8

Statistical Analysis for Automatic Identification of Ovarian Cancer Protein-Biomarkers Based on Fast Fourier Transform Infrared Spectroscopy

Marcano A.[1,2,3], D. Pokrajac[1,4], A. Lazarevic[1],
M. Smith[3], Y. Markushin[1,2] and N. Melikechi[1,2,3]
[1]*Center for Research and Education in Optical Sciences and Applications,*
[2]*Center for Applied Optics for Space Science,*
[3]*Department of Physics and Pre-Engineering*
[4]*Department of Computer and Information Sciences,*
Delaware State University, 1200 North Dupont Highway, Dover, DE 19901
United States of America

1. Introduction

Fast Fourier transform infrared (FTIR) spectroscopy has been used widely for the study of vibrations of protein molecules (Arrondo et al. 1993; Goormaghtigh et al. 1994; Haris & Chapman, 1994; Siebert, 1995; Barth & Zscherp, 2002; Petibois et al., 2006). Valuable information can be obtained of the secondary structure of the protein since peak positions and their relative amplitude are affected by the number of hydrogen bridges that sustain this secondary structure (Byler & Sussi, 1989; Fabian et al., 2001; Fabian et al., 2002). However, the spectral lines of proteins are usually broadened due to different molecular interactions thus making the identification of the structure difficult. Furthermore, identification of a particular protein within a complex matrix like a blood or a serum sample based on FTIR spectra is particularly challenging. Namely, direct application of automatic classification techniques is not a simple task, due to large numbers of attributes (measurements at different wavenumbers). Recently, principal component analysis (PCA) has been used as a statistical method for the feature extraction in the analysis of spectroscopic data aimed at detection of several complex organic samples (Hybl et al., 2003; Melikechi et al.; 2008, Lazarevic et al., 2009). In these methods, the spectroscopic data can be represented in a three-dimensional (or arbitrary dimension) space of eigenvector projections of the matrices corresponding to a series of experimental data measured for different selected wavelengths (Massart et al., 2003). In this regard, each point of this space represents a full set of spectroscopic measurements corresponding to one sample. Differences between the spectra can be then visualized graphically as different points in the space of eigenvectors. Linear discriminant analysis (LDA) or support vector machines (SVM), an advanced machine learning technique, can be subsequently used for automatic observing these differences between spectra. LDA generates linear models that separate classes based on the assumption that class-wise distributions are multivariate Gaussian with the same

covariance matrix (independent of the class label). SVM are classification algorithms that automatically assign a class label to a vector of data with theoretically best generalization (ability to predict the class outside the training data), independently of the data distribution. SVM generate a hyperplane in the transformed feature space (a non-linear transformation applied to the original data) such that the separation plane is as far from the data closest to it as possible. By using non-linear transformation, the likelihood that the training data can be separated by a hyperplane increases. By maximizing the distance between the data and the hyperplane, we achieve the smallest complexity of the classifier and hence, according to computational learning theory, maximize the generalization capability of the classification model. In this study, we propose to use the output of PCA analysis as input of LDA and SVM and to perform an automatic identification of protein molecules based on their FTIR spectra.

We use the proposed methodology to distinguish among the fast Fourier transform infrared (FTIR) spectra of proteins reported as possible biomarkers of ovarian cancer: monoclonal antibodies (MAB) and antigens (AG) of ovarian cancer marker CA125, Osteopontin (OPN), Leptin and insulin-like growth factor II (IGF2) (Mor et al., 2005; Schorge et al., 2004; Sutphen et al., 2004). We also complete a similar study on the common protein Bovine Serum Albumin (BSA) and human plasma samples for comparison purposes. We show that despite the presence of broadening mechanisms and evident similarities in the FTIR spectra of these proteins, the proposed method provides an automatic and effective identification of the proteins with almost perfect accuracy. This statistical procedure can also be applied to other spectroscopic methods such as fluorescence, NIR-VI absorbance spectroscopy and laser-induced breakdown spectroscopy.

As an important application we also perform deuteration of proteins and study the differences in the FTIR spectra introduced by this process using the PCA and LDA methods. FTIR spectra of deuterated versions of the proteins have been used extensively for the study of the secondary structure (Baenziger & Methot, 1995; Dave et al., 2002; Nie et al., 2005). Deuteration occurs by simple dilution of proteins in heavy water that contains the deuterium isotope of hydrogen (^2H). We have studied in details the changes induced by deuteration in the FTIR spectra of BSA and ovarian cancer biomarkers referred above. We have also explored the use of temperature and ultrasound to increase the changes. We use PCA and LDA methods to differentiate undeuterated and deuterated versions of the same protein. We propose that these methods can be used for identification of proteins within a matrix containing a large variety of proteins like a blood or serum sample. Furthermore, we propose a FTIR based immunoassay that uses the developed data analysis method and deuterated versions of the corresponding monoclonal antibodies for detection of protein biomarkers contained in a complex matrix like blood, plasma or serum samples.

2. Experimental method

For measuring the FTIR spectra we use an attenuated total reflection (ATR) FTIR spectrophotometer NICOLET 6700 (Thermo Industries, Inc). Drops of the samples are deposited over an aperture on the top of the device. This aperture connects to the surface of a diamond prism where the total reflection occurs. Samples under study are distilled and deionized water, heavy water (99.8% purity Deuterium oxide from Alfa Easer) and high

purity proteins (Sigma): BSA, 15 mM saline solution of MAB to AG CA125, AG CA125, Leptin, OPN and IGF2. Usually water masks most of the contribution from the proteins. To eliminate water peaks the samples are dried through simple evaporation of the solvent before collecting data. A drop of 5 μL of the solution is deposited over the aperture of the spectrophotometer. The samples are then left to dry at room temperature during 30 minutes. The drying process is monitored by taking spectra every 5 minutes until solvent (water of heavy water) contribution is depleted. When the drying process is complete the spectra do not show further changes. The dried protein sample forms a film over the aperture of several tens of micrometers good enough for total reflection spectroscopy. The spectra are collected with a resolution of 4 cm^{-1}. One hundred scans are averaged for each spectrum. The spectra show high reproducibility and a signal to noise value usually larger than 100. To collect data for the data analysis we repeat the spectroscopy experiment 40 times for each specimen. The deuteration of the proteins is performed using the dilution method at different concentrations and different dilution times. For solid samples like BSA we prepare directly a heavy water solution of the protein. For the solution samples we mix equal volumes of D_2O and the original protein solution. Deuteration can be improved by adding additional drops to the previously dried sample. Deuteration can also be improved by changing the temperature or using ultrasound. For this purpose we use an ultrasound cleaner with temperature control (Fisher Scientific FS20). The temperature is monitored with an independent thermocouple.

3. Classification methodology

We propose a statistical framework for automatic classification of the FTIR spectra of different proteins. The framework is illustrated in figure 1.

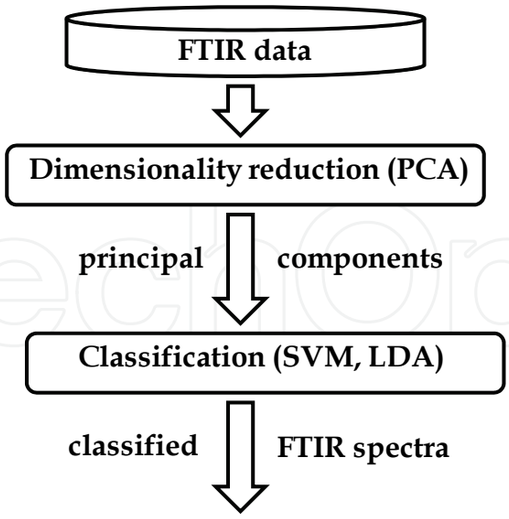

Fig. 1. Statistical framework for automatic FTIR spectra classification

The first step in the framework corresponds to dimensionality reduction, in which we reduce the number of frequencies in FTIR spectra using PCA. Principal components obtained through PCA are them used as an input to the classification module, which provides final classification of particular FTIR spectra.

3.1 Principal component analysis (PCA)

PCA is a powerful technique for dimensionality reduction in machine learning and data mining (Jolliffe, 2002). The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables (with non-diagonal covariance matrix), while retaining as much as possible of the variation present in the data set. This reduction is achieved by transforming to a new set of variables, which are called the principal components (PC). The PC are uncorrelated, and ordered in such a way that the first few retain most of the variation present in all of the original variables.

