Identification of Differentially Expressed Proteins Using MALDI-TOF Mass Spectra

Qiuhua Liu,[1] Balaji Krishnapuram,[1] Pallavi Pratapa,[2]
Xuejun Liao,[1] Alexander Hartemink,[2] and Lawrence Carin[1]

[1]Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708
[2]Department of Computer Science, Duke University, Durham, NC 27708

Abstract **In the search for diagnostic and therapeutic strategies for lung cancer, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) has been evinced as a new and promising discovery platform to generate protein expression profiles in search of overexpressed proteins in lung tumors. Data from MALDI-TOF spectra require considerable signal processing such as noise removal and baseline level error correction. In this study, we discuss our preliminary approaches to these issues. We further present a novel algorithm that identified two proteins that are differentially expressed between tumor and normal tissues in patients with lung cancer. The overexpression of these two proteins was confirmed by immunoblotting, and their localization to tumor cells was studied by immunohistochemistry. Our results indicate that the proposed schemes enable accurate diagnosis based on the overexpression of a small number of identified proteins with high sensitivity, suggesting that these proteins would be helpful in further study of drug design.**

I. Introduction

Lung cancer is the leading cause of cancer mortality throughout the world, with more than 170,000 new cases in the United States each year alone. Despite extensive research in genomics, drug discovery, and better understanding of its biology and causes, the overall 5-year survival remains about 14% [1]. Recent research has demonstrated that using a MALDI-TOF platform to generate protein expression profiles from lung cancer lystates is an alternative promising strategy in the search for new diagnostic and therapeutic molecular targets. The MALDI-TOF spectra need considerable signal processing such as noise removal and baseline error correction. In this study, we discuss our preliminary approaches to these issues.

Matching pursuits (MP) is a well-known technique for representing a signal as a linear expansion of basis functions that are selected from a potentially redundant dictionary [2]. Based on MP, Kernel Matching Pursuits was introduced in [3] as a novel classification-prediction model. The KMP classification algorithm enables us to perfectly classify the MALDI-TOF mass spectra of tumor and normal tissues from patients with lung cancer. From this classification algorithm, we also obtained two mass spectrometry peaks that distinguished between these two different tissues. Overexpression of these two protein peaks were confirmed in lung cancer specimens by western blotting and localized to the tumor cells by immononhistochemistry.

II. Principle Of Maldi-tof Mass Spectrometry

Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry has become an important tool of choice for large molecular analyses, especially for proteins [4]. The schematic setup of a linear MALDI-TOF instrument is shown in Figure 1. First, the samples are mixed with an organic compound that acts as a matrix to facilitate the desorption and ionization of compounds in the sample. The analyte molecules are distributed throughout the matrix so that they are completely isolated from each other. Some of the energy incident on the sample plate is absorbed by the matrix, causing rapid vibrational excitation. The analyte molecules can become ionized by simple protonation by the photo-excited matrix, leading to the formation of the typically singly charged ions. Some multiply charged ions are also formed, but this is rarer. The analyte ions are then accelerated by an electrostatic field to a common kinetic energy. If all the ions have the same kinetic energy, the ions with low mass to charge ratio (m/z) travel faster than those with higher m/z values, therefore, they are separated in the flight tube and the number of ions reaching the detector at the end of the flight tube is recorded as the intensity of the ions. For MALDI, normally the charge is equal to one or two.

