters for control and cancer samples. Therefore, PLS-DA and OSC-PLS-DA models were applied to the DART-MS spectra (normalized and mean-centered), using a dummy Y matrix. We performed cross-model validation to examine the classification error rates of the models. An alternative model was then used to combine the two datasets: the single-component Y matrix was chosen to be the PC1 score from the NMR data, because it achieved a better separation of the cancer and control samples than the PC1 score from the DART-MS data. The X matrix was composed of the DART-MS spectral data. Both X and Y matrices were normalized and mean-centered prior to multivariate analysis using Matlab to ensure that the PC1 values could be compared. PLS and OSC-PLS models were fitted using the above Y and X matrices. Results from cross-model validation were compared to the results from models using one dataset only. A receiver operating characteristics (ROC) graph and area under the ROC curve (AUC) were calculated for each model to compare their performance.

3. Results and discussion

3.1. NMR and DART-MS spectra

^{1}H NMR spectra of the serum samples from a healthy control and a breast cancer patient are presented in Supplemental Fig. S1a and b, respectively. It can be clearly seen that the aliphatic region dominates both control and disease spectra. Compounds identified in Fig. 1 include lactate (4.14 ppm, q, 1.35 ppm, d), creatine (3.94 ppm, s, 3.04 ppm, s), methionine (3.86 ppm, t), glucose (3.75 ppm, m), glycine (3.57 ppm, s), *myo*-inositol/glycerol (3.56 ppm), taurine (3.43 ppm, t, 3.25 ppm, t), trimethylamine oxide (TMAO)/betaine (3.27 ppm, s), acetone (2.25 ppm, s), alanine (1.49 ppm, d), and hydroxybutyrate (1.17 ppm, d). Several metabolites, such as taurine, acetone, alanine, and hydroxybutyrate, are observed to have large variations between the two spectra in Supplemental Fig. S1. Assignment of these metabolites was based on the use of KnowItAll software (Bio-Rad Laboratories, Inc., Hercules, CA) and previous studies [41–44].

DART-MS spectra from the same two samples, healthy control and breast cancer patient, are shown in Fig. 1a and b, respectively. Many peaks are spread over a wide mass range (m/z 100–1000), which demonstrates the ability of DART to ionize small metabolites (typically <600 Da) as well as some larger molecules. Though there are certain differences in a number of peak intensities, the DART-MS spectra are quite similar, and in general there are no prominent peaks which can be used to differentiate control and breast cancer samples by visual comparison.

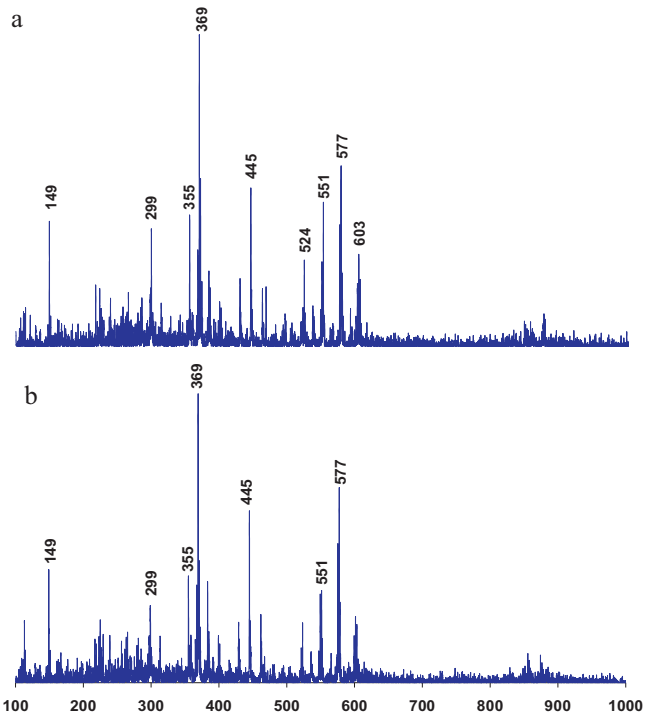

Fig. 1. DART-MS spectra of the same serum samples from: (a) healthy control and (b) breast cancer patient.

Furthermore one metabolite may produce several peaks and introduce some correlation among the peaks. Multivariate data analysis is thus necessary to extract subtle changes in the DART-MS data set.