Suppose that **X** is a N-dimensional matrix of k-dimensional random variables $[x_1; x_2; x_N]$, and that the variances of the k random variables and the structure of the covariances or correlations between the k variables are of interest. Assume that we intend to approximate the vector x_i as a linear combination of m<k predetermined variables. In other words, assume that we would like to determine $\hat{\mathbf{x}}_i = \sum_{j=1}^{m} a_{ij}\mathbf{v}_j$, such that the mean square error $E\left(\left|\hat{\mathbf{x}}_i - \mathbf{x}_i\right|^2\right)$ is minimized. It can be proven that the mean square error is minimized when $\mathbf{v}_i$, i = 1,…,m are eigenvectors corresponding to m largest eigenvalues of the covariance matrix of **X**, and when a_{ij} are principal values projections of vector $\mathbf{x}_i$ with respect to its mean and first m eigenvectors. The vector $\hat{\mathbf{x}}_i$ contains m variables and thus it is typically stated that m "most significant" features are extracted out of k original coordinates.

The covariance matrix C of **x** can be estimated as:

$$C = \frac{1}{N-1} X^T X \ , \tag{1}$$

Its eigenvectors (column vectors) and corresponding eigenvalues satisfy the following condition:

$$C\,\mathbf{v}_j = \lambda_j \mathbf{v}_j \quad , \tag{2}$$

We can formally define eigenvector matrix and the diagonal matrix of eigenvalues respectively as:

$$\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2 \ldots\ \mathbf{v}_k] \ , \tag{3}$$

$$\Lambda = \mathrm{diag}\left(\lambda_1, ..., \lambda_k\right) . \tag{4}$$

Therefore we can compute the coefficients a_{ij} as $a_{ij} = \mathbf{x}_i\ \mathbf{v}_j$. Computation of PCA for high-dimensional data by definition may be very cumbersome, since it has $O(k^3)$ complexity, where k is the number of dimensions. The reason for such computational complexity is the requirement to compute eigenvalues and eigenvectors of a k*k matrix $C = \frac{1}{N-1} X^T X$. To make computation feasible, we will follow recently proposed approach (Bishop, 2006) that can extract up to N-1 principal components with the largest eigenvalues, when N < k.

Define $C^* = \frac{1}{N-1} X X^T$, and let U and Λ^* be respectively matrices of eigenvectors and eigenvalues of C^* such that:

$$C^* U = U \Lambda^* \ , \tag{5}$$

If we assume $V = X^T U$ and consider the estimated covariance matrix C, we can easily obtain that:

$$CV = \frac{1}{N-1} X^T X X^T U = X^T C^* U = X^T U \Lambda^* = V \Lambda^* \quad , \tag{6}$$

In other words, V is the eigenvector matrix of C and Λ^* is corresponding diagonal matrix of eigenvalues. Since C^* is of size N*N, using this technique would allow huge computational savings when N<<k. Note that vectors in V are not necessary normalized (to have a unit norm). Hence, to achieve orthonormal eigenvectors, an additional normalization step is required.

3.2 Linear Discriminant Analysis (LDA)

LDA (Krzanowski, 1988; Seber, 1984) is a statistical technique that classifies objects by computing the logarithm of the likelihood function (likelihood is the probability of the class given the observed data). Here, data from each class is assumed to belong to a multivariate Gaussian distribution. The Gaussian distributions corresponding to different classes are assumed to have the different means but the same covariance matrix, leading to the linear discrimination.

Formally, given the estimates of the prior probabilities p_j, and means $\boldsymbol{\mu}_j$ for each class j, and the estimate of the covariance matrix *C*, the logarithmic likelihood for a sample specified by a vector $\mathbf{y}_i$ can be computed as

$$l(j) = -\frac{1}{2}\ln|C| + \ln p_j - \frac{1}{2}\left(y_i - \boldsymbol{\mu}_j\right)^T C^{-1}\left(y_i - \boldsymbol{\mu}_j\right) \quad , \qquad j = 1,...,c \ , \tag{7}$$

where c is the total number of classes.

Using eq. (7), the classification of an example from a test set, specified by vector $\mathbf{y}_{new}$, is performed according to:

$$c_{new} = \arg\max_j l(j) = \arg\max_j \left(\ln p_j - \frac{1}{2}\boldsymbol{\mu}_j C^{-1} \boldsymbol{\mu}_j^T + \boldsymbol{\mu}_j C^{-1} \mathbf{y}_{new}^T \right) . \tag{8}$$

Hence, the separation plane between classes i and j can be described as a hyperplane:

$$f_{ij}(\mathbf{y}) = \left(\boldsymbol{\mu}_i C^{-1} - \boldsymbol{\mu}_j C^{-1}\right) y^T + \left(\ln p_i - \ln p_j - \frac{1}{2}\boldsymbol{\mu}_i C^{-1} \boldsymbol{\mu}_i^T + \frac{1}{2}\boldsymbol{\mu}_j C^{-1} \boldsymbol{\mu}_j^T \right) = 0 . \tag{9}$$

For each class, we can define a decision margin as a minimal distance between a sample from a class and the separation planes.

Let $\mathbf{y}_{i,j}$, i =1,…,n_j be row feature vectors from the training set belonging to class j and let n_j be the number of vectors in class j. We estimate the class priors, means, and the covariance matrix as:

$$p_j = \frac{n_j}{\sum_{j'=1,...,c} n_{j'}}$$

$$\boldsymbol{\mu}_j = \frac{1}{n_j}\sum_{i=1,\ldots,n_j} \mathbf{y}_{i,j} \tag{10}$$

$$C = \frac{1}{\sum_{j'=1,\ldots,c} n_{j'} - 1}\sum_{j'=1,\ldots,c} \frac{C_{j'}}{n_{j'} - 1}$$

where:

$$C_j = \frac{1}{n_j - 1}\sum_{i=1,\ldots,n_j} \left(\mathbf{y}_{i,j} - \boldsymbol{\mu}_j\right)^T \left(\mathbf{y}_{i,j} - \boldsymbol{\mu}_j\right) . \tag{11}$$

3.3 Support Vector Machines (SVM)

LDA provides optimal classification with linear decision boundaries if its assumption of class-specific Gaussian distributions with identical covariances is satisfied. However, if this assumption is not satisfied, the optimal decision boundaries could be obtained by using SVM (Vapnik, 2000). The main idea of SVM is to construct a separation hyperplane, which optimally separates data examples belonging to two classes, such that the minimal distance between points and the separation hyperplane is maximized. Such constructed hyperplane provides the best generalization of unknown examples. SVM use structural risk minimization principle and aim to achieve zero training error while minimizing the complexity of the model. However, if linear separation is not possible, SVM work towards minimization of the number of misclassified examples on the training set by introducing the slack variables and regularization. Formally, SVM learning can be represented as the following quadratic programming problem (Bishop, 2006):

$$\begin{aligned} &\min_{\mathbf{w},\boldsymbol{\xi}_i,d_0} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\boldsymbol{\xi}_i\right) \text{s.t.} \\ &\left(\mathbf{w}^T\mathbf{y}_i + d_0\right)\cdot c_i \geq 1 - \boldsymbol{\xi}_i, i = 1,\ldots,N \\ &\boldsymbol{\xi}_i \geq 0, i = 1,\ldots,N. \end{aligned} \tag{12}$$

where, $\mathbf{w}$ is vector defining the separation hyperplane, d_0 is the intercept of the separation hyperplane, $c_i \in \{-1,1\}$ is a class label of the i^{th} example determined by attribute vector $\mathbf{x}_i$, ξ_i is the slack variable corresponding to the i^{th} example, C is a preset regularization constant, and $\mathbf{y}_i = \mathbf{f}(\mathbf{x}_i)$ is a vector representing a non-linear function of $\mathbf{x}_i$. In general, $\mathbf{x}_i$ and $\mathbf{w}$ can be infinitely dimensional.