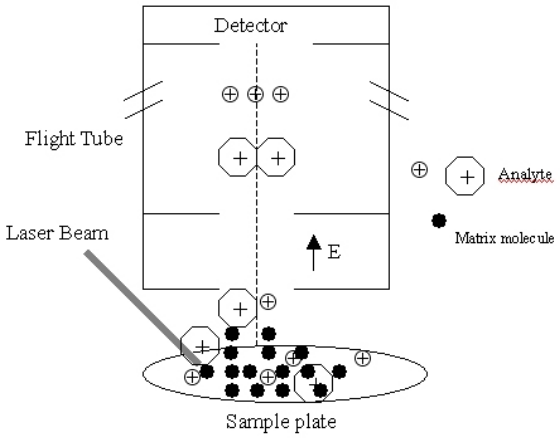

Figure 1. Principle of MALDI-TOF MS

For the MALDI-TOF mass spectra processed in this study, the analytes are tissue samples from both tumor and adjacent normal lung cells of patients with lung cancer. The matrix used is sinapinic acid. In total, there are 34 patients, and for each patient, both the normal and tumor tissues were measured 10 times independently by MALDI-TOF platform, yielding 340 mass spectra for normal tissues and 340 mass spectra for tumor tissues.

III. Pre-processing Of The Mass Spectra

In this section, we discuss our preliminary approaches for pre-processing of the MALDI-TOF mass spectra. One example of an original spectrum is shown in Figure 2(a). From it, we can see that the peaks are very noisy and the baseline at the low m/z values is much higher than the baseline at the higher m/z values. Before picking the peaks from this spectrum, we first need to improve the signal to noise ratio of the peaks and correct the baseline error. For noise removal, we smooth the spectrum with moving-average filtering. Comparing Figure 2(b) to 2(a), we can see the signal to noise ratio of the smoothed spectrum is highly improved while the peaks are all retained. For baseline correction, we compute the convex hull of the spectrum (Figure 2(c)), and subtract the convex hull from the original spectrum to get the baseline-corrected spectrum, shown in Figure 2(d).

After baseline correction, the next step is to pick the peaks from the spectrum, which will be used in the classification algorithm KMP as features. Using the spectrum after baseline correction, we declare a point in the spectrum to be a peak if the intensity is a local maximum, its absolute value is larger than a threshold (t1), and the intensity is larger than a threshold (t2) times the average intensity in the window surrounding this point. The peak picking results are shown in Figure 3. From this figure, it can be seen that we essentially picked all the principal protein peaks with little noise.

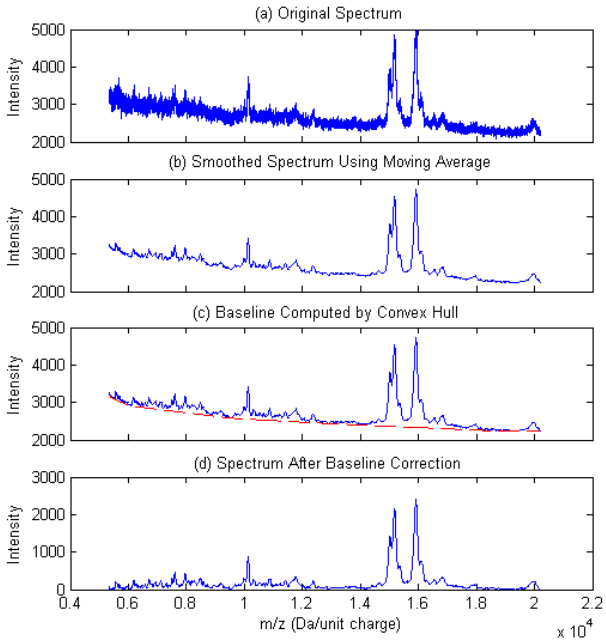

Figure 2. Pre-processing of MALDI-TOF spectra: (a) the original spectrum; (b) the smoothed spectrum; (c) the baseline of the spectrum; (d) the spectrum after baseline correction.