3.2. PCA results for individual NMR and DART-MS data

We selected 5 PCs for the NMR spectra which explained 86.1% of the total variance. We then used a k-nearest neighbor classifier to examine how many samples were misclassified using the five principal components. The k-nearest neighbor classifier classifies an object based on the majority vote of its k-nearest neighbors. The value of k is chosen to minimize the number of misclassified objects. With $k = 3$ and using leave-one-out cross validation, there were 6 misclassified cancer samples and 3 misclassified control samples (see Table 1). Fig. 2a presents the PC1 and PC2 scores of the NMR spectra from the 57 human serum samples. In the score plot and later in other figures, two ellipses indicate 95% confidence regions for the normal and breast cancer samples individually. The 95% confidence regions are centered at the mean values of each group. Their size is determined by the variance of each group and the 95th per-

Table 1
Classification results.

Model	Misclassified cancer samples total = 27	Misclassified control samples total = 30
PCA (NMR)	6	3
PCA (DART)	7	6
PLS-DA (DART)	11	11
OSC-PLS-DA (DART)	10	6
PLS-DA (NMR)	5	5
OSC-PLS-DA (NMR)	3	5
PC direct PLS-DA	5	6
PC directed OSC-PLS-DA	3	1
LV directed OSC-PLS-DA	1	2

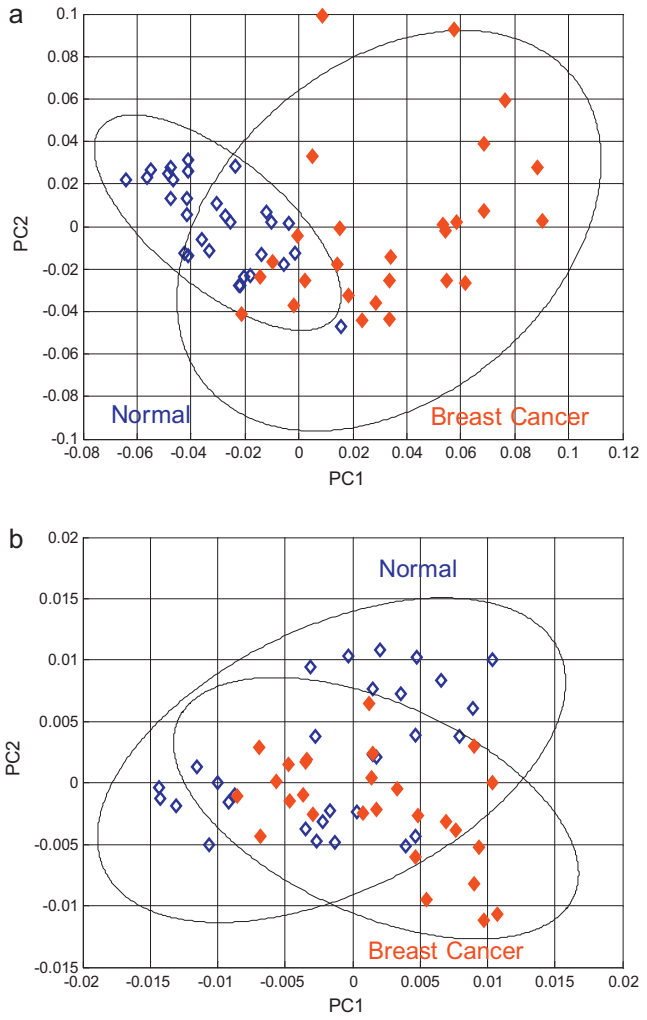

Fig. 2. Score plots from the results of PCA of (a) NMR and (b) DART-MS spectra of the same 57 human serum samples. Open diamonds represent normal samples and red solid diamonds represent breast cancer samples. Ellipses in the score plot of this figure and all other figures illustrate the 95% confidence regions of the corresponding groups.

centile of a chi-square distribution. One outlier (normal sample) is observed along PC1 and another (cancer sample) along PC2 at the 95% confidence level.

The two groups are separated mainly along the PC1 direction, which carries 39.4% of the total variance in the NMR data. However, there is no distinct boundary between the two groups. In the corresponding loading plot (Supplemental Fig. S2), several metabolites, such as taurine, lactate, glucose and several others can be identified. These molecules are among those that have been reported previously to be correlated with breast cancer development [45–48].