Using the Karush-Kuhn-Tucker (KKT) theorem (Karush, 1939), SVM learning can be represented as the optimization in dual space of Lagrangian multipliers λ_i. In this case, the learning phase reduces to the following optimization problem:

$$\begin{aligned} &\max_{\lambda}\left(\sum_{i=1}^{N}\lambda_i - \sum_{l=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j c_i c_j \mathbf{y}_i{}^T\mathbf{y}_j\right) \text{ s.t.} \\ &\sum_{i=1}^{N}\lambda_i c_i = 0 \\ &\lambda_i \geq 0, i = 1,\ldots,N \\ &\lambda_i \leq C, i = 1,\ldots,N. \end{aligned} \tag{13}$$

An example $\mathbf{x}_{new}$ from the test set is subsequently classified according to the following equation:

$$c_{new} = \text{sign}\left(\mathbf{w}^T \mathbf{y}_{new} + d_0\right) \quad , \tag{14}$$

where $\mathbf{y}_{new} = f(x_{new})$, which can be expressed using the Lagrangian multipliers as:

$$c_{new} = \text{sign}\left(\sum_{i:\lambda_i>0} \lambda_i c_i \mathbf{y}_i{}^T \mathbf{y}_{new} + \frac{1}{N_s}\left(\frac{1}{c_i} - \sum_{j:\lambda_j>0} \lambda_j c_{ji}{}^T \mathbf{y}_i{}^T \mathbf{y}_j\right)\right), \tag{15}$$

where N_s denotes the number of support vectors—points closest to the separation hyperplane (i.e., number of non-zero Lagrangian multipliers).
Similar as in the case of LDA, with SVM we can explicitly calculate the separation planes in the transformed space specified by:

$$w^T y + d_o = 0 \quad , \tag{16}$$

where

$$w = \sum_{i:\lambda_i>0} \lambda_i y_i c_i \quad , \tag{17}$$

$$d_0 = \frac{1}{N_s} \sum_{i:\lambda_i>0} \left(\frac{1}{c_i} - w^T \mathbf{y}_i\right).$$

If we select features in the transformed space to be proportional to eigenvalues and eigenfunctions of a symmetric non-negative definite kernel, then, due to the Mercer's theorem, we can write $\mathbf{y}_i^T \mathbf{y}_j = K\left(\mathbf{x}_i, \mathbf{x}_j\right)$ where K is symmetric non-negative-definite function of two vectors (Bishop, 2006). Then, due to the Mercer's spectral theorem for non-negative definite symmetric kernels, SVM learning and classification can be stated directly using original (non-transformed) feature vectors as:

$$\begin{aligned} &\max_{\lambda}\left(\sum_{i=1}^{N}\lambda_i - \sum_{l=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j c_i c_j K\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) \text{ s.t} \\ &\sum_{i=1}^{N}\lambda_i c_i = 0 \\ &\lambda_i \geq 0, i = 1,...,N \\ &\lambda_i \leq C, i = 1,...,N. \end{aligned} \tag{18}$$

$$c_{new} = \text{sign}\left(\sum_{i:\lambda_i>0} \lambda_i c_i K\left(\mathbf{x}_i, \mathbf{x}_{new}\right) + \frac{1}{N_s}\left(\frac{1}{c_i} - \sum_{j:\lambda_j>0} \lambda_j c_{ji}{}^T K\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)\right), \tag{19}$$

This makes possible using implicit and infinitely dimensional transformation **f**. Popular choices of kernel function include:

- Linear kernel: $K\left(\mathbf{u}, \mathbf{v}\right) = \mathbf{u}^T \mathbf{v}$;

- Polynomial kernel (p is a prespecified parameter): $\mathrm{K}(\mathbf{u},\mathbf{v}) = \left(1+\mathbf{u}^{\mathrm{T}}\mathbf{v}\right)^{\mathrm{p}}$;
- Exponential kernel (σ is a pre-specified parameter): $\mathrm{K}(\mathbf{u},\mathbf{v}) = \mathrm{e}^{-\frac{1}{2\sigma^2}\|\mathbf{u}-\mathbf{v}\|^2}$.

The original SVM technique is designed for a two-class problem. For a multiclass problem (i.e., c>2) we use Directed Acyclic Graph SVM (DAG-SVM) method (Platt et al., 2000) which for a c-class problem trains c(c-1)/2 two-class support machines and the class decision is performed based on successive elimination of classes as a result of a two-class comparison. In comparison to one-to-rest classifiers, the application of DAG-SVM is more practical, since it does not result in imbalanced training sets (Hsu & Lin, 2002; Jiang et al. 2005).

3.4 Classification accuracy evaluation

To validate the accuracy of the classification model on the data unseen during the learning process, we use a four-fold cross validation, that can be described as follows (Bishop, 2006): 1) split randomly the dataset into four subsets; 2) set aside one of the subsets as the test set while the other three subsets are chosen to form the training set; 3) Utilize the training set to learn the classification model and employ the test set to evaluate the accuracy of classification on data unseen during the learning process; 3) repeat the process four times so that each of the four subsets has a chance to be a test set; 4) use averaged results from the four classification experiments as an overall measure of the model performance.

As a measure of performance, we utilize overall classification accuracy, the ratio of correctly classified samples for all classes versus the number of all classified samples in the test set (Bramer, 2007), defined as:

$$\text{Overall Accuracy} = \frac{\text{correctly classified samples from all classes}}{\text{total number of samples}} \times 100[\%]. \qquad (20)$$

4. Classification of proteins using PCA and SVM analysis of their FTIR spectra

Figure 2 depicts the FTIR spectra from dried MAB to CA125, BSA, human plasma, MAB to ILGF2, MAB to Leptin and MAB to OPN. All spectra exhibit a similar structure. The origin of the peaks has been well documented in the literature (reviewed by Barth & Zscherp, 2002). The spectra have several distinctive regions. The first region corresponds to the interval 2800-3500 cm^{-1}. A NH_2 region around 3200 cm^{-1} is strongly overlapped with OH stretching band. The region 1800-2700 cm^{-1} is relatively free of peaks. Amide bands are characteristics in the region 1200-1700 cm^{-1}. Those arise from the amide bonds that link the amino acids. The amide I centered about 1740 cm^{-1} corresponds to the stretching mode of the C=O bond of the amide. It may have some contribution from CN stretching and CCN deformations. The amide peak II centered around 1550 cm^{-1} corresponds to the bending mode of the NH bond of the amide with contributions from C=O in plane bending and NC stretching. Amide III mode is the in-phase combination of NH in-plane bending and CN stretching. Other smaller peaks corresponding to CC stretching and CO bending are observed in this region. The characteristics of these peaks provide information about the

secondary structure of the proteins since the hydrogen bonds that establish this structure, are mostly associated to the CO and NH bonds. The wide peak in the region 400-800 cm^{-1} corresponds to librations with contribution from other rotational and low energy vibrational lines. In figure 3 we show the results of the use of the first two PCA variables to represent the data presented in figure 2. Despite the evident similarities between the spectra the data are perfectly separable even with the use of only the first two PCA variables. For MAB to Leptin, ILGF2 and OPN the separation is larger.

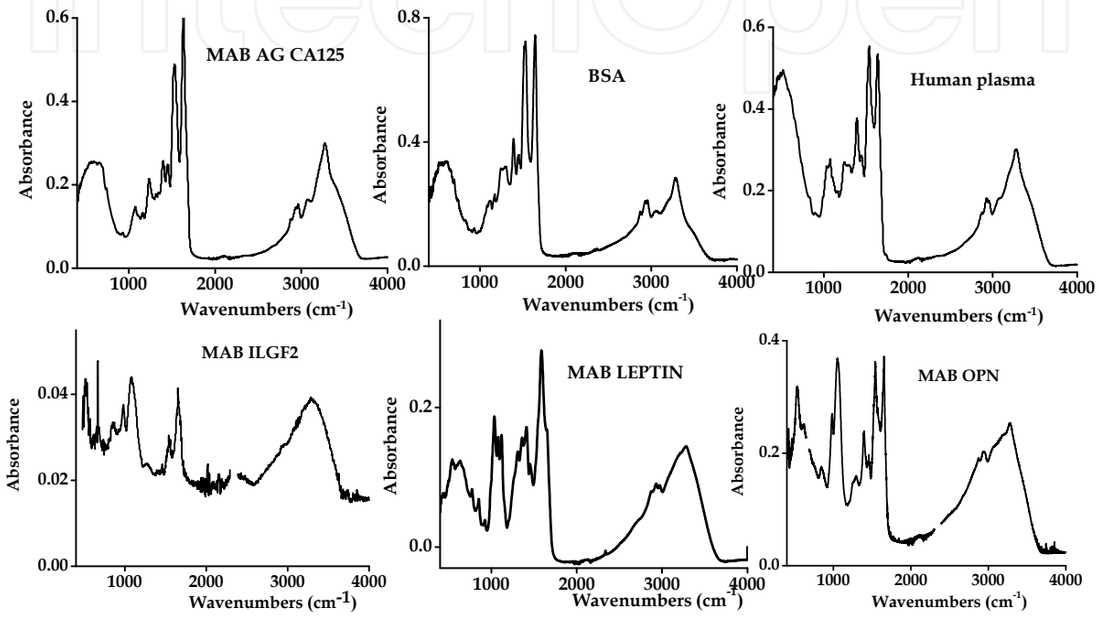

Fig. 2. FTIR spectra of MAB to AG CA125, BSA, human plasma, MAB to ILGF2, MAB to Leptin and MAB OPN