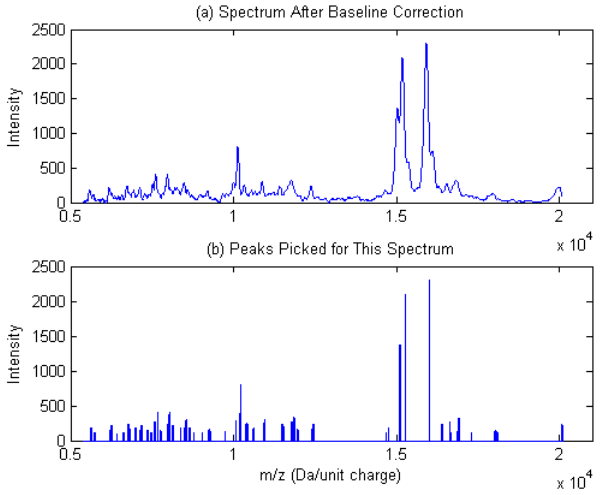

Figure 3. Peak picking from the MALDI-TOF mass spectra: (a) the spectrum after baseline correction; (b) the peaks picked

IV. Kernel Matching Pursuits (KMP) Classifier

After identifying the peaks for each of the spectra, the next step is putting these peaks into a feature selection algorithm to find the protein peaks that can distinguish the tumor tissues and the normal ones for lung cancer diagnosis. In this section, we introduce the kernel matching pursuits (KMP) classifier.

A. Estimation of the Weights

The KMP implements a set of functions of the form

$$f_n(\mathbf{x}) = \sum_{i=1}^{n} w_{n,i} K(\mathbf{c}_i, \mathbf{x}) + w_{n,0} = \mathbf{w}_n^T \boldsymbol{\phi}_n(\mathbf{x}) \quad (1)$$

where $w_{n,0}$ is the bias term, and

$$\boldsymbol{\phi}_n(\cdot) = [1, K(\mathbf{c}_1, \cdot), K(\mathbf{c}_2, \cdot), ..., K(\mathbf{c}_n, \cdot)]^T \quad (2)$$

with $K(\mathbf{c}_i, \cdot)$ the kernel-induced basis function centered at $\mathbf{c}_i$,

$$\mathbf{w}_n = [w_{n,0}, w_{n,1}, w_{n,2}, \cdots, w_{n,n}]^T \quad (3)$$

are the weights that combine the basis functions in the summation, and the subscript n is used to denote the number of basis functions being used.

Assume we are given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$ of size N, where $\mathbf{x}_i$ is the i-th input and y_i its expected output, the weighted sum of squared errors between the expected output and the KMP output given in (1) is [5]:

$$\begin{aligned} e_n &= (1/\textstyle\sum_{i=1}^N \beta_i) \sum_{i=1}^N \beta_i [y_i - f_n(\mathbf{x}_i)]^2 \\ &= (1/\textstyle\sum_{i=1}^N \beta_i) \sum_{i=1}^N \beta_i [y_i - \mathbf{w}_n^T \boldsymbol{\phi}_n(\mathbf{x}_i)]^2 \end{aligned} \quad (4)$$

where β_i is a constant responsible for the importance of the *i*-th training sample $(\mathbf{x}_i, y_i)$. For example, $1/\beta_i$ may represent the variance of the *i*th measurement. In addition, if one has *a priori* knowledge that some data $\mathbf{x}_i$ are ìbetterî representatives of the system being modeled, this can be accounted for in the parameter β_i. The unknowns in (4) are the centers $\mathbf{c}_i$ of the basis functions in $\boldsymbol{\phi}_n$, and the weights $\mathbf{w}_n$. The determination of $\mathbf{c}_i$ is addressed separately in Sec. IV. B. At the moment we suppose $\mathbf{c}_i$ and consequently $\boldsymbol{\phi}_n$ are known and aim at solving for $\mathbf{w}_n$. Then the value of $\mathbf{w}_n$ that minimizes (4) is easily found to be [5]

$$\mathbf{w}_n = \mathbf{M}_n^{-1}\{\beta_t \boldsymbol{\phi}_{n,i} y_i\}_i \tag{5}$$