We selected 6 PCs for the DART-MS spectra which explained 75.1% of the total variance. The *k*-nearest neighbor classifier with $k = 5$ and using leave-one-out cross validation returned 7 misclassified cancer samples and 6 misclassified control samples (see Table 1). The PC1 and PC2 score plot is shown in Fig. 2b. No well-separated clusters could be identified from the PC1 and PC2 scores which captured 62.5% of the total variance in the DART-MS data. The poor classification observed is in agreement with the observation that there is no obvious differentiation when comparing individual normal and disease DART-MS spectra.

3.3. PLS and OSC-PLS analysis of DART-MS spectra

For PLS-DA and OSC-PLS-DA models fitted to DART-MS spectra alone, we used leave-one-out cross validation to select the number of latent variables (LVs) based on the cross validation prediction error sum of squares (PRESS) curve. (Q^2, an important predictive ability parameter, equals 1 minus the ratio of PRESS and the total sum of squares, where the total sum of squares is a constant.)

PLS-DA was then applied to the DART-MS spectra, and the score plot is shown in Supplemental Fig. S3. To build the PLS-DA model, the presence of breast cancer was used as the input for the Y-matrix data ("1" for the presence of cancer and "0" for its absence). 5 LVs were selected. The results of this analysis are similar to the PCA results of Fig. 2b, in that there is no obvious improvement in the separation between the two groups.

To assess the classification error of the PLS-DA model we performed a 4-fold cross-model validation. In each trial 75% of the samples (75% from each group) were used as calibration data set and the remaining 25% as an independent test data set. The calibration data set and the test data set were each normalized and mean-centered separately. Leave-one-out cross-validation was performed with the calibration data set to determine the number of LVs for each fold and construct a model. The resulting model was applied to the test data set to compute the predicted Y values. The above process was repeated 4 times so that every sample served as an independent test sample once and only once. We then used the predicted Y values produced by the 4-fold cross-model validation to examine the classification error. The predicted Y values obtained from the cross-model validation returned 11 misclassified cancer samples and 11 misclassified control samples.

Next, OSC-PLS-DA was applied to the DART-MS data to remove the impact from potentially confounding factors such as diet, medications, and environment. Although PCA and PLS-DA failed to distinguish the cancer samples from normal, it is still possible that the DART-MS data contains valuable biochemical information. The OSC-PLS-DA model used the same dummy Y matrix as the PLS-DA model. One component in the OSC filter was chosen to remove the information in the X matrix that is uncorrelated to the class information. The model was constructed using 3 LVs.

Fig. 3a shows the score plot for the above OSC pretreated PLS model. Compared to the score plot from PCA (Fig. 2b) and PLS-DA (Supplemental Fig. S3), samples in Fig. 3a were better separated with reduced overlap along a diagonal direction between LV1 and LV2 directions. These results indicate that the OSC-processed DART-MS spectra are correlated with breast cancer and the variations caused by other effects can be reduced. Fig. 3b shows the weight plot from the OSC-PLS model, where Weight 1 corresponds to the weight of LV1. Several breast cancer-related peaks can be seen in the weight plot, including those at $m/z = 149$, 369, 445, 575, 577, and 601 are labeled in Fig. 3b. Further MS/MS experiments are needed to identify these metabolites.

Again we performed a 4-fold cross model validation. The predicted Y values performed better with control samples: there were 10 misclassified cancer samples and 6 misclassified control samples. The OSC-PLS-DA results give evidence for the presence of metabolic differences that may be detected using DART-MS. Since PLS can extract the maximum variance from two matrices and OSC is effective in focusing the analysis more exclusively on the variation of interest, it is helpful to use the OSC filter prior to PLS to combine the merits of both NMR and MS. Low-concentration molecules can be examined using the DART-MS data as the X matrix. One issue in the OSC-PLS analysis is to determine which data should be chosen as the Y matrix. From the NMR PCA score plot (Fig. 2a), the separation between breast cancer and normal samples

