http://dx.doi.org/10.21577/0103-5053.20170159

J. Braz. Chem. Soc., Vol. 29, No. 3, 472-481, 2018.
Printed in Brazil - ©2018 Sociedade Brasileira de Química

Article

Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry

Camilo L. M. Morais and Kássio M. G. Lima*

Química Biológica e Quimiometria, Instituto de Química,
Universidade Federal do Rio Grande do Norte, 59072-970 Natal-RN, Brazil

Mass spectrometry (MS) is a powerful technique that can provide the biochemical signature of a wide range of biological materials such as cells and biofluids. However, MS data usually has a large range of variables which may lead to difficulties in discriminatory analysis and may require high computational cost. In this paper, principal component analysis with linear discriminant analysis (PCA-LDA) and quadratic discriminant analysis (PCA-QDA) were applied for discrimination between healthy control and cancer samples (ovarian and prostate cancer) based on MS data sets. In addition, an identification of prostate cancer subtypes was performed. The results obtained herein were very satisfactory, especially for PCA-QDA. Selectivity and specificity were found in a range of 90-100%, being equal or superior to support vector machines (SVM)-based algorithms. These techniques provided reliable identification of cancer samples which may lead to fast and less-invasive clinical procedures.

Keywords: mass spectrometry, classification, ovarian cancer, prostate cancer, QDA

Introduction

Mass spectrometry (MS) is an analytical technique that is used for determining the chemical composition of a given sample, to quantify compounds,[1] and to help elucidate molecular structures.[2,3] This technique has been increasingly utilized in biomedical and clinical research,[4] since it can overcome many limitations of classical immunoassays[5,6] and supports the development of fast and less-invasive clinical procedures.[7-9]

MS is usually coupled with chromatography such as liquid chromatography (LC-MS) and gas chromatography (GC-MS). Other techniques such as surface-enhanced laser desorption ionization time-of-flight (SELDI-TOF) and matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) are often used in MS applications, including disease screening and diagnosis.[5] Some examples of MS applications includes toxicology screening and toxic drug quantification using quadrupole MS/MS;[10] identification of inborn errors in metabolism or genetic defects in newborns for prenatal screening programs using electrospray tandem MS;[11] detection of drug-induced hepatotoxicity using MS-based metabolomics;[12] and identification and quantification of bleomycin in serum and tumor tissue by high resolution LC-MS.[13] MS-based techniques have been largely employed for cancer identification, such as for breast cancer,[14] prostate cancer,[15,16] ovarian cancer,[17] lung cancer,[18] and pancreatic cancer;[19] as well as for identifying many biomarkers.[18,20-24]

One of the main fields using MS data is metabolomics, which aims to identify and quantify small molecules involved in metabolic reactions.[25] Metabolomics studies have been applied in several areas, especially cancer.[26] These analyses are typically performed in either targeted or untargeted approaches.[25] The target approach aims to identify and quantify specific metabolites or metabolite class; whereas in the untargeted analysis a new hypothesis for further tests is generated by measuring all the metabolites in a biological system.[25] To make this possible, multivariate statistical analysis is commonly employed in metabolomics studies by means of unsupervised or supervised classification techniques.[25]

Various types of chemometric algorithms have been reported for pattern recognition and classification of MS data, especially for discriminating between healthy control and cancer samples, or discriminating cancer subtypes. For instance, there are several papers reporting the use of partial least squares discriminant analysis (PLS-DA),[14,18,20] hierarchical cluster analysis (HCA),[14,27,28] principal component analysis (PCA),[14,29] support vector machines

*e-mail: kassiolima@gmail.com

(SVM),[17,29] artificial neural networks (ANN),[28] principal-component analysis followed by linear discriminant analysis (PCA-LDA),[15] principal component directed partial least squares (PC-PLS),[30] and backward variable elimination partial least squares discriminant analysis (BVE-PLSDA).[31]

Principal component analysis (PCA) is a method of exploratory analysis capable of reducing the original data into a few variables.[32] PCA reduces the data into a few principal components (PCs), where each one represents a piece of the original information. The first PC has the largest explained variance; therefore, they represent most of the information present in the original data. Using PCA, for instance, it is possible to reduce a large MS data set of thousands of variables into a few PCs representing the majority of the original information in just a few seconds. The PCA scores can be used as discriminant variables in conjunction with supervised classification techniques, such as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). LDA is one of the most common algorithms used in supervised classification of 1st order spectral data, especially for spectroscopy applications in discriminatory analysis of cancer samples.[33] On the other hand, there are only a few applications of QDA algorithm for discriminatory analysis reported in literature, and even fewer for QDA coupled to other chemometric techniques.[33] QDA is a very simple algorithm, and differently from LDA, it computes the variance structures for each class separately,[34] creating a more powerful discrimination rule for classes with different covariance matrices, such as for biological spectra sets in which the variability within classes is a key issue.

LDA has been reported in many MS applications, including analysis of *N*-glycans of human serum α_1-acid glycoprotein (AGP) in cancer and healthy individuals;[35] differentiation of vegetable oils;[36] ovarian cancer detection based on proteomics;[37] estimating false discovery rate (FDR) in phosphopeptide identifications;[38] discrimination of ionic liquid types (ILs);[39] and gasoline classification.[40] QDA applications are fewer, and include characterization of ILs;[39] identification of ovarian cancer;[41] and gasoline classification.[40]

In this paper, principal component analysis followed by linear discriminant analysis (PCA-LDA) and quadratic discriminant analysis (PCA-QDA) were compared for discrimination between healthy controls and cancer (ovarian and prostate) samples. In addition, a further classification between benign subtypes of prostate cancer (serum PSA (prostate-specific antigen) 4-10 ng mL^{-1} and serum PSA > 10 ng mL^{-1}) was performed. These algorithms take advantage of the power of MS-based techniques for clinical analysis and provide a simple, fast, and reliable way to identify cancer samples.