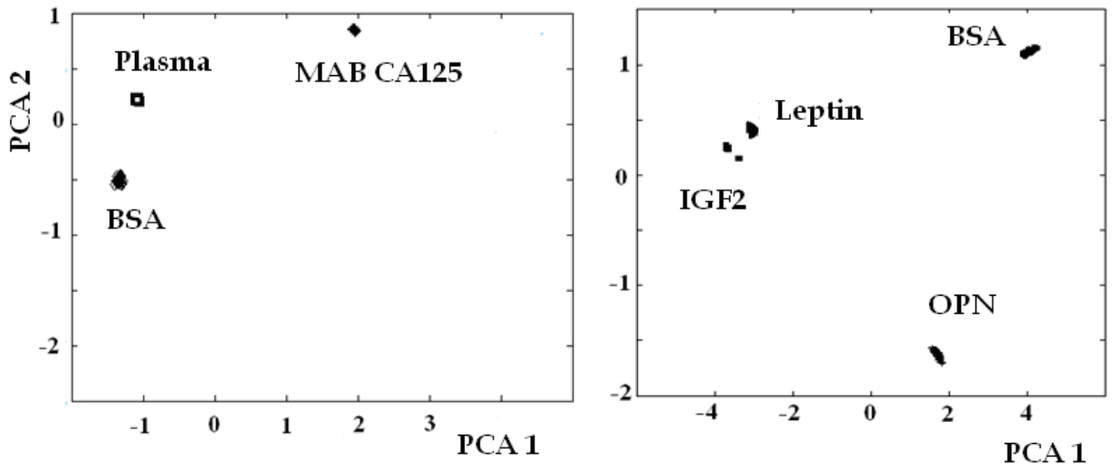

Fig. 3. Two-dimensional PCA of the data presented in figure 2

From figure 4 we can see that even the first two principal components are sufficient to achieve perfect separation of protein classes. Therefore, even the application of linear SVM is able to provide perfect accuracy on both training and test sets (100% accuracy). The result does not depend on the number k of principal components used (k>1). The average number of support vectors per class is relatively small (Figure 4) and practically does not depend on the number of principal components used, which indicates good generalization and stability of the proposed technique. The results demonstrate the possibility of automatic classification of proteins using PCA and linear SVM with accuracy of nearly 100%. Hence, this justifies the application of the conceptually simpler LDA technique. Namely, LDA is also capable of achieving 100% accuracy using as little as 2 principal components. Hence, below we discuss the use of LDA to separate the FTIR spectra of proteins and their deuterated versions aimed at the development of a FTIR based immunoassay.

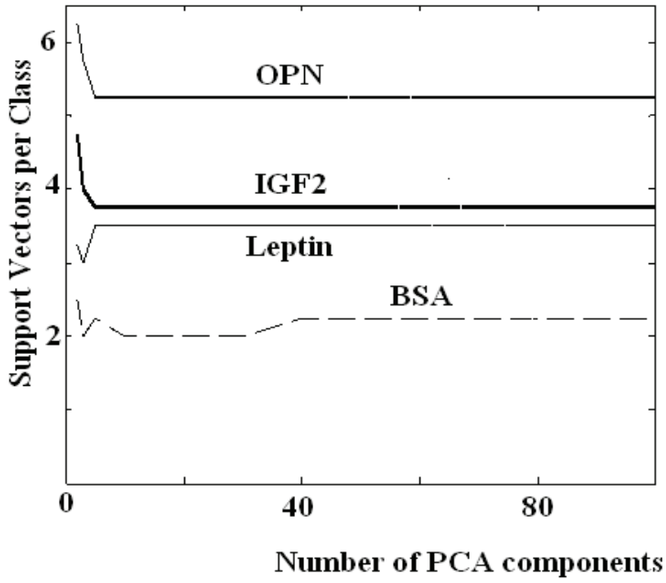

Fig. 4. Support vector per class as a function of the number of PCA components after Lazarevic et al. (2009). Reproduction authorized by the International Society for Optical Engineering SPIE

5. FTIR of deuterated proteins

Deuterium is a stable isotope and can be used as a labeling agent. Deuteration occurs by simple dilution of proteins in heavy water that contains the deuterium isotope of hydrogen (^2H). Hydrogen atoms on the surface of the protein are exposed to a fast exchange with deuterium atoms while hydrogen atoms deeply buried within the protein molecule exchange at a low pace. As an effect by substituting hydrogen atoms by deuterium atoms vibration modes of OH (hydroxyl peaks), NH_2 (amide peaks), NH and/or CH can be affected. Deuteration also induces the appearance of a strong peak in the region around 2400 cm^{-1}. This region is usually free of peaks for most of the proteins. Besides its evident advantages and extensive use for the study of the secondary structure deuteration of proteins can have another important application still not considered in the

literature. Indeed, the FTIR spectrum of a deuterated protein is different from the non-deuterated one, and this can be used for their identification within a matrix containing a large variety of proteins, e.g., a blood, plasma or serum sample. Furthermore, as we demonstrate, the use of PCA and LDA can identify these differences automatically with high accuracy.

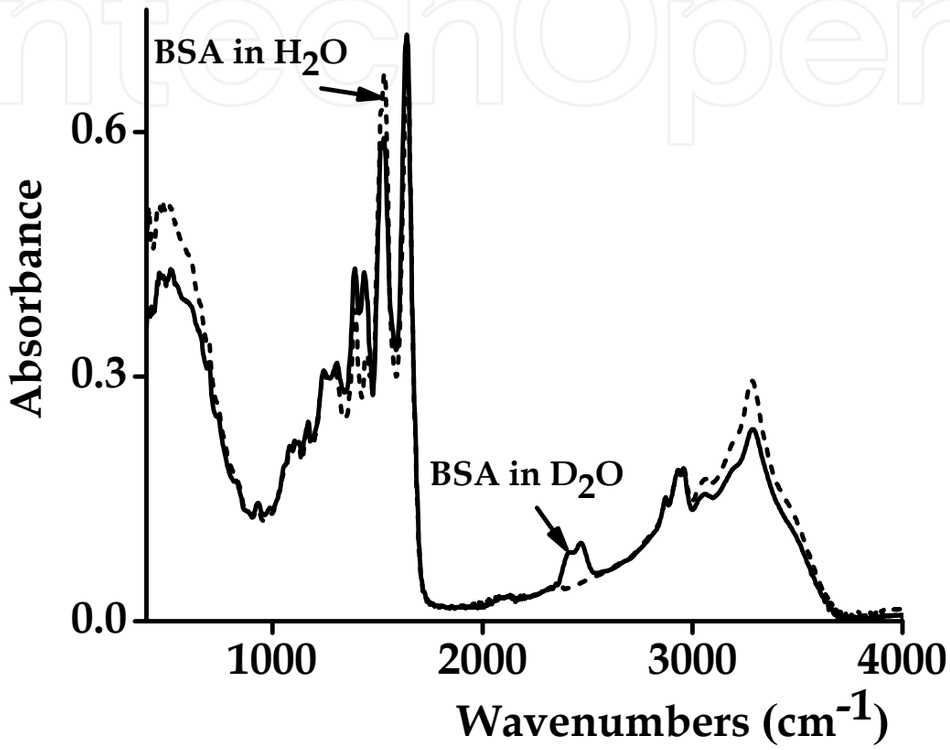

Fig. 5. FTIR spectra of non-deuterated BSA and partially deuterated BSA

As it is well known, the spectrum of water exhibits the stretching symmetric and antisymmetric modes at 3253 cm^{-1} and 3315 cm^{-1}, respectively, the bending mode at 1647 cm^{-1} and libration peaks observed below 700 cm^{-1}. The spectrum also exhibits a small peak at 2094 cm^{-1} that corresponds to the interaction between bending and libration modes. The heavy water spectrum exhibits a similar structure of peaks but all the peaks are shifted by a factor of 1.37 in close correspondence to the factor of 1.41 calculated from the differences of masses between hydrogen and deuterium atoms. The spectral widths of the D_2O lines are also reduced by a factor of 1.35 comparing to those of water. Of special interest are the peaks of the stretching vibration of deuterium oxide molecule which are centered at 2401cm^{-1} and 2471 cm^{-1}. Proteins FTIR spectra are usually free of peaks in this area (see figure 2). Deuterated proteins can have peaks in this region, a feature that can be used for protein identification and calibration. As a consequence, shifts of spectral lines, changes in relative amplitudes and changes in the spectral widths are expected for deuterated proteins (Marcano et al., 2008).