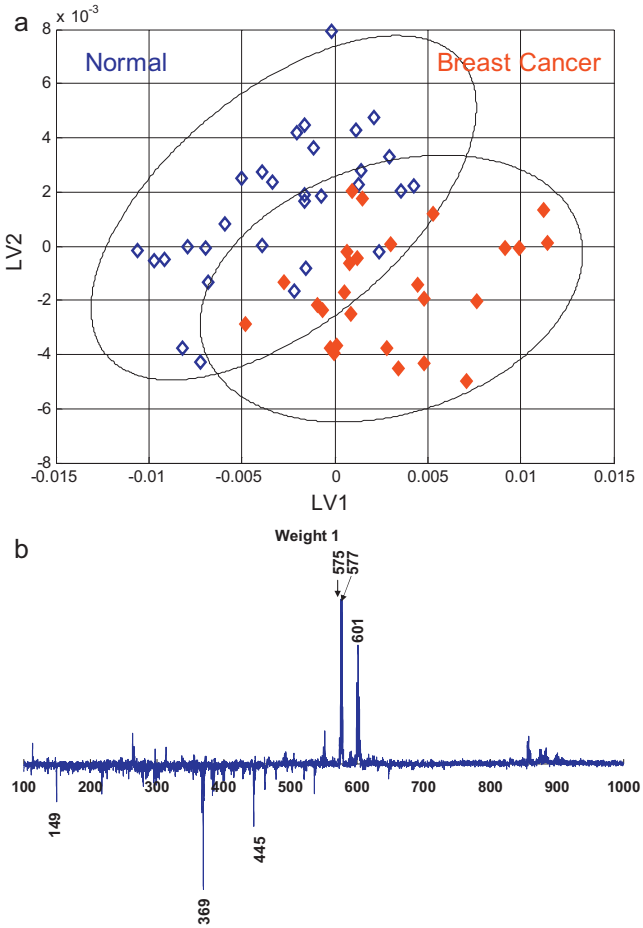

Fig. 3. OSC-PLS-DA results of the DART-MS data from 57 patient samples: (a) score plot with symbols indicating the same sample types as Fig. 2; (b) weight plot for LV1.

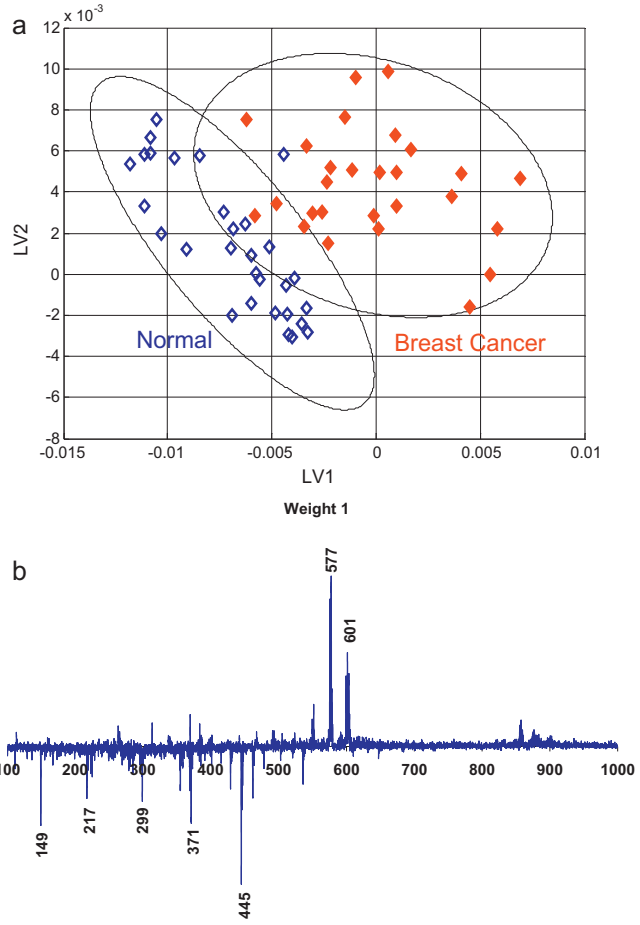

Fig. 4. The OSC-PLS results of DART-MS spectra regressed against the PC1 scores of NMR spectra: (a) score plot; (b) weight plot for LV1.

is mostly along PC1, in other words, the perturbation caused by breast cancer could be mainly represented by the NMR PC1 scores.

3.4. PC directed PLS: combined NMR and DART analysis

To take advantage of this correlation, OSC-PLS analysis was performed by regression of DART-MS against the PC1 scores from PCA of the NMR data, and the resulting score plot is shown in Fig. 4a. 4 LVs were selected. One component in the OSC filter was chosen to be excluded to remove the information in the X matrix that is uncorrelated to the response Y. It is evident that the classification in Fig. 4a is significantly improved compared to that in the DART-MS score plots in Fig. 2b (PCA), Supplemental Fig. S3 (PLS-DA) and even Fig. 3a (OSC-PLS). In Fig. 4a, only one (normal) sample deviates (outside the 95% confidence region) from either confidence region. It is interesting to notice that the classification in Fig. 4a is also better than that in Fig. 2a (PCA score plot of NMR spectra), which indicates that the metabolic profile extracted from the combination of NMR and MS should be more informative and meaningful than that from each individual analytical tool.