Experimental

Samples

Data set 1: ovarian cancer

This data set is public available by Guan *et al.*[17] It is composed of LC/TOF-MS mass spectra (positive mode) from 35 healthy control (H.C.) and 37 ovarian cancer (O.C.) samples based on serum metabolomics. Retention time was not considered as a factor for chemometric modeling, thus the entire mass spectra (*m/z* values varying from 134.9919 to 1.4879×10^3, having 360 variables) was integrated into an interval of retention time of 0-180 min. The control population consisted of patients with histology considered within normal limits and women with non-cancerous ovarian conditions; and the ovarian cancer samples were composed of patients with papillary serous ovarian cancer (stage I-IV). More details about the sample acquisition can be found in Guan *et al.*[17]

Data set 2: prostate cancer

This data set is public available by Petricoin III *et al.*[16] It is composed of SELDI-TOF mass spectra from 63 healthy control (H.C.) and 69 prostate cancer (P.C.) samples based on serum proteomics. The *m/z* values varied from 0 to 1.9996×10^4, having 15,153 variables. The control population was composed of men with no previous history of prostate cancer and serum PSA < 1 ng mL^{-1}. The prostate cancer samples were acquired from patients with serum PSA ≥ 4 ng mL^{-1}, digital rectal exam (DRE) evidence and single sextant biopsy evidence of prostate cancer (Gleason scores 4-9). More details about the sample acquisition can be found in Petricoin III *et al.*[16]

Data set 3: subtypes of prostate cancer

This data set was also obtained from Petricoin III *et al.*[16] It is composed of SELDI-TOF mass spectra from 26 prostate cancer samples with PSA 4-10 ng mL^{-1} (low grade) and 43 prostate cancer samples with PSA > 10 ng mL^{-1} (high grade). These data are derived from data set 2 (*m/z* values varying from 0 to 1.9996×10^4, having 15,153 variables) and more details about the sample acquisition can be found in Petricoin III *et al.*[16]

Computational analysis

The data treatment and chemometric analysis were performed using MATLAB® software R2012b[42]

(MathWorks, USA) with PLS Toolbox 7.9.3 (Eigenvector Research, Inc., USA). All data sets were normalized by Euclidian norm and baseline corrected using automatic Whittaker filter ($\lambda = 100$, $p = 0.001$).[43] Data sets 2 and 3 were mass drift corrected by using the icoshift algorithm[44,45] in the *m/z* range of 3000-10000. Mean-centering scaling was applied to the data before chemometric modelling.

The samples for each data set were divided into training (ca. 70%), validation (ca. 15%) and prediction (ca. 15%) sets by using the Kennard-Stone uniform sample selection algorithm.[46] Table 1 summarizes the number of samples for training, validation and prediction in each data set.

Table 1. Number of samples in the training, validation and prediction sets for each data set

	Training	Validation	Prediction
Data set 1	50	10	12
Data set 2	92	19	21
Data set 3	48	10	11

The chemometric models of PCA-LDA and PCA-QDA were built by firstly performing a principal component analysis (PCA),[32] and then the A firstly scores selected were utilized as classification variables in a linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) model. The LDA classification score (L_{ik}) and the QDA classification score (Q_{ik}) are calculated for a given class k by the following equations:[47,48]

$$L_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T \sum\nolimits_{pooled}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) - 2\log_e \pi_k \quad (1)$$

$$Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T \sum\nolimits_{k}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) + \log_e |\Sigma_k| - 2\log_e \pi_k \quad (2)$$

where $\mathbf{x}_i$ is the vector containing the classification variables for sample i; $\bar{\mathbf{x}}_k$ is the mean vector of class k; Σ_{pooled} is the pooled covariance matrix; and π_k is the prior probability of class k. The pooled covariance matrix Σ_{pooled} and the prior probability π_k are calculated as follows:[47,48]

$$\Sigma_{pooled} = \frac{1}{n}\sum\nolimits_{k=1}^{K} n_k \Sigma_k \quad (3)$$

$$\pi_k = \frac{n_k}{n} \quad (4)$$

where n is the total number of objects in the training set; K is the number of classes; n_k is the number of objects of class k; and Σ_k is the variance-covariance matrix of class k, estimated by:[48]

$$\Sigma_k = \frac{1}{n_k - 1}\sum\nolimits_{i=1}^{n_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T \quad (5)$$

The LDA and QDA classification scores (equations 1 and 2, respectively) were calculated based on the Mahalanobis distance modified by the fraction of samples in each class. In that case, they do not depend of scale, thus being dimensionless. These scores were used to calculate the discriminant function (DF) between the two classes as follows:[48]

$$DF_{LDA} = L_{i1} - L_{i2} \quad (6)$$

$$DF_{QDA} = Q_{i1} - Q_{i2} \quad (7)$$

where Q_{i1} and Q_{i2} are the quadratic classification scores for classes 1 and 2, respectively.

If the DF result is positive for a given sample, the sample is closer to class 2, therefore it is classified as class 2; and if the DF result is negative for a given sample, the sample is closer to class 1, therefore being classified as class 1. In this sense, on the DF plot the class 2 is constituted of all positive values; whereas class 1 is constituted of all negative values. A flowchart illustrating the MS data processing is shown in Figure 1.

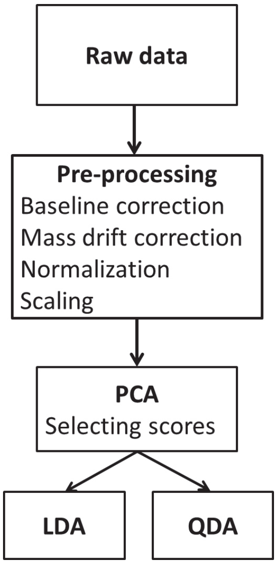

Figure 1. Flowchart illustrating MS data processing.

Although both LDA and QDA are based on a Mahalanobis distance calculation, the QDA algorithm forms a separated variance model for each class, not assuming that classes have similar variance-covariance matrices as LDA does.[34] Therefore, QDA is more suitable to build classification models of data having different variance structures, such as what happens in many biological data sets.