where $\boldsymbol{\phi}_{n,i}$ is an abbreviation of $\boldsymbol{\phi}_n(\mathbf{x}_i)$, $\{\cdot\}_i \overset{Def}{=} \sum_{i=1}^{N}(\cdot)$, and

$$\mathbf{M}_n = \sum_{i=1}^{N} \beta_i \boldsymbol{\phi}_n(\mathbf{x}_i)\boldsymbol{\phi}_n^T(\mathbf{x}_i) = \{\beta_i \boldsymbol{\phi}_{n,i}\boldsymbol{\phi}_{n,i}^T\}_i \tag{6}$$

is the Fisher information matrix [5]. It is known that in the case that (1) is an exact model of $\int y dP(y|\mathbf{x})$ and β_i is the reciprocal of the variance of y conditional on $\mathbf{x}_i$, $i = 1,2,...,N$, (5) is the best linear unbiased estimate (BLUE) of $\mathbf{w}_n$ [5]. Note that for (5) to be BLUE, we make no assumptions with regard to the statistics of *y* conditioned on **x**, other than that of a finite second moment [5].

B. *Sequential Selection of Basis Functions*

An *n*th order KMP employs *n* basis functions. According to the definition in (1), the (*n*+1)-th order KMP is inductively written as

$$f_{n+1}(\mathbf{x}) = \mathbf{w}_{n+1}^T \boldsymbol{\phi}_{n+1}(\mathbf{x}) \tag{7}$$

where

$$\begin{aligned}\boldsymbol{\phi}_{n+1}(\cdot) &= [1, K(\mathbf{c}_1,\cdot), K(\mathbf{c}_2,\cdot),..., K(\mathbf{c}_n,\cdot), K(\mathbf{c}_{n+1},\cdot)]^T \\ &= \begin{bmatrix}\boldsymbol{\phi}_n(\cdot) \\ \phi_{n+1}(\cdot)\end{bmatrix}\end{aligned} \tag{8}$$

with $\phi_{n+1}(.) = K(\mathbf{c}_{n+1},\cdot)$ a new basis function centered at $\mathbf{c}_{n+1}$. The weighted sum of squared errors of the (*n*+1)-th order KMP is

$$e_{n+1} = (1/\textstyle\sum_{i=1}^{N}\beta_i)\sum_{i=1}^{N}\beta_i[y_i - f_{n+1}(\mathbf{x}_i)]^2 \tag{9}$$

Assuming the basis functions in $\boldsymbol{\phi}_{n+1}$ are all known, then according to (5)

$$\mathbf{w}_{n+1} = \mathbf{M}_{n+1}^{-1}\{\beta_i \boldsymbol{\phi}_{n+1,i} y_i\}_i \tag{10}$$

minimizes (9), where the Fisher information matrix $\mathbf{M}_{n+1}$ is given as

$$\mathbf{M}_{n+1} = \{\beta_i \boldsymbol{\phi}_{n+1,i}\boldsymbol{\phi}_{n+1,i}^T\}_i \tag{11}$$

It can be shown that $\mathbf{w}_{n+1}$, and e_{n+1} are respectively related to $\mathbf{w}_n$ and e_n as [6]:

$$\mathbf{w}_{n+1} = \begin{bmatrix}\mathbf{w}_n + \mathbf{M}_n^{-1}\{\beta_i \boldsymbol{\phi}_{n,i}\phi_{n+1,i}\}_i b^{-1}[\{\beta_i \boldsymbol{\phi}_{n,i}^T \phi_{n+1,i}\}_i \mathbf{w}_n - \{\beta_i \phi_{n+1,i} y_i\}_i] \\ -b^{-1}\{\beta_i \boldsymbol{\phi}_{n,i}^T \phi_{n+1,i}\}_i \mathbf{w}_n + b^{-1}\{\beta_i \phi_{n+1,i} y_i\}_i\end{bmatrix} \tag{12}$$