The peaks in the corresponding weight plot shown in Fig. 4b contribute to the classification in Fig. 4a, and several of the important peaks, at m/z 149, 217, 299, 371, 445, 577, and 601 are labeled. Many, but not all of these are the same as those in the weight plots after OSC-PLS-DA was performed on the DART-MS data alone. Collision induced dissociation (CID) MS/MS confirmation experiments are needed in order to identify the metabolites.

We then performed a 4-fold cross-model validation. The PC1 scores were computed from the NMR calibration data set with 75% of the samples in each trial. Since the Y matrix was no longer binary, we used the LV scores of the calibration data set as a training set to classify the LV scores of the test data set using k-nearest neighbor analysis. Cross-model validation resulted in 1 misclassified control sample and 3 misclassified cancer samples (see Table 1).

To indicate the effectiveness of the OSC filter in excluding information unrelated to PC1 scores of NMR spectra, the score plot of common PLS (without the OSC filter) of DART-MS spectra regression against PC1 scores of NMR spectra is shown in Fig. S4. 8 LVs were chosen for the common PLS model based on leave one out cross validation PRESS. There is only a small change in the score plot (Fig. S4) compared to the PCA score plot of DART-MS spectra (Fig. 2b), along with a certain dimensional rotation that is probably caused by the PLS regression. The classification in Fig. S4 is worse than that in Fig. 4a, which indicates that the OSC-processed DART-MS spectra are more correlated with breast cancer. There were 6 misclassified control samples and 5 misclassified cancer samples based on cross-model validation.

Fig. 5 shows a ROC graph comparing the four PLS based models. The ROC curve for one model is constructed by plotting the true positive rate against the false positive rate. Then the ROC curves for different models are superposed on the same graph to compare their performance. Furthermore the area under the ROC curve (AUC) is calculated for each model. The PC directed OSC pretreated PLS model has the largest AUC (0.99), indicating its effectiveness to separate the control and cancer classes.

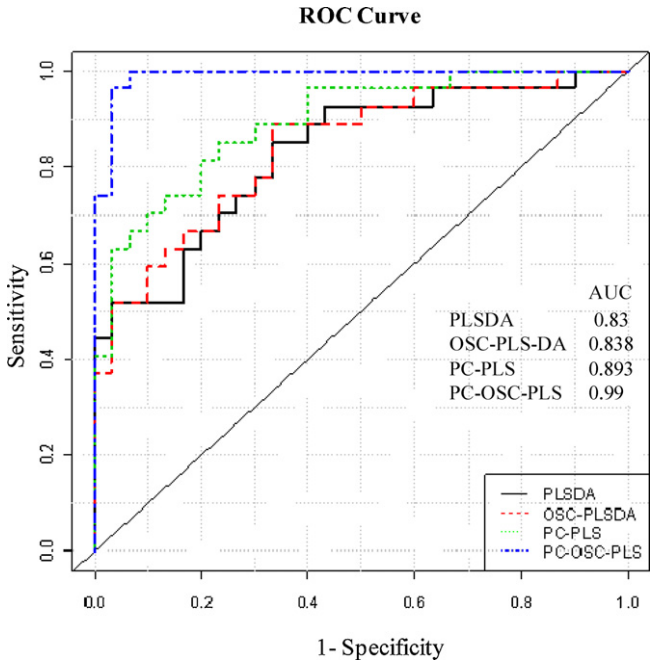

Fig. 5. ROC graphs for the comparison of four PLS based models. The PLS-DA model has AUC = 0.83; the OSC-PLS-DA model has AUC = 0.838; the PC directed common PLS model has AUC = 0.893; and the PC directed OSC pretreated PLS model has AUC = 0.99.

3.5. Further PCA analysis

We also implemented the Consensus PCA algorithm based on the work by Westerhuis et al. [49] to combine information from both datasets. We point out the fact that the DART-MS spectra have many more variables than do the NMR spectra. However, the NMR and DART-MS spectra were properly scaled in the analysis such that DART-MS spectra did not dominate the results [49]. CPCA results were worse than the results of simple PCA using one dataset. CPCA explained a small percentage of the variance in each dataset using 6 PCs and the cancer and normal scores were highly overlapped.