Quality performance

The performances of the employed algorithms were evaluated according to the following quality metrics: accuracy, sensitivity, specificity, positive and negative predictive value, Youden's index, and positive and negative likelihood ratios. Accuracy is related to the percentage of correct classification;[49] sensitivity (SENS) is the confidence that a positive result for a sample of the labeled class is obtained; specificity (SPEC) is the confidence that a negative result for a sample of the non-labeled class is obtained; positive predictive value (PPV) measures the proportion of positives that are correctly assigned; negative predictive value (NPV) measures the proportion of negatives that are correctly assigned; Youden's index (YOU) evaluates the classifier's ability to avoid failure; positive likelihood ratio (LR+) is the ratio between the probability of predicting an example as positive when it is truly positive and the probability of predicting an example as positive when it is not positive; and negative likelihood ratio (LR–) is the ratio between the probability of predicting an example as negative when it is actually positive and the probability of predicting an example as negative when it is truly negative.[33] The equations of these quality parameters are shown in Table 2.

Table 2. Quality parameters

Parameter	Equation
Accuracy / %	$100 - \frac{y}{N} \times 100$
Sensitivity / %	$\left(\frac{TP}{TP + FN}\right) \times 100$
Specificity / %	$\left(\frac{TN}{TN + FP}\right) \times 100$
Positive predictive value / %	$\left(\frac{TP}{TP + FP}\right) \times 100$
Negative predictive value / %	$\left(\frac{TN}{TN + FN}\right) \times 100$
Youden's index / %	SENS – (100 – SPEC)
Positive likelihood ratio	$\frac{SENS}{100 - SPEC}$
Negative likelihood ratio	$\frac{SPEC}{100 - SENS}$

y = total number of samples incorrectly classified for a set of N samples; TP: true positive; TN: true negative; FP: false positive; FN: false negative; SENS: sensitivty; SPEC: specificity.

Results and Discussion

Data set 1: ovarian cancer

Ovarian cancer encompasses a heterogeneous group of tumors having differences in epidemiological and genetic risk factors, precursor lesions, spread patterns, molecular events during oncogenesis, response to chemotherapy and prognosis. Most ovarian cancers (90%) are malignant epithelial tumors named carcinomas, and the remaining are germ cells and sex cord-stromal tumors.[50] This type of cancer is the leading cause of death from gynecological malignances, and its mortality is a consequence of late presentation and diagnosis at stages III or IV, resulting in five-year survival rates of 20 and 6%, respectively.[33] A study using serum metabolomics by MS-based techniques could lead to a faster and more robust classification of cancer and non-cancer patients. In this data set, the baseline corrected LC/TOF-MS mass spectra of healthy control (H.C.) and ovarian cancer (O.C.) samples are shown in Figure 2a. As can be seen, the signals are very superposed and no visual differentiation between H.C. and O.C. can be made.

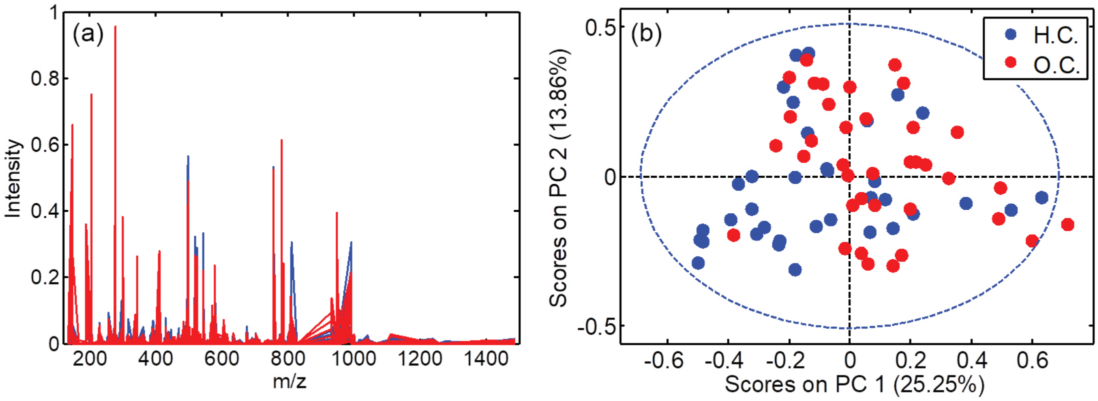

Figure 2. (a) Baseline corrected mass spectra for healthy control (H.C.) and ovarian cancer (O.C.) samples; (b) PCA scores on PC1 *versus* scores on PC2 for healthy control (H.C.) and ovarian cancer (O.C.) samples, where the percentage of total variance described by each PC is described inside parenthesis. The circled blue line is the confidence ellipse of 95%.

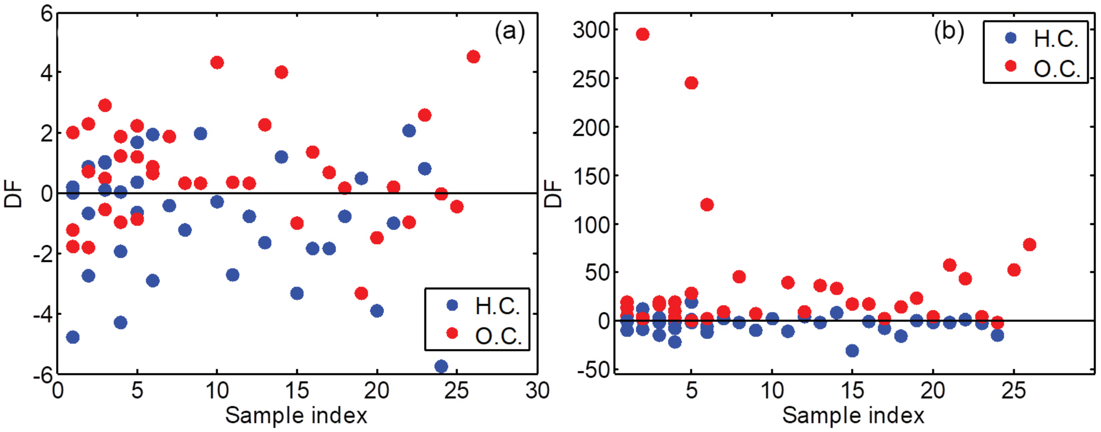

Figure 3. DF plot for (a) PCA-LDA and (b) PCA-QDA models for discriminating healthy control (H.C.) and ovarian cancer (O.C.) samples. The DF scale for the QDA-based models were zoomed to improve visualization.