$$e_{n+1} = e_n - \delta e(K, \mathbf{c}_{n+1}) \tag{13}$$

where

$$\delta e(K, \mathbf{c}_{n+1}) = (1/\textstyle\sum_{i=1}^{N}\beta_i) b^{-1}[\{\beta_i \boldsymbol{\phi}_{n,i}^T \phi_{n+1,i}\}_i \mathbf{w}_n - \{\beta_i \phi_{n+1,i} y_i\}_i]^2 \tag{14}$$

and

$$b = \{\beta_i \phi_{n+1,i}^2\}_i - \{\beta_i \phi_{n+1,i}\boldsymbol{\phi}_{n,i}^T\}_i \mathbf{M}_n^{-1}\{\beta_i \boldsymbol{\phi}_{n,i}\phi_{n+1,i}\}_i \tag{15}$$

with $\phi_{n+1,i} = K(\mathbf{c}_{n+1}, \mathbf{x}_i)$.

It can be shown that b^{-1} is a diagonal element of $\mathbf{M}_{n+1}^{-1}$ [6]. With sufficient training data points, we can always make $\mathbf{M}_{n+1}$ positive definite, as can be seen from (11). Then $\mathbf{M}_{n+1}^{-1}$ is also positive definite and it holds that $b^{-1} > 0$, which guarantees $\delta e(K, \mathbf{c}_{n+1})$ is always non-negative. Therefore, from (13), $e_{n+1} < e_n$, which means appending a new basis function to the KMP generally leads to decrease of the representation error.

As $\delta e(K, \mathbf{c}_{n+1})$ is dependent on the center $\mathbf{c}_{n+1}$ of the new basis function, we obtain different values of $\delta e(K, \mathbf{c}_{n+1})$ by selecting different $\mathbf{c}_{n+1}$. If we confine $\mathbf{c}_{n+1}$ to be selected from the training data, we then can conduct a ìgreedyî search in the training set but with the previously selected data excluded to avoid repetition, and select the datum that maximizes (14). Formally, we have

$$\mathbf{c}_{n+1} = \mathbf{x}_{i_{n+1}} = \arg\max_{\substack{k \neq i_1,...,i_n \\ 1 \le k \le N}} \delta e(K, \mathbf{x}_k) \tag{16}$$

After $\mathbf{c}_{n+1}$ is determined, we update the weights using (12) and the Fisher information matrix using (11) and (8).

C. Kernel Optimization

From (14), $\delta e(K, \mathbf{c}_{n+1})$ depends on the functional form of the kernel $K(\cdot,\cdot)$ as well as on $\mathbf{c}_{n+1}$. This allows us to optimize the kernel to gain further error reduction. A simple approach to take is to first conduct a ìgreedyî search of $\mathbf{c}_{n+1}$ in the training set, for a fixed kernel, and then fix $\mathbf{c}_{n+1}$ and optimize the parameters of the kernel. For radial basis function (RBF) kernels, the only parameter other than $\mathbf{c}_{n+1}$ is the kernel width, thus optimization of RBF kernels with $\mathbf{c}_{n+1}$ fixed is a one-dimensional search for the kernel width. It is also possible to optimize $\mathbf{c}_{n+1}$ and the kernel width simultaneously, but then $\mathbf{c}_{n+1}$ is treated as a free parameter and no longer confined to the training set. Another possibility is optimization over kernels of different functional forms, which offers greater diversity of the basis functions available to the KMP.

V. Finding Differentially Expressed Proteins

Using the peaks selected in Sec. III as features for the KMP algorithm, we try to find those peaks that can distinguish the tumor tissues from the normal ones. Figure 4 shows the KMP classification results when trained with features of (a) m/z larger than 5,000 Da/unit charge; (b) m/z larger than 10,000 Da/unit charge and (c) m/z larger than 15,000 Da/unit charge. The amplitude of a given m/z value is used as the feature. From Figure 4, we can see with only a single feature, the correct classification rate for (a), (b), and (c) is 78%, 69%, and 59% respectively. Therefore, the corresponding peaks should

be able to distinguish the tumor and the normal tissues effectively. The corresponding m/z values for the first six features selected by KMP of each of the cases in Figure 4 are listed in Table 1.