NMR and DART-MS spectra may contain unique breast cancer related metabolic information not found in the other dataset. Unfortunately we cannot separate cancer and normal samples and confirm the selected peaks using one dataset alone. The proposed model combines two datasets and uses PC1 scores of NMR data to supervise model building. The resulting metabolic profile may not be exhaustive but it provides important information for early breast cancer detection.

It is worth mentioning that the algorithm used in this paper, utilizing DART-MS spectra in the OSC-PLS regression against PC1 scores of NMR spectra, should have broader applicability and can be extended in metabolomics studies to correlate any two spectroscopically orthogonal data sets, such as NMR, MS, and even Raman. It is recommended to use PC(s) or LV(s) for the Y matrix to carry the variance of interest from a quantitative and reproducible analytical tool. It is also suggested that a limited number of PCs or LVs be used for the Y matrix in the regression, and they might carry less than the total variation in the corresponding spectra data. By this means, systematic errors can be greatly reduced, and one can be assured that only variations of interest are used especially when statistical methods for orthogonal exclusion are applied.

3.5.1. Further PLS analysis of NMR spectra

We also constructed PLS-DA and OSC-PLS-DA models for the NMR spectra alone using a dummy Y matrix. Again 4-fold cross-model-validation was performed for both models. There were 5 misclassified control samples and 5 misclassified cancer samples that resulted from the PLS-DA model, as well as 5 misclassified control samples and 3 misclassified cancer samples from the OSC-PLS-DA model (see Table 1). Although the performance of PLS-DA models for NMR spectra using a dummy Y matrix alone is better than that for DART spectra, we still observed some misclassified samples.

The PLS analysis of the NMR spectra suggests an alternative Y matrix. Instead of using the PC1 score of NMR spectra as Y matrix to replace the dummy 0/1 Y matrix, we can use the LV1 score of the PLS-DA model for NMR spectra or the LV1 score of the OSC-PLS-DA model for NMR spectra as the Y matrix. We then have a LV directed PLS-DA model along with a PC directed PLS-DA model.

Next we used LV1 score from OSC-PLS-DA model of NMR spectra as the Y matrix and DART spectra as X matrix to construct the OSC-PLS-DA model. The LV1 scores of the two classes have only minor overlap. This turns out to also be a good choice for the Y matrix. We performed 4-fold cross-model-validation again. There are 2 misclassified control samples and 1 misclassified cancer sample (Table 1).

Given the very similar results, we recommend using the PC1 score of the NMR spectra over the LV1 score as Y matrix, because the PC1 score is more robust than the LV1 score. This recommendation assumes there is at least some visible separation along PC1 (or possibly PC2), otherwise the LV1 score would need to be used. However, the LV1 score is subject to the individual user's choice for PLS-DA model, i.e. the number of LVs and whether to apply OSC or not. Different users can have different LV1 scores from the same dataset. LV1 score from a mis-specified model can potentially damage the performance of the LV directed PLS-DA model. On the other hand, different users will have the same PC1 score from a particular dataset. The classification results of two PCA analyses and 7 PLS based models are compared in Table 1.

Regarding the mechanism for how our new method works, we note that the binary class labels such as "cancer" and "control," i.e., the true sample classifications, have no numerical meaning themselves. Using a 1 and 0 (or 1 and −1) to represent two classes has become common practice because it seems there is no alternative method to assign numerical values to the objects in two different classes. However, given the continuum of disease states, it makes sense that a continuous variable can instead be used to model breast cancer, and this approach may provide an improved methodology to describe the disease heterogeneity. Using a PC1 score or a LV1 score from one set of spectral data allows us to accurately evaluate every object numerically. The approach is in essence similar to the construction of support vector machine target function. In this sense, a minor overlap is tolerated in the response Y to improve the final classification result.