Using PCA for exploratory analysis of this data set, the scores plot on the 1st and 2nd PCs is depicted in Figure 2b. Although PCA technique could be used as a classification tool, the lack of discrimination pattern in this scores plot leads to the use of supervised discriminant analysis. PCA-LDA and PCA-QDA were applied to the 10 first PCs (cumulative explained variance of 86.33%) and its DF plots are shown in Figures 3a and 3b, respectively. These figures show a better discriminant pattern for differentiating H.C. and O.C. samples.

The PCA-QDA DF plot also suggests a difference in variance structures between the classes, where the ovarian cancer sample set has a higher covariance matrix since this class has higher DF values than the other. This is probably caused by the high complexity of ovarian cancer disease as mentioned earlier. The quality performance parameters found for these chemometric models are shown in Table 3.

As shown in Table 3, the best quality parameters were obtained for PCA-QDA (accuracy in prediction set = 91.67%). On the other hand, PCA-LDA only achieved accuracy of 58.33% in prediction and 30% in the validation set. The low accuracy in the validation set suggests that the model is not well fitted, reflecting its poor prediction ability. PCA-QDA probably had superior performance because the classes' variance structures are very different due to the high composition variability of the ovarian cancer samples, which increases the power of QDA compared to LDA. The accuracy in prediction set of PCA-QDA is close to what was obtained in literature using SVM, a more robust algorithm.[17] Sensitivity and specificity were also equal to 91.67%, being superior to the results achieved by linear and non-linear SVM classifiers applied to this data set (sensitivity = 78.4 and 83.8%, respectively; and specificity = 74.3 and 77.1%, respectively).[17] In addition, the classification results using PCA-QDA were superior than those ones found by applying PCA-SVM using a radial bases function (RBF) kernel to this data set. PCA-SVM shown accuracy, sensitivity and specificity all equal to 75%, therefore being an algorithm with intermediary performance between PCA-LDA and PCA-QDA to classify H.C. and O.C. samples. Moreover, the high value of LR+ and the low value of LR– prove that PCA-QDA is superior for identifying cancer, since these parameters are directly related to the clinical concept of "ruling-OUT" and "ruling-IN" disease, respectively.[33]

Table 3. Quality performance parameters found for PCA-LDA and PCA-QDA models for discriminating healthy control and ovarian cancer samples.

Parameter	Model	
	PCA-LDA	PCA-QDA
Accuracy		
Training set / %	70.00	84.00
Validation set / %	30.00	70.00
Prediction set / %	58.33	91.67
Sensitivity / %	58.33	91.67
Specificity / %	58.33	91.67
PPV / %	58.33	91.67
NPV / %	58.33	91.67
YOU / %	16.67	83.33
LR+	1.40	11.00
LR–	0.71	0.09

PCA-LDA: principal component analysis with linear discriminant analysis; PCA-QDA: principal component analysis with quadratic discriminant analysis; PPV: positive predictive value; NPV: negative predictive value; YOU: Youden's index; LR+: positive likelihood ratio; LR–: negative likelihood ratio.

From 360 variables present in this MS data set, only 31 were found to be statistical significant between the two classes ($p < 0.05$) (see Figure S7 in Supplementary Information (SI)). Among these variables, seven presented mean intensity variations (ΔI) higher than 1%. These *m/z* values were 279.1263 (ΔI = –14.71%), 496.3121 (ΔI = 6.97%), 496.3139 (ΔI = 8.99%), 520.3164 (ΔI = 4.98%), 520.3169 (ΔI = 4.59%), 524.3463 (ΔI = 4.33%) and 991.6178 (3.75%). The negative signal implies that the peak is more intense in O.C. class, while the positive signal implies that the peak is more intense in H.C. class. The *m/z* values of 496.3121, 496.3139, 520.3164, 520.3169 and 524.3463 are associated with types of lysophosphatidylcholine (LysoPC),[51] a metabolite identified in plasma that is directly related to the presence of ovarian cancer.[52] The other *m/z* values have not been reported or associated with any cancer metabolite according to the Human Metabolome Database (HMDB).[51]

Data set 2: prostate cancer

Prostate cancer is the most commonly diagnosed male malignant cancer in the world. It has an incidence rate of 214 cases *per* 100,000, and a mortality rate from metastatic disease of 30 in 100,000.[53] Prostate tissue is structurally complex, being primarily constituted of glandular ducts lined by epithelial cells and supported by heterogeneous stroma. Its identification is very invasive and analyst-dependent, being subject to intra- and inter-observer errors.[54] A study using serum proteomics by MS-based techniques could lead to a faster and more robust classification of cancer and non-cancer patients. In this data set, SELDI-TOF mass spectra of healthy control (H.C.) and prostate cancer (P.C.) samples were utilized. Figure 4a shows the baseline corrected mass spectra for these two classes. The signal complexity present in Figure 4a shows how difficult it is to differentiate one class from another, therefore requiring pattern recognition algorithms. Initially, PCA was utilized as exploratory analysis, and its scores plot is shown in Figure 4b.

No clear discriminant pattern is observed in the PCA scores graph. On the other hand, the results improved significantly by applying LDA and QDA to the PCA scores. PCA-LDA and PCA-QDA DF plots are shown in Figures 5a and 5b, respectively. 10 PCs were utilized (cumulative explained variance of 81.11%) for classification.