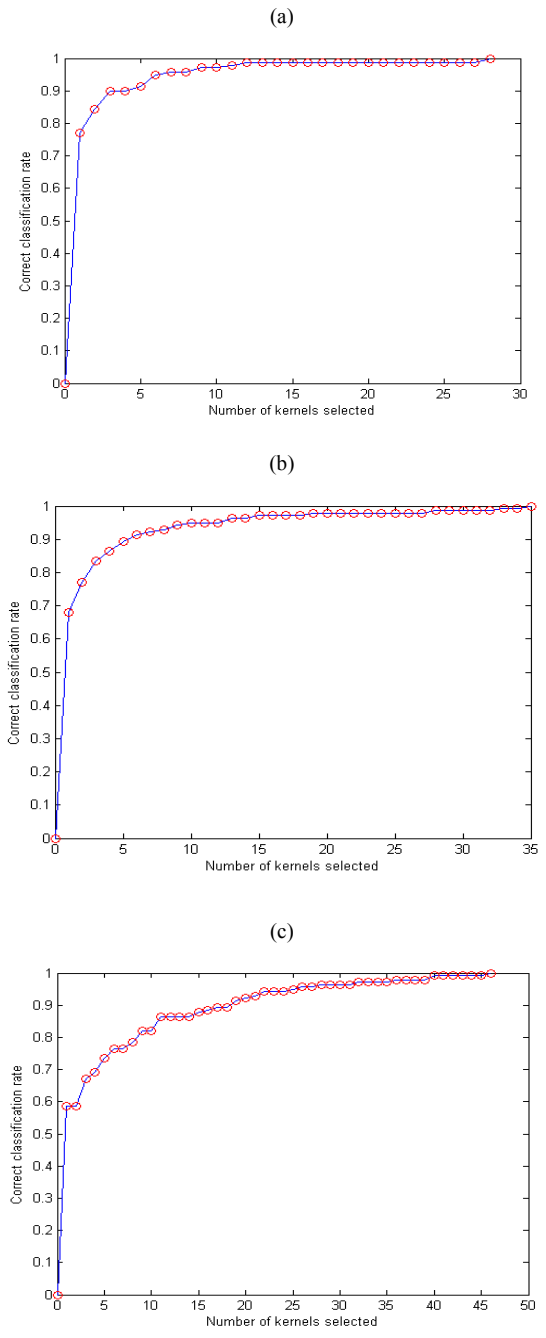

Figure 4. KMP Feature Selection Results (a) Training KMP with m/z larger than 5,000 Da/charge; (b) Training KMP with m/z larger than 10,000 Da/charge; (c) Training KMP with m/z larger than 15,000 Da/charge

Table 1. The features selected by KMP as potential protein peaks to distinguish normal and tumor tissues

Case in Figure 4	The corresponding m/z values of the features					
	The first feature	The second feature	The third feature	The fourth feature	The fifth feature	The sixth feature
(a)	6203	6210	6195	9771	8490	7818
(b)	12379	12394	10330	10991	13009	12376
(c)	17966	17755	15010	17148	17951	17907

According to the mass accuracy of the MALDI spectra (about 0.4% in this data set), in those features listed in Table 1, the peaks with m/z values 6203, 6210, and 6195 are from one protein; the peaks with 12379, 12394, and 12376 are from one protein; and the peaks with m/z 17966, 17755, and 17907 are from one protein. We then checked the three peaks from the original spectra and found that those peaks only exist in the spectra from tumor tissues. By immunohistochemical analysis, the two differentially expressed protein peaks around m/z 12,379 and 17,966 were identified in lung tumor tissues as the proteins macrophase migration inhibitory factor and cyclophilin A, which have a mass 12,338 and 17,882 Da respectively. The overexpression of both proteins was confirmed by Western blotting. We did not find a particular protein that has a mass around 6203 Da in the tumor tissues. However, the value of 6203 is approximately one half of the value 12,338 and we can see these two peaks are most probably from the same protein macrophase migration inhibitory factor, but the ions were doubly-charged during ionization for the 6203 peak and singly-charged for the 12,338 peak when doing the MALDI-TOF experiment. Figure 5 shows a close look at the smoothed mass spectra from normal and tumor tissues around m/z 12,338 and 17,882 Da/charge for a particular patient, where the spectrum for tumor tissue was shifted upward by 300 for visual clarity. The mass spectrum from the patientís tumor lung tissue shows the two protein peaks at these two m/z values and the mass spectrum from the patientís normal lung tissue does not show these two peaks.

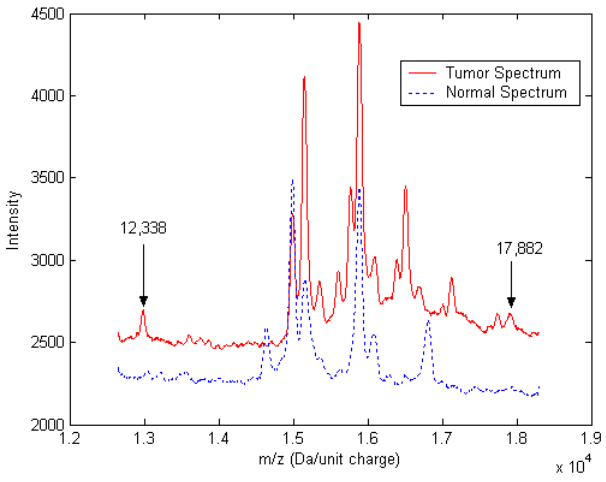

Figure 5. A close look at the MALDI-TOF mass spectra with the two identified protein peaks labeled

VI. Conclusions

In this study, MALDI-TOF mass spectrometry was used as a platform to generate protein expression profiles for tissues of lung cancer patients. The preprocessing algorithms and the KMP feature selection algorithm proposed in this study found two differentially expressed proteins that only exist in the spectra of tumor tissues. These two proteins were identified by immnohistochemistry and confirmed to be expressed in the tumor cells by Western blotting. These results demonstrate the feasibility of using MALDI-TOF mass spectra and the proposed protein finding algorithm to identify potential molecular targets for cancer diagnostics and therapeutics.

Acknowledgment

The mass spectra analyzed in this study were provided by Dr. Michael J. Campa and Dr. Edward F. Patzís research group at Duke University Medical Center and measured at the Duke University Department of Chemistry. We thank the various individuals in these groups for their useful discussion and encouragement.

References

[1] B.K. Edwards, H.L. Howe, L.A. Ries, M.J. Thun, H.M. Rosenberg, R. Yancik, P.A. Wingo, A. Jemal, and E.G. Feigal, ìAnnual report to the nation on the status of cancer, 1973-1999, featuring implications of age and aging on U.S. cancer burdenî, Cancer, 94: 2766-2792, 2002.

[2] P. Vincent and Y. Bengio. ìKernel matching pursuitî, Machine Learning, 48: 165-187, 2002.

[3] S. Chen, F. Cowan, and P. Grant, ìOrthogonal least squares learning algorithm for radial basis function networksî, IEEE Transactions on Neural Networks, Vol. 2, No. 2: 302-309, 1991.

[4] http://www.srsmaldi.com/MALDI.html.

[5] V.V. Fedorov, Theory of Optimal Experiments, Academic Press, 1972.

[6] X. Liao, H. Li, and B. Krishnapurum, ìAn M-ary classifier for multi-aspect target classificationî, submitted to the 2004 International Conference on Acoustics, Speech, and Signal Processing.