4. Conclusions

A new method for combining NMR and MS for metabolite studies that is based on the OSC-PLS regression is illustrated in this paper. The use of complementary spectroscopies, in this case NMR which is quantitative and reproducible and DART-MS which is highly sensitive, is shown to improve classification. In this study, according to the PLS and PCA results, DART-MS did not extract as much variation related to breast cancer as did NMR. However, it was shown that the DART-MS data could be used to separate the breast cancer and normal samples in the score plot of OSC-PLS regression when the PC1 (or LV1) scores of the NMR data are used as the Y matrix. Since an OSC filter is utilized, it is anticipated that effects of other confounding factors on the spectra such as diet and medication intake are also removed or reduced, and thus an improved metabolic profile can be expected. The combination of

these two analytical techniques will have powerful capabilities in the areas such as disease detection and biomarker discovery. 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. Structural identification of the DART-MS peaks is currently underway, but is beyond the scope of the current paper. Finally, biochemical validation will be highly useful to identify the mechanisms of breast cancer development and provide further evidence for the validity of the obtained metabolite species of interest. Follow-on metabolomics studies will concentrate on the metabolite identification and on the aberrant biological processes of breast cancer.

Acknowledgments

This work is supported by the NIH/NIGMS (Grant 1R01 GM085291-01), The Purdue University Center for Cancer Research, the Purdue Research Foundation, and the Oncological Sciences Center in Discover Park at Purdue University. The authors also thank Dr. Greg Banik at Bio-Rad Laboratories for his assistance with the KnowItAll software. The authors also thank the reviewer for suggesting the use of LV1 scores as a choice for the Y matrix data.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.aca.2010.11.040.