In figure 5 we show the FTIR spectra of the dried BSA sample from solutions of distilled double dionized water (dash line) and D_2O (solid line) prepared at concentration of 42 μg/mL. The amide I peak (1644 cm^{-1}) remains almost unaltered while the amplitude of the amide II peak (1530 cm^{-1}) decreases. This peak corresponds to NH bending vibrations which are strongly affected by substitution of hydrogen by deuterium atoms. Amplitude increase is observed for other peaks in the region 1200-1300 cm^{-1}. Remarkable is the presence of the peaks in the region 2400-2500 cm^{-1} which is free of peaks not only for BSA but also for a number of antibody proteins (see figure 2). These peaks correspond to stretching OD vibration. Correspondingly, the hydroxyl peaks in the region 3000-3500 cm^{-1} are reduced.

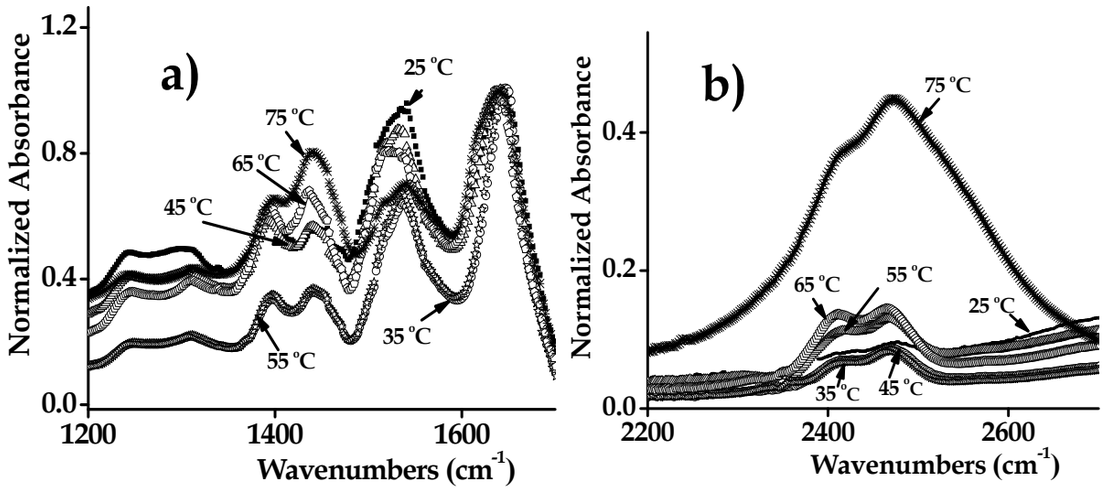

Fig. 6. Details of the FTIR spectra of BSA as a function of temperature in the regions 1200-1700 cm^{-1} and 2200-2600 cm^{-1}

Increase in temperature deepens the changes observed. In figure 6 we show details of the FTIR spectra of BSA in the regions 1200-1700 cm^{-1} (a) and 2200-2600 cm^{-1} (b), respectively, after heating the solution from 25 °C up to 75 °C. High temperature breaks the hydrogen bonds opening the protein molecule and exposing it to wide deuteration. The effect is small up to a certain temperature (60 °C in our case) but when the thermal energy is enough to break the hydrogen bonds the effect increases substantially. For the sample heated up to 75 °C the peaks at 2400 cm^{-1} are more than 5 times larger than the one for 55 °C (see figure 6b). The amide II peak is depleted as well as other peaks at 1200 cm^{-1}(see figure 6a). The depletion is also remarkable for the hydroxyl peaks at 3500 cm^{-1}. Deuteration can be significantly increased by the use of ultrasound. Ultrasound shakes the molecule exposing its hydrogen bonds to deuteration. In figure 7 we show the results of deuteration of BSA in D_2O at concentration of 500 mg/mL by using ultrasound during 120 minutes at room temperature. After exposing to ultrasound the solution is put to rest for long term dilution (1 week) at 6 °C. The changes in the absorbance FTIR are remarkable. The peak at 2466 cm^{-1} dominates the center of the spectrum. The amide II peaks is almost depleted and the peak at 1434 cm^{-1} triples its amplitude value.

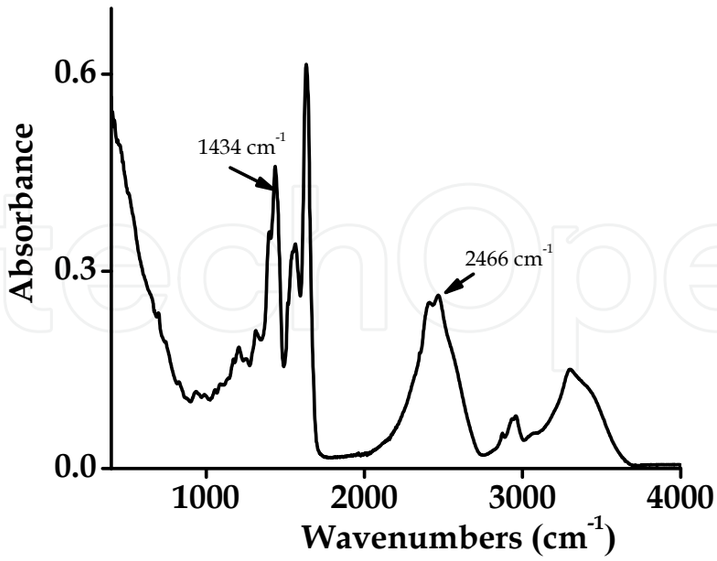

Fig. 7. FTIR spectrum of highly deuterated BSA obtained using ultrasound

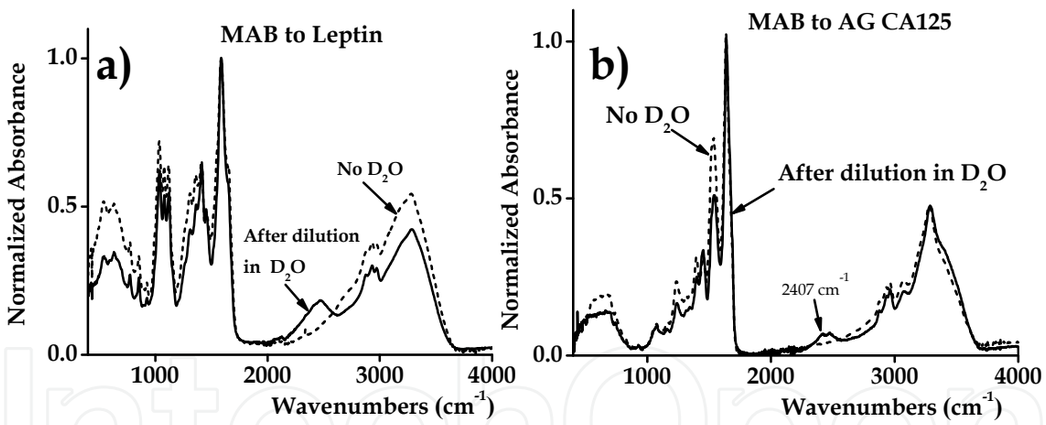

Fig. 8. FTIR spectra of deuterated MAB to Leptin and MAB to AG CA125

We have studied the effects of deuteration on MAB to Leptin, MAB to AG CA125, AG CA125, Leptin, OPN and IGF2. All these proteins are originally diluted in saline solution by the supplier of chemicals. For deuteration we use 5 μL of the sample and diluted it into 5 μL of heavy water. One drop of this diluted solution is then deposited to dry over the spectrometer. In figure 8 we show the FTIR spectra of MAB to Leptin (8a) and MAB to AG CA125 (8b) from the original saline solution (dot lines) and heavy water dilution (solid lines). The spectra are normalized with respect to the amplitude of the amide I peak. Again we observe the surge of the DO peak in the region around 2400 cm^{-1} and also changes in the relative amplitudes of hydroxyl and amide II peaks. Reduction in the spectral width of the peak is also observed.