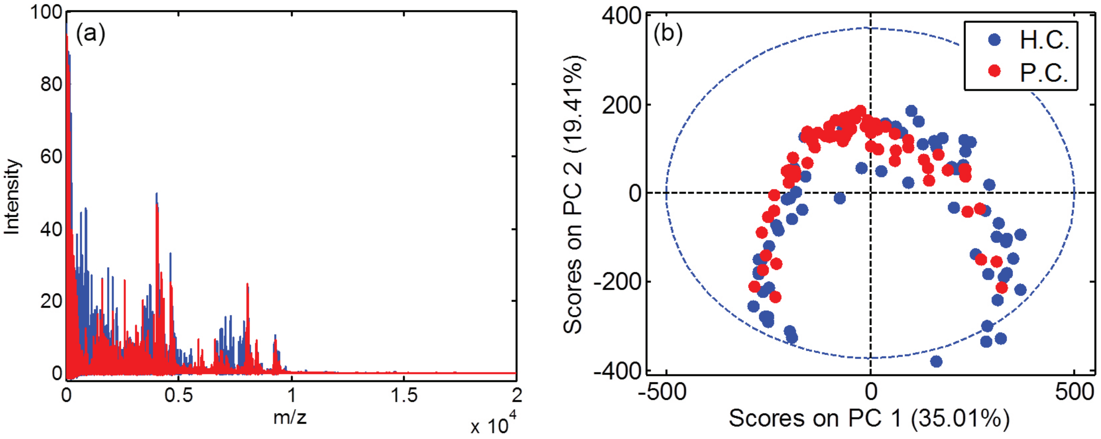

Figure 4. (a) Baseline corrected mass spectra for healthy control (H.C.) and prostate cancer (P.C.) samples; (b) PCA scores on PC1 *versus* scores on PC2 for healthy control (H.C.) and prostate cancer (P.C.) samples, where the percentage of total variance described by each PC is described inside parenthesis. The circled blue line is the confidence ellipse of 95%.

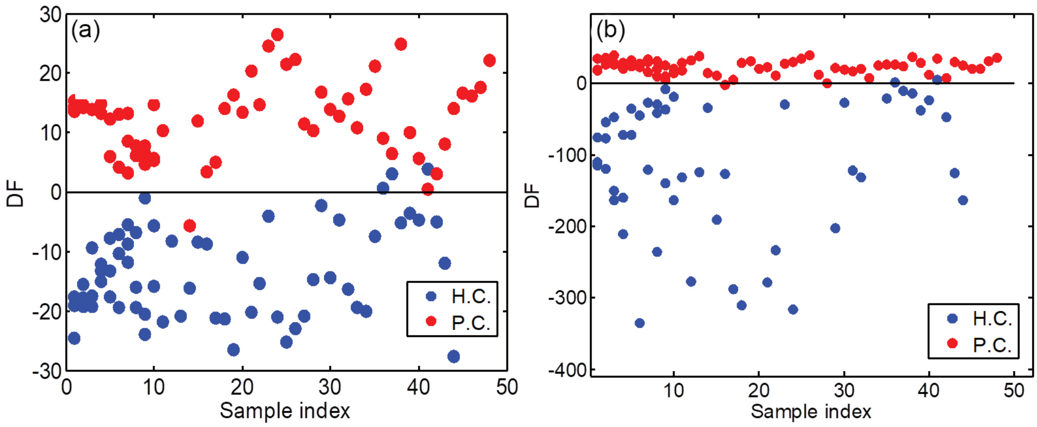

Figure 5. DF plot for (a) PCA-LDA and (b) PCA-QDA models for discriminating healthy control (H.C.) and prostate cancer (P.C.) samples. The DF scale for the QDA-based models were zoomed to improve visualization.

Figure 5 shows a clear separation between the two classes using both PCA-LDA and PCA-QDA, where PCA-QDA had a slightly better classification. As seen in the DF plot of PCA-QDA, the healthy control samples have a higher variance structure than prostate cancer samples. This variability within this biological class may be related to different habits and lifestyles of the patients.[53] The quality performance parameters found for PCA-LDA and PCA-QDA models are shown in Table 4.

Table 4. Quality performance parameters found for PCA-LDA and PCA-QDA models for discriminating healthy control and prostate cancer samples

Parameter	Model	
	PCA-LDA	PCA-QDA
Accuracy		
Training set / %	95.65	96.74
Validation set / %	100	100
Prediction set / %	100	100
Sensitivity / %	100	100
Specificity / %	100	100
PPV / %	100	100
NPV / %	100	100
YOU / %	100	100
LR+	Inf	Inf
LR–	0	0

PCA-LDA: principal component analysis with linear discriminant analysis; PCA-QDA: principal component analysis with quadratic discriminant analysis; PPV: positive predictive value; NPV: negative predictive value; YOU: Youden's index; LR+: positive likelihood ratio; LR–: negative likelihood ratio; Inf: infinite.

Table 4 shows the notable performance of the tested algorithms. PCA-LDA and PCA-QDA had accuracy in the prediction set of 100%, being 5% above the value found in literature for prostate cancer detection based on this data set.[16] The LR+ values equal to infinite are a consequence of LR+ equation shown in Table 2, because when the specificity is close to 100%, this parameter tends to infinite. The sensitivity and specificity of PCA-LDA and PCA-QDA were equal to 100%, being above the values found using a bioinformatics algorithm based on cluster analysis of topological feature maps (sensitivity = 95%, specificity = 71%).[16] Using PCA-SVM with RBF kernel, an accuracy, sensitivity and specificity of 100% were also found. However, the complexity degree employed during SVM is much higher than LDA and QDA, meaning that with simpler algorithms the same classification performance can be obtained.

From a total of 15,153 variables in the original data, 5,583 were found to be statistical significant between the two classes ($p < 0.05$) (see Figure S8 in SI). The larger number of variables as well as the untargeted procedure and the complexity of this proteomic data make nearly impossible to identify important molecules based on these 5,583 variables.

The use of PCA-LDA and PCA-QDA in this data set of serum proteomics provides a reliable, non-analyst dependent and less-invasive differentiation between patients with no evidence of prostate cancer and patients with prostate cancer. This can be a powerful tool for clinical screening, avoiding patients to suffer unnecessary surgical procedures, for instance.

Data set 3: subtypes of prostate cancer

This data set is derived from data set 2, where the cancer samples were divided into two classes: class 1 having cancer samples with serum PSA 4-10 ng mL^{1} (low grade); and class 2 having cancer samples with serum PSA > 10 ng mL^{-1} (high grade). This data set was created to evaluate the power of the algorithms to differentiate cancer samples according to its stage. Although PSA is not a final indicator of prostate cancer, it is important to differentiate low and high PSA levels, since the PSA indicates during clinical screening if a patient will need a more robust/invasive exam or not. Usually, patients with low PSA levels but with suspicion of prostate cancer undergo an additional DRE exam. However, it is recommended that patients with high PSA levels undergo additional DRE tests, such as transrectal ultrasound and cystoscopy.[55,56] The baseline corrected mass spectra of the low grade and high grade cancer samples are shown in Figure 6a.

Figure 6b shows the PCA scores for low and high grade samples, where no discriminant profile is seen. By applying PCA-LDA and PCA-QDA to the data (10 PCs, cumulative explained variance of 86.29%), the differentiation between the two classes improves significantly, as shown in Figures 7a and 7b, respectively. An almost perfect separation between the two classes is obtained in the PCA-QDA DF plot.

The coefficients in the PCA-QDA DF plots show that the variances of the low and high grade classes are similar to each other, with a bit higher covariance matrix for the high grade samples. Table 5 shows the quality parameters found by the chemometric models applied to this data set.

For classification purposes, the PCA-LDA and PCA-QDA models had very similar performances, with sensitivity and specificity of 100% each. The training ability of PCA-QDA was better than PCA-LDA, but the algorithm had worst performance in the validation set. The poorer classification in the training and validation set for both algorithms when compared to the prediction set is a possible result of the reduced number of samples. Nevertheless, the maximum results obtained in the prediction set with

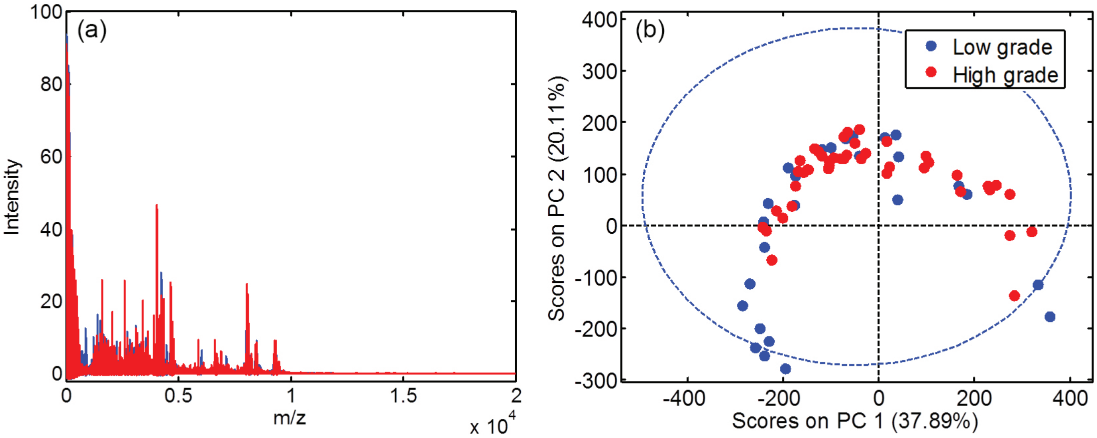

Figure 6. (a) Baseline corrected mass spectra for low grade prostate cancer and high grade prostate cancer samples; (b) PCA scores on PC1 *versus* scores on PC2 for low grade and high grade prostate cancer samples, where the percentage of total variance described by each PC is described inside parenthesis. The circled blue line is the confidence ellipse of 95%.

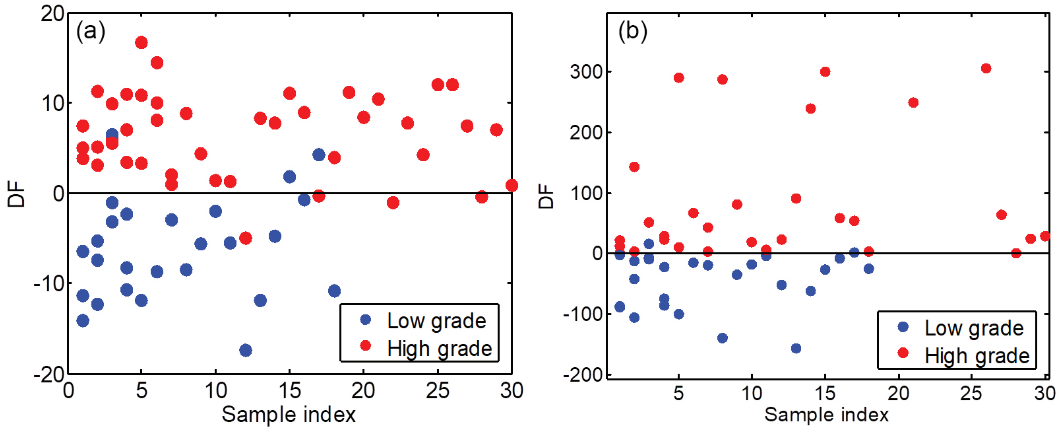

Figure 7. DF plot for (a) PCA-LDA and (b) PCA-QDA models for discriminating low grade and high grade prostate cancer samples. The DF scale for the QDA-based models were zoomed to improve visualization.

Table 5. Quality performance parameters found for PCA-LDA and PCA-QDA models for discriminating low and high grade prostate cancer samples

Parameter	Model	
	PCA-LDA	PCA-QDA
Accuracy		
Training set / %	87.50	97.92
Validation set / %	90.00	80.00
Prediction set / %	100	100
Sensitivity / %	100	100
Specificity / %	100	100
PPV / %	100	100
NPV / %	100	100
YOU / %	100	100
LR+	Inf	Inf
LR–	0	0

PCA-LDA: principal component analysis with linear discriminant analysis; PCA-QDA: principal component analysis with quadratic discriminant analysis; PPV: positive predictive value; NPV: negative predictive value; YOU: Youden's index; LR+: positive likelihood ratio; LR–: negative likelihood ratio; Inf: infinite.

PCA-LDA and PCA-QDA provided good quality metrics, showing the ability of both algorithms to differentiate stages of prostate cancer based on its PSA level.

From a total of 15,153 variables present in the original data, 2,765 were found to be statistical significant ($p < 0.05$) (see Figure S9 in SI). As occurred in data set 2, the larger number of variables combined with the untargeted procedure and the complexity of this proteomic data inhibit the identification of important molecules based on these 2,765 variables.

The performance of PCA-LDA and PCA-QDA algorithms were equal to PCA-SVM using RBF kernel (accuracy, sensitivity and specificity of 100%), showing the capability of PCA-LDA and PCA-QDA to properly classify this data set.

Conclusions

The use of PCA-LDA and PCA-QDA provided very satisfactory classification models for MS data, as demonstrated for MS-based serum metabolomics in the detection of ovarian cancer; and also MS-based serum proteomics for the detection of prostate cancer and its subtypes according to the PSA level. The LDA and QDA-based algorithms are very simple compared to many other algorithms utilized in literature, such as SVM, and can also provide very solid classification results; especially PCA-QDA, which models the data considering different variance structures between the classes. Apart from the very satisfactory classification results found for the tested data sets (sensitivity and specificity > 90%), these algorithms also significantly reduce the data, which considerably speeds up the computational analysis, enabling a supervised classification of an MS data set of thousands of variables in less than one minute, for example. The speed and solid classification results found by these algorithms for the tested applications show that they combine very well with the power of MS-based techniques, thus being capable to be utilized in other types of applications in the future. The combination of MS-based serum analysis and these types of chemometric techniques can provide very acceptable findings for developing fast, very accurate, less-invasive, and non-analysis dependent clinical procedures, especially for screening purposes.

Supplementary Information

Supplementary information is available free of charge at http://jbcs.sbq.org.br as PDF file.

Acknowledgments

Camilo L. M. Morais would like to acknowledge the financial support from CAPES/PPGQ/UFRN for his research grant. K. M. G. Lima acknowledges the CNPq grant (305962/2014-0) for financial support.

References

1. Vogester, M.; Seger, C.; *Clin. Biochem.* **2016**, *49*, 947.
2. Kind, T.; Fiehn, O.; *Bioanal. Rev.* **2010**, *2*, 23.
3. El-Aneed, A.; Banoub, J.; *Rapid Commun. Mass Spectrom.* **2005**, *19*, 1683.
4. Strathmann, F. G.; Hoofnagle, A. N.; *Am. J. Clin. Pathol.* **2011**, *136*, 609.
5. Meng, Q.; *J. Clin. Exp. Pathol.* **2013**, *S6*, e001.
6. Soldin, S. J.; Soukhova, N.; Janicic, N.; Jonklaas, J.; Soldin, O. P.; *Clin. Chim. Acta* **2005**, *358*, 113.
7. Jannetto, P. J.; Langman, L. J.; *Clin. Biochem.* **2016**, *49*, 1032.
8. Krone, N.; Hughes, B. A.; Lavery, G. G.; Stewart, P. M.; Arlt, W.; Shackleton, C. H. L.; *J. Steroid Biochem. Mol. Biol.* **2010**, *121*, 496.
9. Geyer, P. E.; Kulak, N. A.; Pichler, G.; Holdt, L. M.; Teupser, D.; Mann, M.; *Cell Syst.* **2016**, *2*, 185.
10. Liotta, E.; Gottardo, R.; Bertaso, A.; Polettini, A.; *J. Mass Spectrom.* **2010**, *45*, 261.
11. Rashed, M. S.; Bucknall, M. P.; Little, D.; Awad, A.; Jacob, M.; Alamoudi, M.; Alwattar, M.; Ozand, P. T.; *Clin. Chem.* **1997**, *43*, 1129.
12. Dahab, A. A.; Smith, N. W.; *Anal. Methods* **2012**, *4*, 1887.
13. Kosjek, T.; Krajnc, A.; Gornik, T.; Zigon, D.; Groselj, A.; Sersa, G.; Cemazar, M.; *Talanta* **2016**, *160*, 164.
14. Willmann, L.; Schlimpert, M.; Hirschfeld, M.; Erbes, T.; Neubauer, H.; Stickeler, E.; Kammerer, B.; *Anal. Chim. Acta* **2016**, *925*, 34.
15. Kerian, K. S.; Jarmusch, A. K.; Pirro, V.; Koch, M. O.; Masterson, T. A.; Cheng, L.; Cooks, R. G.; *Analyst* **2015**, *140*, 1090.
16. Petricoin III, E. F.; Ornstein, D. K.; Paweletz, C. P.; Ardekani, A.; Hackett, P. S.; Hitt, B. A.; Velassco, A.; Trucco, C.; Wiegand, L.; Wood, K.; Simone, C. B.; Levine, P. J.; Linehan, W. M.; Emmert-Buck, M. R.; Steinberg, S. M.; Kohn, E. C.; Liotta, L. A.; *J. Natl. Cancer Inst.* **2002**, *94*, 1576.
17. Guan, W.; Zhou, M.; Hampton, C. Y.; Benigno, B. B.; Walker, L. D.; Gray, A.; McDonald, J. F.; Fernández, F. M.; *BMC Bioinf.* **2009**, *10*, 259.
18. Callejón-Leblic, B.; García-Barrera, T.; Grávalos-Guzmán, J.; Pereira-Veja, A.; Gómez-Ariza, J. L.; *J. Proteomics* **2016**, *145*, 197.
19. Hingorani, S. R.; Petricoin, E. F.; Maitra, A.; Rajapakse, V.; King, C.; Jacobetz, M. A.; Ross, S.; Conrads, T. P.; Veenstra, T. D.; Hitt, B. A.; Kawaguchi, Y.; Johann, D.; Liotta, L. A.; Crawford, H. C.; Putt, M. E.; Jacks, T.; Wright, C. V. E.; Hruban, R. H.; Lowy, A. M.; Tuveson, D. A.; *Cancer Cell* **2003**, *4*, 437.
20. Zhang, Y.; Liu, Y.; Li, L.; Wei, J.; Xiong, S.; Zhao, Z.; *Talanta* **2016**, *150*, 88.
21. Crutchfield, C. A.; Thomas, S. N.; Sokoll, L. J.; Chan, D. W.; *Clin. Proteomics* **2016**, *13*, 1.
22. Wu, L.; Qu, X.; *Chem. Soc. Rev.* **2015**, *44*, 2963.
23. Bergman, N.; Bergquist, J.; *Analyst* **2014**, *139*, 3836.
24. Taguchi, A.; Hanash, S. M.; *Clin. Chem.* **2013**, *59*, 119.
25. Zhou, B.; Xiao, J. F.; Tuli, L.; Ressom, H. W.; *Mol. BioSyst.* **2012**, *8*, 470.
26. Armitage, E. G.; Barbas, C.; *J. Pharm. Biomed. Anal.* **2014**, *87*, 1.
27. Sauer, S.; Kliem, M.; *Nat. Rev. Microbiol.* **2010**, *8*, 74.
28. Lasch, P.; Drevinek, M.; Nattermann, H.; Grunow, R.; Stämmler, M.; Dieckmann, R.; Schwecke, T.; Naumann, D.; *Anal. Chem.* **2010**, *82*, 8464.
29. Al Masoud, N.; Xu, Y.; Nicolaou, N.; Goodacre, R.; *Anal. Chim. Acta* **2014**, *840*, 49.

30. Gu, H.; Pan, Z.; Xi, B.; Asiago, V.; Musselman, B.; Raftery, D.; *Anal. Chim. Acta* **2011**, *686*, 57.
31. Deng, L.; Gu, H.; Zhu, J.; Gowda, G. A. N.; Djukovic, D.; Chiorean, E. G.; Raftery, D.; *Anal. Chem.* **2016**, *88*, 7975.
32. Bro, R.; Smilde, A. K.; *Anal. Methods* **2014**, *6*, 2812.
33. Siqueira, L. F. S.; Lima, K. M. G.; *Analyst* **2016**, *141*, 4833.
34. Dixon, S. J.; Brereton, R. G.; *Chemom. Intell. Lab. Syst.* **2009**, *95*, 1.
35. Imre, T.; Kremmer, T.; Héberger, K.; Molnáz-Szöllosi, E.; Ludányi, K.; Pócsfalvi, G.; Malorni, A.; Drahos, L.; Vékey, K.; *J. Proteomics* **2008**, *71*, 186.
36. Jakab, A.; Nagy, K.; Héberger, K.; Vékey, K.; Forgács, E.; *Rapid Commun. Mass Spectrom.* **2002**, *16*, 2291.
37. Hong, Y.; Wang, X.; Shen, D.; Zhen, S.; *Acta Pharmacol. Sin.* **2008**, *29*, 1240.
38. Du, X.; Yang, F.; Manes, N. P.; Stenoien, D. L.; Monroe, M. E.; Adkins, J. N.; States, D. J.; Purvine, S. O.; Camp II, D. G.; Smith, R. D.; *J. Proteome Res.* **2008**, *7*, 2195.
39. González-Álvarez, J.; Mangas-Alonso, J. J.; Arias-Abrodo, P.; Gutiérrez-Álvarez, M. D.; *Anal. Bioanal. Chem.* **2014**, *406*, 3149.
40. Doble, P.; Sandercock, M.; Pasquier, E. D.; Petocz, P.; Roux, C.; Dawson, M.; *Forensic Sci. Int.* **2003**, *132*, 26.
41. Liu, Y.; *Comput. Biol. Med.* **2009**, *39*, 818.
42. https://www.mathworks.com/, accessed in September 2017.
43. Eilers, P. H. C.; *Anal. Chem.* **2003**, *75*, 3631.
44. Savorani, F.; Tomasi, G.; Engelsen, S. B.; *J. Magn. Reson.* **2010**, *202*, 190.
45. Tomasi, G.; Savorani, F.; Engelsen, S. B.; *J. Chromatogr. A* **2011**, *1218*, 7832.
46. Kennard, R. W.; Stone, L. A.; *Technometrics* **1969**, *11*, 137.
47. Wu, W.; Mallet, Y.; Walczak, B.; Penninckx, W.; Massart, D. L.; Heuerding, S.; Erni, F.; *Anal. Chim. Acta* **1996**, *329*, 257.
48. Næs, T.; Isaksson, T.; Fearn, T.; Davies, T.; *A User-Friendly Guide to Multivariate Calibration and Classification*; NIR Publications: Chichester, UK, 2002.
49. de Carvalho, L.; de Morais, C. L. M.; de Lima, K. M. G.; Cunha Junior, L. C.; Nascimento, P. A. M.; de Faria, J.; Teixeira, G. A.; *Anal. Methods* **2016**, *8*, 5658.
50. Theophilou, G.; Lima, K. M. G.; Martin-Hirsch, P. L.; Stringfellow, H. F.; Martin, F. L.; *Analyst* **2016**, *141*, 585.
51. Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M.-A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; MacInnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L.; *Nucleic Acids Res.* **2007**, *35*, D521.
52. Okita, M.; Gaudette, D. C.; Mills, G. B.; Holub, B. J.; *Int. J. Cancer* **1997**, *71*, 31.
53. Theophilou, G.; Lima, K. M. G.; Briggs, M.; Martin-Hirsch, P. L.; Stringfellow, H. F.; Martin, F. L.; *Sci. Rep.* **2015**, *5*, 13465.
54. Siqueira, L. F. S.; Lima, K. M. G.; *Trends Anal. Chem.* **2016**, *82*, 208.
55. Lorentzen, T.; Nerstrom, H.; Iversen, P.; *Prostate Suppl.* **1992**, *4*, 11.
56. Oesterling, J. E.; Rice, D. C.; Glenski, W. J.; Bergstralh, E. J.; *Urology* **1993**, *42*, 276.

Submitted: March 23, 2017
Published online: September 5, 2017