References

[1] J.K. Nicholson, J.C. Lindon, E. Holmes, Xenobiotica 29 (1999) 1181.
[2] O. Fiehn, Plant Mol. Biol. 48 (2002) 155.
[3] J.K. Nicholson, J. Connelly, J.C. Lindon, E. Holmes, Nat. Rev. Drug Discov. 1 (2002) 153.
[4] J. Van der Greef, A.K. Smilde, J. Chemometr. 19 (2005) 376.
[5] X. Zhang, D. Wei, Y. Yap, L. Li, S. Guo, F. Chen, Mass Spectrom. Rev. 26 (2007) 403.
[6] G.A.N. Gowda, S. Zhang, H. Gu, V. Asiago, N. Shanaiah, D. Raftery, Expert Rev. Mol. Diagn. 8 (2008) 617.
[7] K. Dettmer, P.A. Aronov, B.D. Hammock, Mass Spectrom. Rev. 26 (2007) 51.
[8] J.K. Nicholson, I.D. Wilson, Nat. Rev. Drug Discov. 2 (2003) 668.
[9] S.G. Villas-Boas, S. Mas, M. Akesson, J. Smedsgaard, J. Nielsen, Mass Spectrom. Rev. 24 (2005) 613.
[10] N.J. Serkova, C.U. Niemann, Expert Rev. Mol. Diagn. 6 (2006) 717.
[11] Z. Pan, D. Raftery, Anal. Bioanal. Chem. 387 (2007) 525.
[12] H. Chen, Z. Pan, N. Talaty, R.G. Cooks, D. Raftery, Rapid Commun. Mass Spectrom. 20 (2006) 1577.
[13] R.D.L. Beger, K. Schnackenberg, R.D. Holland, D.H. Li, Y. Dragan, Metabolomics 2 (2006) 125.
[14] D.J. Crockford, E. Holmes, J.C. Lindon, R.S. Plumb, S. Zirah, S.J. Bruce, P.C. Rainville, C.L. Stumpf, J.K. Nicholson, Anal. Chem. 78 (2006) 363.
[15] H. Gu, H. Chen, Z. Pan, A.U. Jackson, N. Talaty, B. Xi, C. Kissinger, C. Duda, D. Mann, D. Raftery, R.G. Cooks, Anal. Chem. 79 (2007) 89.
[16] Z. Pan, H. Gu, N. Talaty, H. Chen, N. Shanaiah, B.E. Hainline, R.G. Cooks, D. Raftery, Anal. Bioanal. Chem. 387 (2007) 539.
[17] E.C.Y. Chan, P.K. Koh, M. Mal, P.Y. Cheah, K.W. Eu, A. Backshall, R. Cavill, J.K. Nicholson, H.C. Keun, J. Proteome Res. 8 (2009) 352.
[18] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Anal. Chem. 78 (2006) 2700.
[19] J.H. Granger, A. Bake, R.S. Plumb, J.C. Perez, I.D. Wilson, Drug Metab. Rev. 36 (2004) 252.
[20] R.G. Cooks, Z. Ouyang, Z. Takats, J.M. Wiseman, Science 311 (2005) 1566.
[21] H. Chen, A. Wortmann, R. Zenobi, J. Mass Spectrom. 42 (2007) 1123.
[22] R.B. Cody, J.A. Laramee, H.D. Durst, Anal. Chem. 77 (2005) 2297.
[23] C.Y. Pierce, J.R. Barr, R.B. Cody, R.F. Massung, A.R. Woolfitt, H. Moura, H.A. Thompson, F.M. Fernandez, Chem. Commun. 8 (2007) 807.
[24] S.X. Yu, E. Crawford, J. Tice, B. Musselman, J.T. Wu, Anal. Chem. 81 (2009) 193.
[25] R.B. Cody, Anal. Chem. 81 (2009) 1101.
[26] M. Rantalainen, O. Cloarec, O. Beckonert, I.D. Wilson, D. Jackson, R. Tonge, R. Rowlinson, S. Rayner, J. Nickson, R.W. Wilkinson, J.D. Mills, J. Trygg, J.K. Nicholson, E. Holmes, J. Proteome Res. 5 (2006) 2642.
[27] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986.
[28] J.C. Lindon, E. Holmes, J.K. Nicholson, Prog. Nucl. Magn. Reson. Spectrosc. 39 (2001) 1.
[29] S. Wold, H. Antti, F. Lindgren, J. Ohman, J. Chemometr. Intell. Lab. Syst. 44 (1998) 175.
[30] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. Van Velzen, J.P.M. Van Duijnhoven, F.A. Van Dorsten, Metabolomics 4 (2008) 81.
[31] O. Cloarec, M.E. Dumas, J. Trygg, A. Craig, R.H. Barton, J.C. Lindon, J.K. Nicholson, E. Holmes, Anal. Chem. 77 (2005) 517.
[32] C. Stella, B. Beckwith-Hall, O. Cloarec, E. Holmes, J.C. Lindon, J. Powell, F. Van der Ouderaa, S. Bingham, A.J. Cross, J.K. Nicholson, J. Proteome Res. 5 (2006) 2780.
[33] G. Musumarra, V. Barresi, D.F. Condorelli, C.G. Fortuna, S. Scirè, Comput. Biol. Chem. 29 (2005) 183.
[34] S. Wold, S. Hellberg, T. Lundstedt, M. Sjostrom, H. Wold, Proceed. Sympos. PLS Model Building: Theory and Application, Frankfurt am Main Germany, 1987, p. 23.
[35] C.L. Gavaghan, I.D. Wilson, J.K. Nicholson, FEBS Lett. 530 (2002) 191.
[36] H. Mao, M. Xu, B. Wang, H. Wang, X. Deng, D. Lin, Acta Chim. Sin. 65 (2007) 152.
[37] J. Trygg, S. Wold, J. Chemometr. 16 (2002) 119.
[38] H. Wold, in: F. David (Ed.), Research Papers in Statistics, Wiley, New York, 1996, p. 411.
[39] M. Stone, J. R. Stat. Soc. Ser. B 36 (1974) 111.
[40] E. Anderssen, K. Dyrstad, F. Westad, H. Martens, Chemometr. Intell. Lab. Syst. 84 (2006) 69.
[41] J.K. Nicholson, P.J.D. Foxall, M. Spraul, R.D. Farrant, J.C. Lindon, Anal. Chem. 67 (1995) 793.
[42] W.M.T. Fan, Prog. Nucl. Magn. Reson. Spectrosc. 28 (1996) 161.
[43] J. Feng, X. Li, F. Pei, X. Chen, S. Li, Y. Nie, Anal. Biochem. 301 (2002) 1.
[44] M.A. Constantinou, E. Papakonstantinou, D. Benaki, M. Spraul, K. Shupis, M.A. Koupparis, E. Mikros, Anal. Chim. Acta 511 (2004) 303.
[45] U. Sharma, A. Mehta, V. Seenu, N.R. Jagannathan, Magn. Reson. Imaging 22 (2004) 697.
[46] P. Shen, Y. Kang, Y. Cheng, Chem. J. Chin. Univ. 26 (2005) 1798.
[47] T.L. Whitehead, T. Kieber-Emmons, Prog. Nucl. Magn. Reson. Spectrosc. 47 (2005) 165.
[48] T.L. Whitehead, B. Monzavi-Karbassi, T. Kieber-Emmons, Metabolomics 1 (2005) 269.
[49] J.A. Westerhuis, T. Kourti, J.F. MacGregor, J. Chemometr. 12 (1998) 301.