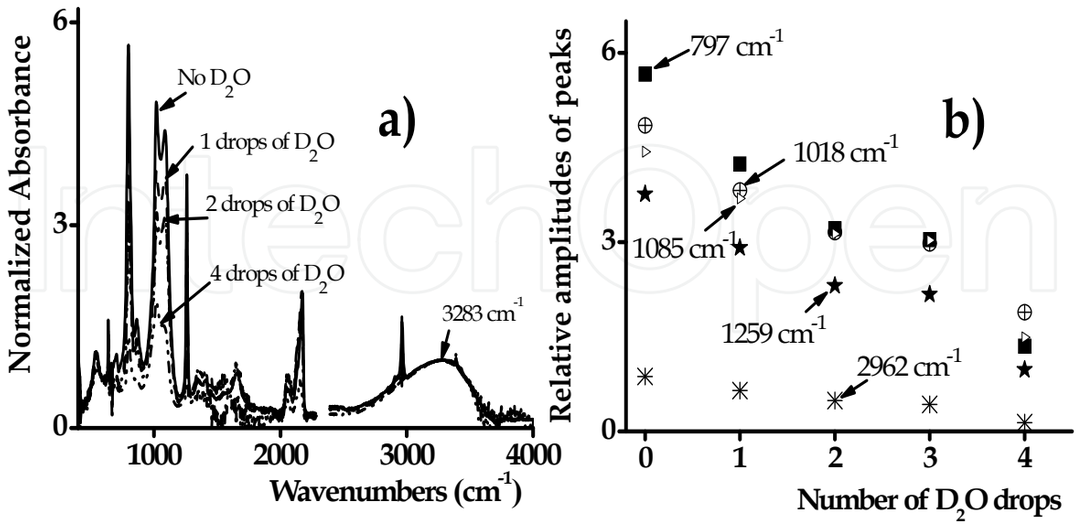

Fig. 9. FTIR spectra of AG CA125 exposed to different amount of D_2O

We have also studied the effect of deuteration of ovarian cancer AG CA125, OPN, IGF2 and leptin. In figure 9 we show the effect of deuteration over the ovarian cancer AG CA125 after adding drops of D_2O (figure 9a) subsequently. The sample is left to dry after addition of each drop and before recording the spectra. We observe decrease in the relative amplitude of several peaks all over the spectrum. In figure 9b we plot the relative amplitudes of five of these peaks as a function of the amount of D_2O used. Similar results are obtained for leptin, OPN and IGF2.

Figures 5-9 demonstrate that deuteration of proteins is relatively easy to achieve by simple dilution in heavy water for both the monoclonal antibody proteins and their corresponding antigens. If required the impact of deuteration can be increased by increasing the temperature or by applying ultrasound.

6. Use of PCA method for detection of deuterated proteins

As expected, using PCA to perform feature extraction can lead to distinguish with high efficiency between deuterated and undeuterated versions of the same protein. In figure 10 we plot the tridimensional principal projection of the FTIR data from AG Leptin exposed to 1, 2 and 3 drops of D_2O. A good separation between the data is obtained. The absolute distance between the data increases with the number of D_2O drops as correspond to larger deuteration effect as suggested by figure 10. However, the effect depends on the type of protein. In figure 11 we show the result for BSA. In this case, the distance between the unexposed sample PCA data and the exposed ones does not change monotonically with the numbers of D_2O drops. This may be related to parasitic D-H exchange with water of the surrounding the sample atmosphere over the time of the experiment. Nevertheless, a very clear separation between BSA samples with different numbers of D_2O drops is achieved with 100% classification accuracy using LDA with four-fold cross-validation.

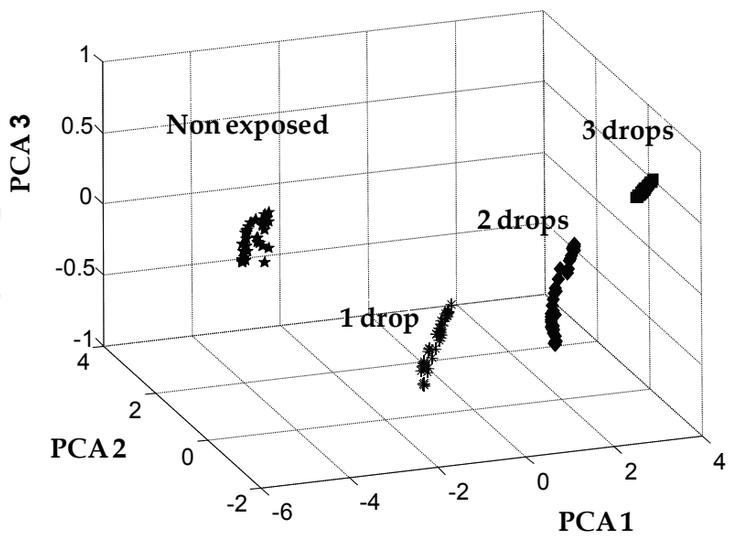

Fig. 10. Tridimensional PCA representation of the FTIR data from AG Leptin exposed to successive drops of D_2O. Non-exposed to D_2O data are included for comparison

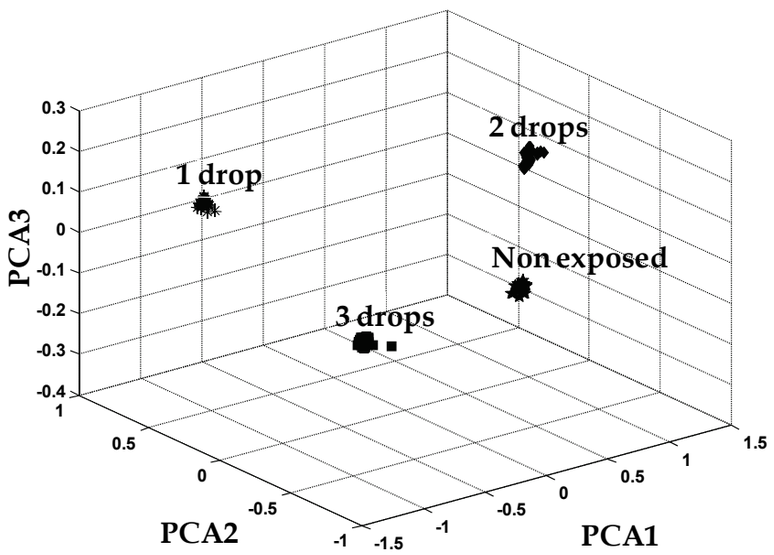

Fig. 11. Tridimensional PCA of FTIR spectra of BSA exposed to 1, 2 and 3 drops of D_2O. The non-exposed to D_2O samples is included for comparison

Different regions of the spectra contribute differently to the separation. In figure 12 we show the plot of the PCA loadings (absolute value of components for eigenvectors $\mathbf{v}_1$-$\mathbf{v}_3$ from Eq. 2 as functions of the wave-numbers. The larger the absolute value of the PCA loading is the larger is the importance of the corresponding spectral region for a more efficient separation.

In the figure we see the importance of the low wavenumbers region (400-600 cm^{-1}), the amide peaks region (1300-1600 cm^{-1}) and the DO peak region (2400-2500 cm^{-1}).

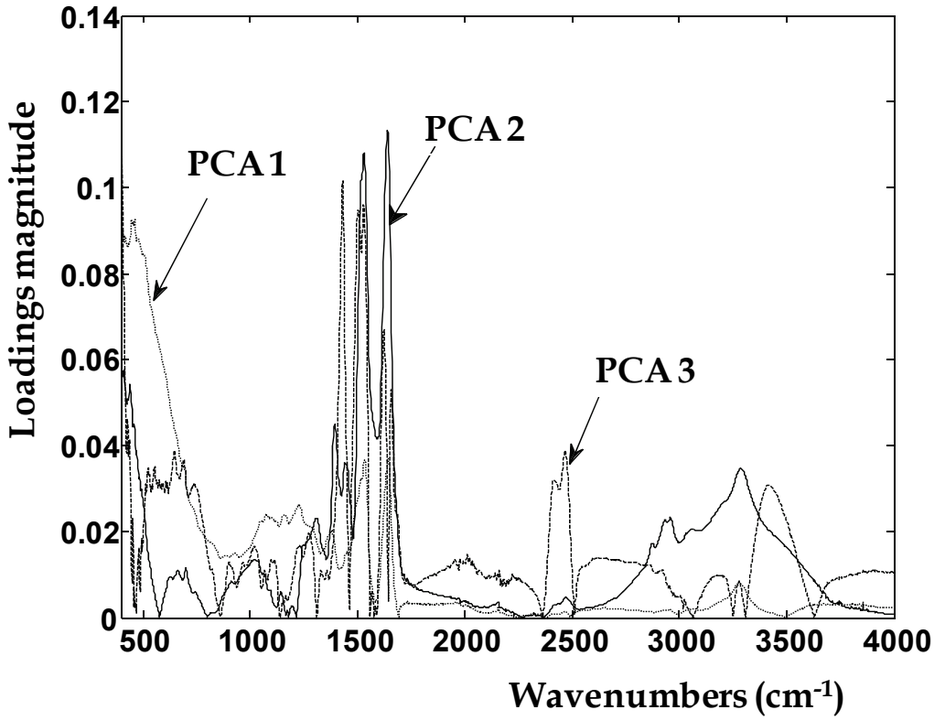

Fig. 12. PCA variables as functions of the wavenumbers corresponding to the data of figure 11

7. Detecting a protein antibody in a complex matrix

The proposed methods can be used to detect the presence of a particular protein in a complex matrix, such as blood, plasma, serum or other biomedical samples. The efficiency of separation can be increased by using deuterated versions of the proteins. In figure 13a we show the FTIR spectra of pure human plasma (stars), plasma with added non-deuterated (diamonds) and deuterated (squares) MAB to CA125 at a concentration of 50 μg/ml normalized by the amplitude of the peak at 3275 cm^{-1}. The differences are more remarkable for the deuterated samples comparing to the non deuterated ones. In figure 13b we show the tridimensional PCA plot corresponding to these data. Despite the similarities of spectra, the data are separable in the PCA coordinates space. The separation is larger for the deuterated version of the protein. The absolute distance between the centers of data cluster in the PCA coordinates space corresponding to plasma and the cluster corresponding to deuterated MAB to CA125 is more than twice larger than the distance of the center of the plasma cluster to the non-deuterated protein data cluster. Although the concentration used is relatively high, the result demonstrates possibilities of detection of proteins embedded in a complex matrix and the increase in sensitivity when using deuterated versions of the proteins.

The spectral changes in deuterated proteins and the statistical and data mining methods used for their analysis can be applied to develop new kinds of immunoassays for detection of antigen proteins. The immunoassays are aimed at detection of a particular antigen protein embedded in a complex matrix, such as blood, serum or a plasma sample. Figure 14 describes two such immunoassays. Figure 14a depicts an immunoassay where non-deuterated protein antibodies are deposited over a glass substrate (step 1). The plate is then exposed to the sample. The antibody proteins on the plate capture their corresponding antigens from the sample (step 2). Finally, the system is exposed to the presence of the deuterated versions of the antibody proteins. These deuterated antibodies are then attached to the trapped antigens forming a sandwich structure which is wash away to remove non captured proteins (step 3). This sandwiched structure can then be analyzed using an FTIR spectrophotometer. In a second type of immunoassay the first step is the same (see figure 14b). Then, the plate is exposed to the sample which has been previously diluted in heavy water. Deuterated antigens can then be captured by their antibody protein deposited on a plate. The rest of the sample can be washed away. Finally, the presence of the antigens can be detected by performing the FTIR experiment over the treated plate.

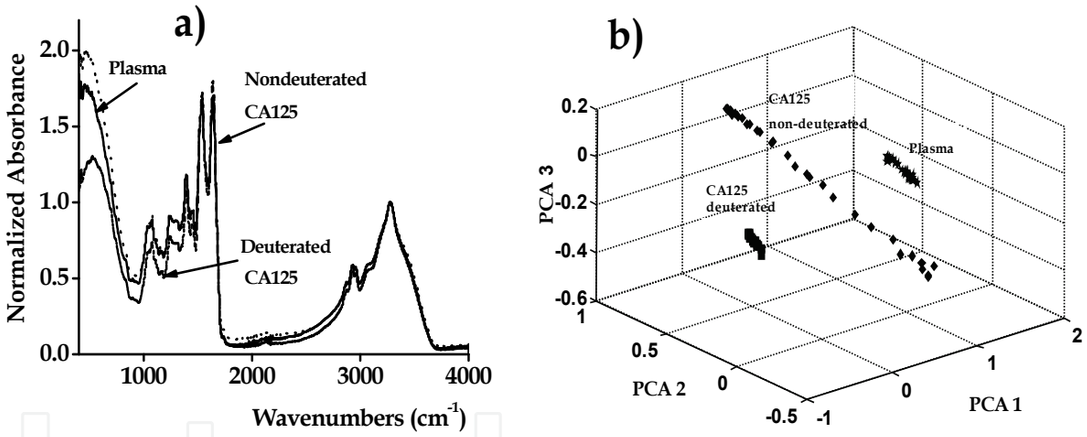

Fig. 13. a) Normalized FTIR spectra of plasma and plasma containing deuterated and non-deuterated MAG AG CA125, b)Tridimensional PCA plot of FTIR data from plasma and deuterated and non-deuterated versions of the protein MAB to CA125

Several steps need to be completed before developing a FTIR-based immunoassay of practical use. Deuteration can affect the bioactivity of the proteins. In this regard, the affinity constant between the antibody and antigen can depend on the level of deuteration. The level of deuteration can be also affected by parasitic D-H exchanges that can mask the real results. Practical comparison with well established immunoassays such as ELISA must be completed. However, we show that the use of deuteration techniques in combination with statistical methods, such as PCA, LDA and SVM, will play a crucial role in the developing of this new kind of FTIR-based immunoassays aimed at detection of a targeted protein in a complex biosample.

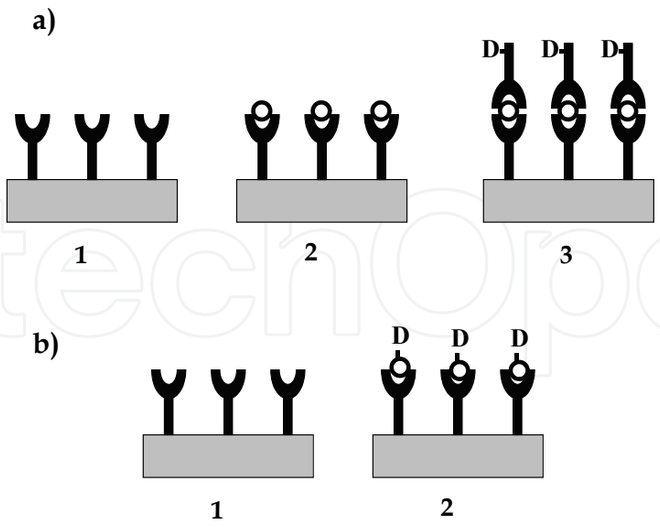

Fig. 14. Proposed FTIR based immunoassays for detection of a given antigen

8. Conclusions

In this study, we demonstrate that using combination of PCA analysis and statistical and data mining classification techniques (LDA and SVM) it is possible to automatically determine a class of the FTIR sample of an unknown protein using a small number of principal components. In such a case, using conceptually simpler LDA analysis (that relies on stronger assumptions about the data than SVM) is justified. The full advantage of non-linear SVM could, however, be expected in case of more complex and noisy spectroscopic data. The proposed data analysis technique is computationally fast and can in principle be applied in on-line learning classification framework. Work in progress includes testing the proposed technique on larger datasets to exclude the small variability of samples (the sample bias) as a potential reason for extremely high classification accuracy. We show that the techniques distinguish between different proteins with similar FTIR spectra and between deuterated and non-deuterated versions of the same protein. Furthermore, we demonstrate the use of the method for separation and identification of proteins embedded in a complex matrix of proteins such as plasma. We show that deuteration increases the sensitivity of the method. Finally, we propose an immunoassay that is aimed to utilize the demonstrated sensitivity of the methodology to detect a particular antigen protein in a complex biosample.

9. Acknowledgements

This research has been possible thanks to the support of the National Science Foundation (NSF-CREST grant Nº 0630388 and NSF-MRI grant Nº 0922587), National Institute of Health (NIH Grant Nº P20 RR016472), Department of Defense (DoD/DoA Grants Nº 45395-MA-ISP, and Nº 54412-CI-ISP) and of the National Aeronautics and Space Administration (NASA URC 5 grant Nº NNX09AU90A). We also would like to thank Blood Bank of Delmarva, Delaware, for providing the human plasma samples.

10. References

Arrondo, J. L. R.; Muga, A.; Castresana, J. & Goñi, F. M. (1993). Quantitative studies of the structure of proteins in solution by Fourier-transform infrared spectroscopy. *Prog. Biophys. Molec. Biol*., Vol. 59, No. 1, 23–56, ISSN 0079-6107.

Baenziger, J. E. & Methot, N. (1995). Fourier transform infrared and hydrogen/deuterium exchange reveal an exchange-resistant core of alpha-helical peptide hydrogens in the nicotinic acetylcholine receptor. *J. Biol. Chem*., Vol. 270, No. 49, 29129-29137, ISSN 0021-9258.

Barth, A. & Zscherp, C. (2002). What vibrations tell us about proteins. *Quarterly Reviews of Biophysics,* Vol. 35, No. 4, 369–430. Cambridge University Press, DOI: 10.1017/S0033583502003815. ISSN 0033-5835.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer + Business Media LLC, ISBN-10 0-38731073-8, Singapore.

Bramer, C. M. (2007), *Principles of Data Mining (Undergraduate Topics in Computer Science)*, Springer-Verlag, ISBN-10: 1-84628-765-0, London.

Byler, D. M. & Susi, H. (1989). Examination of the secondary structure of proteins by deconvolved FTIR spectra. *Biopolymers,* Vol. 25, 469-487, ISSN 1097-0282.

Dave, N.; Lórenz-Fonfrı□a, V. A.; Villaverde J.; Lemonnier R.; Leblanc G. & Padrós, E. (2002). Study of Amide-proton Exchange of *Escherichia coli* Melibiose Permease by Attenuated Total Reflection-Fourier Transform Infrared Spectroscopy: Evidence of structure modulation by substrate binding. *J. Biol. Chem.* Vol. 277, 3380-3387, ISSN 0021-9258.

Fabian H. & Shultz, C. P. (2001). *Encyclopedia of Analytical Chemistry,* R. A. Meyers (Ed.), Wiley, 5779-5803. ISBN: 978-0-471-97670-7, Chichester.

Fabian H. & Mantele W. (2002). *Handbook of Vibrational Spectroscopy*, J. M. Chalmers, P. R. Griffiths (Eds.) Wiley, 3399-3425, ISBN: 978-0-471-98847-2, Chichester

Goormaghtigh, E.; Cabiaux, V. & Ruysschaert, J. M. (1994). Determination of soluble and membrane protein structure by Fourier transform infrared spectroscopy I. Assignments and mode compounds. *Subcell. Biochem.* Vol. 23, 329–362, ISSN -0306-0225.

Haris, P. I. & Chapman, D. (1994). Analysis of polypeptide and protein structures using Fourier transform infrared spectroscopy. *Methods in Molecular Biology, Microscopy, Optical Spectroscopy, and Macroscopic Techniques*, vol. 22 (eds. C. Jones, B. Mulloy & A.H. Thomas), 183–202. Humana Press Inc. ISBN 0-89603-232-9, Totowa, NJ.

Hsu, C. W & Lin, C. J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Networks*, vol. 13, No. 2, 415-425, ISSN 1045-9227.

Hybl, J. D.; Lithgow, G. A. & Buckley, S. G. (2003). Laser-induced breakdown spectroscopy detection and classification of biological aerosols. Appl. *Spectros*. Vol. 57, 1207-1215.

Jiang, Z. Q.; Fu, H. G. & Li, L. J. (2005). Support vector machine for mechanical faults classification. *Journal of Zhejiang University Science* (Online) 6A(5). http://www.zju.edu.cn/jzus/2005/A0505/A050513.pdf.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Verlag, ISBN 0-387-95442-2, New York.

Karush, W. (1939). *Minima of Functions of Several Variables with Inequalities as Side Constraints*. M.Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois. Available from http://wwwlib.umi.com/dxweb/details?doc_no=7371591.

Krzanowski, W. J., 1988). *Principles of Multivariable Analysis: A User's Perspective*. Oxford University Press, ISBN 0 198 8507089, New York.

Lazarevic, A.; Pokrajac, D.; Marcano, A. & Melikechi, N. (2009). *Support vector machine based classification of fast Fourier transform spectroscopy of proteins*, Proc. SPIE, Vol. 7169, 71690C; DOI: 10.1117/12.809964, ISBN 9780819474155, San Jose, California, USA, January 2009, SPIE, Bellingham.

Massart, D. L.; Vandenginste, B. G.; Deming, S. N.; Michotte, Y. & Kaufman, L. (2003). *Chemometrics: A Textbook*. Elsevier, ISBN 0-444-42660-4, Amsterdam.

Marcano A.; Markushin, Y.; Melikechi, N. & D. Connolly (2008). Fourier Transform Spectroscopy of Deuterated Proteins. *Linear and Nonlinear Optics of Organic Materials VII by Rachel Jakubiak, Proc. SPIE*, Vol. 7049, pp. 70490z-1-8. ISBN 9780819472694, San Diego, California, USA, August 2008, SPIE Bellingham.

Melikechi, N.; Ding, H.; Rock, S.; Marcano, A. & Connolly, D. (2008). Laser-induced breakdown spectroscopy of whole blood and other Liquid organic compounds. *Optical Diagnostic and Sensing VIII, Editors G. Cote and A. V. Priezzhev, Proc. SPIE*, Vol. 6863, 68630O1-7, DOI:10.1117/12.761901, ISBN 9780819470386, San Jose, California, USA, January 2008, SPIE, Bellingham.

Mor, G.; Visintin, I.; Lai, Y., Zhao, H.; Schwartz, P. Rutherford, T.; Yue, L.; Bray-Ward, P. & Ward D. C. (2005). Serum protein markers for early detection of ovarian cancer. *PNAS*, Vol. 102, No. 21, 7677-7682, on line ISSN 1091-6490.

Nie, B.; Stutzman, J. & Xie, A. (2005). A vibration spectra maker probing the hydrogen-bonding status of protonated Asp and Glu residues. *Biophys. J.,* Vol. 88, No 4, 2833-2847, ISSN 0006-3495.

Petibois, C.; Gionnet, K.; Goncalves, M.; Perromat, A.; Moenner, M. & Deleris, G. (2006). Analytical performances of FT-IR spectrometry and imaging for concentration measurement within biological fluids, cells, and tissues. *Analysis*, Vol. 131, 640-647, ISSN 0003-2638.

Platt, J.; Cristianini, N. & Shawe-Taylor, J. (2000). Large Margin DAGs for Multiclass Classification. *Advances in Neural Information Processing Systems,* Vol. 12, pp. 547-553, ISBN-13 978-0-262-56145-7.

Schorge, J.O.; Drake, R. D.; Lee, H.; Skates, S. J.; Rajanbabu, R.; Miller, D. S.; Kim, J. H.; Cramer, D. W.; Berkowitz, R. S. & Mok, S. C. (2004). Osteopontin as an adjunct to CA125 in detecting recurring ovarian cancer. *Clin. Cancer Res.*, Vol. 10, No. 10, 3474-3478, ISSN 1557-3265.

Siebert, F. (1995). Infrared spectroscopy applied to biochemical and biological problems. *Methods Enzymol.*, Vol. 246, 501–526, ISBN 978-0-12-182147-0.

Sutphen, R.; Xu, Y.; Wilbanks, G. D.; Fiorica, J.; Grendys Jr., E. C.; LaPolla, J. P.; Arango, H.; Hoffman, M.S.; Martino, M.; Wakeley, K.; Griffin, D.; Blanco, R. W.; Cantor, A. B.; Xiao, Y. J. & Krischer, J. P. (2004). Lysophospholipids are potential biomarkers of ovarian cancer. *Cancer Epidemiol. Biomarkers Prev*. Vol. 13, No. 7, 1185-1191, ISSN 1055-9965.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*, Springer Verlag, ISBN 0-397-98780-0, New York.

Fourier Transforms - New Analytical Approaches and FTIR Strategies

Edited by Prof. Goran Nikolic

ISBN 978-953-307-232-6
Hard cover, 520 pages
Publisher InTech
Published online 01, April, 2011
Published in print edition April, 2011

New analytical strategies and techniques are necessary to meet requirements of modern technologies and new materials. In this sense, this book provides a thorough review of current analytical approaches, industrial practices, and strategies in Fourier transform application.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Marcano A., D. Pokrajac, A. Lazarevic, M. Smith, Y. Markushin and N. Melikechi (2011). Statistical Analysis for Automatic Identification of Ovarian Cancer Protein-Biomarkers Based on Fast Fourier Transform Infrared Spectroscopy, Fourier Transforms - New Analytical Approaches and FTIR Strategies, Prof. Goran Nikolic (Ed.), ISBN: 978-953-307-232-6, InTech, Available from: http://www.intechopen.com/books/fourier-transforms-new-analytical-approaches-and-ftir-strategies/statistical-analysis-for-automatic-identification-of-ovarian-cancer-protein-biomarkers-based-on-fast

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen