

2010

Prostate Cancer Region Prediction Using MALDI Mass Spectra

Ayyappa Vadlamudi
Old Dominion University

Shao-Hui Chuang
Old Dominion University


Xiaoyan Sun
Old Dominion University

Lisa Cazares
Eastern Virginia Medical School

Julius Nyalwidhe
Eastern Virginia Medical School

See next page for additional authors

Follow this and additional works at: https://digitalcommons.odu.edu/ece_fac_pubs

 Part of the [Amino Acids, Peptides, and Proteins Commons](#), [Bioimaging and Biomedical Optics Commons](#), [Biomedical Commons](#), and the [Urology Commons](#)

Original Publication Citation

Vadlamudi, A., Chuang, S.-H., Sun, X., Cazares, L., Nyalwidhe, J., Troyer, D., Semmes, O. J., Li, J., & McKenzie, F. (2010) Prostate cancer region prediction using MALDI mass spectra. In N. Karssemeijer & R. M. Summers (Eds.), *Medical Imaging 2010: Computer-Aided Diagnosis, Proceedings of SPIE Vol. 7624* (762422). SPIE of Bellingham, WA. <https://doi.org/10.1117/12.844494>

This Conference Paper is brought to you for free and open access by the Electrical & Computer Engineering at ODU Digital Commons. It has been accepted for inclusion in Electrical & Computer Engineering Faculty Publications by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

Authors

Ayyappa Vadlamudi, Shao-Hui Chuang, Xiaoyan Sun, Lisa Cazares, Julius Nyalwidhe, Dean Troyer, O. John Semmes, Jiang Li, and Frederic D. McKenzie

Prostate Cancer Region Prediction using MALDI Mass Spectra

¹Ayyappa Vadlamudi, ¹Shao-Hui Chuang, ¹Xiaoyan Sun, ²Lisa Cazares, ²Julius Nyalwidhe,

²Dean Troyer, ²O. John Semmes, ¹Jiang Li, ¹Frederic D. McKenzie

¹Dept. Of Electrical and Computer Engineering, Old Dominion University
Norfolk VA, USA, 23509-0246

²Dept. Of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School
Norfolk VA, USA, 23507

ABSTRACT

For the early detection of prostate cancer, the analysis of the Prostate-specific antigen (PSA) in serum is currently the most popular approach. However, previous studies show that 15% of men have prostate cancer even their PSA concentrations are low. MALDI Mass Spectrometry (MS) proves to be a better technology to discover molecular tools for early cancer detection. The molecular tools or peptides are termed as biomarkers. Using MALDI MS data from prostate tissue samples, prostate cancer biomarkers can be identified by searching for molecular or molecular combination that can differentiate cancer tissue regions from normal ones. Cancer tissue regions are usually identified by pathologists after examining H&E stained histological microscopy images. Unfortunately, histopathological examination is currently done on an adjacent slice because the H&E staining process will change tissue's protein structure and it will derivate MALDI analysis if the same tissue is used, while the MALDI imaging process will destroy the tissue slice so that it is no longer available for histopathological exam. For this reason, only the most confident cancer region resulting from the histopathological examination on an adjacent slice will be used to guide the biomarker identification. It is obvious that a better cancer boundary delimitation on the MALDI imaging slice would be beneficial. In this paper, we proposed methods to predict the true cancer boundary, using the MALDI MS data, from the most confident cancer region given by pathologists on an adjacent slice.

Keywords : Prostate cancer, Biomarker, Imaging biomarker, MALDI

1. INTRODUCTION

Prostate cancer is a disease that develops in the prostate gland, a male reproductive element being in the size of a walnut and present beneath the bladder just opposite the rectum. Prostate cancer is a dreaded non-skin cancer that kills one man every nineteen minutes and studies show that every 1 in 6 men suffers from this disease [1]. Prostate cancer is difficult to be cured if it is metastasized to other organs, so an early detection is demanded. Currently, Prostate specific antigen (PSA) test is the most common early diagnosis methods, where a sample blood is taken and the level of PSA is checked [2]. A high level (approx 10ng/ml) is considered to be a victim of prostate cancer. However, recent studies prove that some patients with low PSA concentrations are also affected with prostate cancer, showing that the PSA level is not an efficient tool for the early detection. Therefore, there is a need to discover more accurate and specific detection methods for the prostate cancer.

A biological marker usually refers to a specific protein or their products, which indicate the states of a disease. A prostate cancer biomarker could be a protein or a molecular, or a combination of proteins, which is related to prostate cancer status. Recently, there is more and more research for searching prostate cancer biomarkers using mass spectrometry (MS) based techniques. The two most widely used MS based methods involve surface-enhanced laser desorption/ionization (SELDI) and matrix-assisted laser desorption/ionization (MALDI) time of flight approaches. Every peak in the mass spectrum corresponds to a specific protein having a high concentration, and the goal of prostate cancer biomarker identification is to identify peaks that are related to specific outcomes or malignancy stages of prostate cancer.

Prostate cancer biomarker identification using MALDI MS data consists of several pre-processing steps: 1) baseline adjustment, 2) denoising, 3) normalization, 4) peak detection and 5) peak re-binning. After the pre-processing steps, a set of peaks are available and we need to identify a set of peaks that can discriminate spectra from normal tissue regions and

cancer areas. Feature selection and classification algorithms are usually applied for this task. For example, several feature selection algorithms have been proposed including the one proposed by Dossat et. al [3], J-5 test, simple separability criterion and weighted separability criterion. Classification methods such as support vector machine, multilayer perceptron, Bayesian classifiers have found application for the classification purpose.

One of the important issue for prostate cancer biomarker identification is that the definition of cancer boundary should be accurate. The detection of the true cancer boundary has a handful of challenges to be faced, as the MALDI processing and the histopathological examination by an individual pathologist are not done on the same tissue in the current practice. The MALDI destroys the tissue on which the processing is done and vice versa, and hence an adjacent tissue slice is used for the latter. In this paper, we describe the various preprocessing steps in prostate biomarker identification and we will present a method for predicting the true cancer boundary on the MALDI destroyed tissue slice to guide the biomarker identification process.

2. METHODS

A prostate biopsy tissue was obtained at the Eastern Virginia Medical School (EVMS) from which two adjacent slices were taken. One tissue slice was used to obtain MALDI MS data and another H&E stained slice was utilized for pathological examination. We will briefly introduce each of the steps in this study in the following sections.

2.1 MALDI Imaging. MALDI imaging is a soft ionization technique used in mass spectrometry for the analysis of the bio-molecules such as peptides and proteins. The MALDI laser hits the tissue slice at spots of about 10 to 50 micrometers each in diameter across the entire tissue, yielding a MS spectrum for each of the 974 spots on the slice. As MALDI is a time-of-flight approach, a mass spectrum is represented by a vector whose dimensionality is equal to the number of distinct m/z values recorded and the value of each dimension is the concentration of the corresponding m/z value. From the spectra derived from the MALDI MS, one normal and one cancer spectra are shown in Fig 1a and Fig. 1b, respectively. Coordinates of the spots were then used to constructed an artificial image representing the shape of the tissue (Fig. 1c). The most confident cancer region identified by a pathologist from EVMS was shown in white in the artificial image (white region in Fig. 1c, named as AI1).

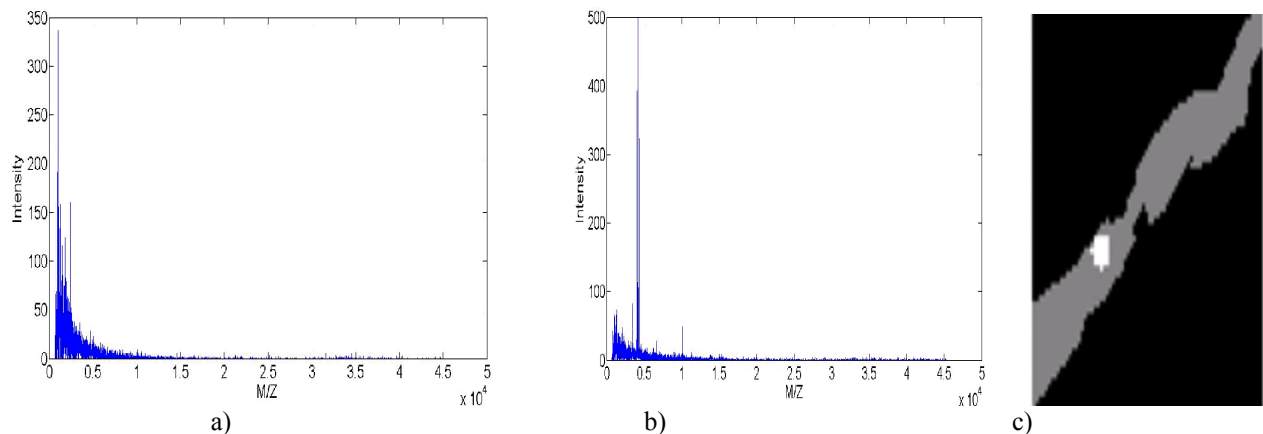


Fig .1a). Normal Spectra, b). Cancer spectra, c). Artificial tissue image generated from the coordinates of MALDI spectra.

2. 2 Histopathological analysis. The H&E stained slice was scanned using a Hirox HI-SCOPE KH-1300 microscope at a magnification of 420 resulting a set of images. Fig. 2a and Fig. 2b show a normal and a cancer images. The whole tissue image was then reconstructed from the scanned small regions with at least 20% overlap with each other as shown in Fig. 2c. The pathologist then analyzed the image and annotated cancer regions as shown in Fig. 2d (named as AI2). For the reason of computational efficiency, a sub-image covering the identified cancer regions and some neighbor normal regions was cropped for subsequent analysis (Fig. 2c).

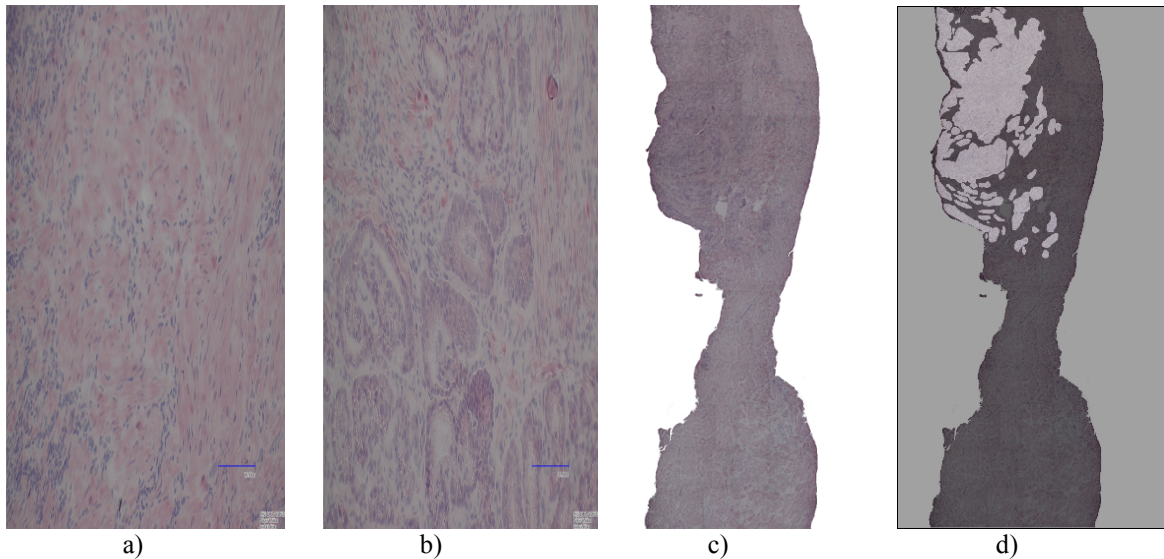


Figure 2. a). Normal region, b). Cancer region, c). Stitched image, d). Stitched image with annotated cancer regions.

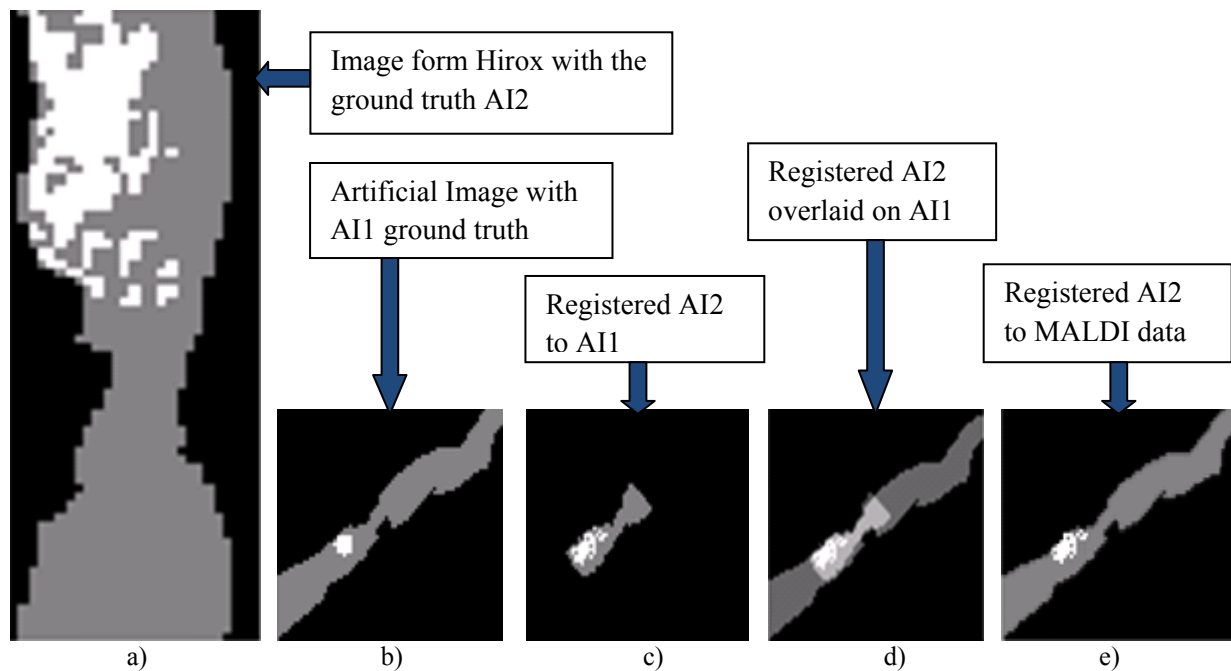


Figure 3. Registration of the high resolution ground truth to MALDI MS data. a) High resolution cancer region mask produced by the pathologist (AI2), b) Artificial tissue image constructed from the coordinates of the MALDI MS data, c) Registered high resolution images, d) Overlaid ground truths AI1 and AI2 and e) Registered AI2 to MALDI MS data.

2. 3 Registration. The Hirox scanned image with annotated cancer regions was registered to the artificial tissue image reconstructed from the coordinates of MALDI MS data, by using a landmark based registration method implemented in Matlab. Cancer regions identified on the Hirox image were mapped to the artificial image (Fig. 3). To confirm an exact match between the two images, AI2 is overlaid on AI1 (Fig. 3d). Since the two images come from two adjacent slices, there are small offsets after the registration. Now the white part (tumor part) from the registered histopathological image is then mapped on to the MALDI image (Fig. 3e).

2. 4 MALDI spectra preprocessing steps: The raw MALDI data we received from the EVMS needs to undergo a series of pre preprocessing steps [5]. They include

a. Baseline adjustment: In general, it is recommended to remove the ion and chemical noise that are usually higher at smaller m/z values. As the baseline characteristics vary from one experiment to another there exists no general solution to this problem. In this paper, A MATLAB function called msbackadj is used. It estimates the baseline within multiple shifted windows and regresses the varying baseline using spline approximation to the window points (Fig. 4a).

b. Smoothing: A wavelet based algorithm is used for denoising and thus enhancing the signal to noise ratio. The m/z values which are lower than 3000 have larger noise, and the m/z values greater than 10,000 have low intensities. So, these particular set of intensities were discarded.

c. Normalization: With multiple spectra, problems like systematic variation among spectra due to varying amount of proteins in the detector sensitivity arouses. It is a good practice to remove these effects. A common factor is used to scale the mass intensities for the same peak from different spectra. For a given peak the area under the peak is computed, the common factor for each peak then is defined as the ratio of the area under this peak to the median of areas of all other peaks in single spectrum.

d. Peak detection: In order to identify and quantify the proteins in mass spectra, the crucial step is to find the m/z values that correspond to higher intensities. This method of peak detection satisfies the criteria that the intensity exceeds a specific threshold of 10, under which all other intensities are zeroed. One example spectra with detected peaks is shown in Fig. 4b. After the process of smoothing and peak detection, a total of 75,719 peaks are obtained from the total 974 spectra.

e. Clustering/re-binning: This step is to align or cluster the same peak from different spectra and to assign a cluster number to every peak found in all spectra with slightly different m/z values. The same peak may have slightly different m/z values due to the fact as the spectra exhibits shifts in the horizontal axis. In this step, we merge all the peaks that have the same m/z values within 0.13% of each other and assigned it as a new peak. This step yields us a total of 820 clusters which represent 820 different peaks. The next step is to back project these peaks onto individual spectra for identifying the biomarkers. In this process of back projection, if there is a peak in one individual spectrum corresponds to a cluster, the intensity of the peak is kept as is. If there is no peak then the cluster point is replaced by zero. After all the raw data are preprocessed we obtain a set of peaks and are denoted as $\{\mathbf{x}_p, i_p\}_{p=1}^{N_v}$, where $\mathbf{x}_p \in R^N$ and $i_p \in R$, \mathbf{x}_p is a vector containing all peaks detected from p^{th} spectrum and i_p is a class ID (1: normal, 2: cancer) associated with this spectrum. The total number of peaks from each of the spectra is N (820 in our study) and the total number of spectra is N_v (974 in this paper). The class ID is obtained from the pathological analysis results and is considered as the ground truth. Note that we have different resolution ground truths so that a spectra might have different class ID based on which ground truth we will consider. Fig. 4 shows the results of the above preprocessing steps.

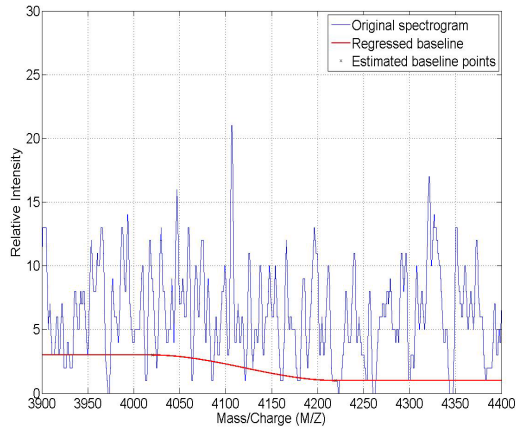
2. 5 Feature selection: This feature selection algorithm[6, 7] helps us choose the best features or the combination of features which correlates to cancer. This algorithm consists of three important components: a piecewise linear classifier, an output reset algorithm and a floating search algorithm.

1.

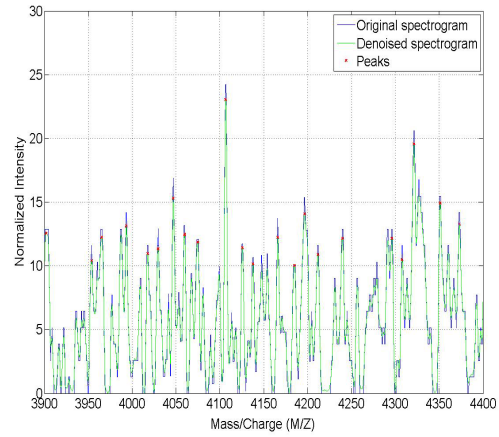
Piecewise linear classifier (PLC): The piece wise linear classifier is a neural classifier[8], which is usually designed by minimizing the standard training error given by

$$E = \sum_{i=1}^{N_{class}} E(i) = \frac{1}{N_v} \sum_{i=1}^{N_{class}} \sum_{p=1}^{N_v} [t_p(i) - y_p(i)]^2 \quad (1)$$

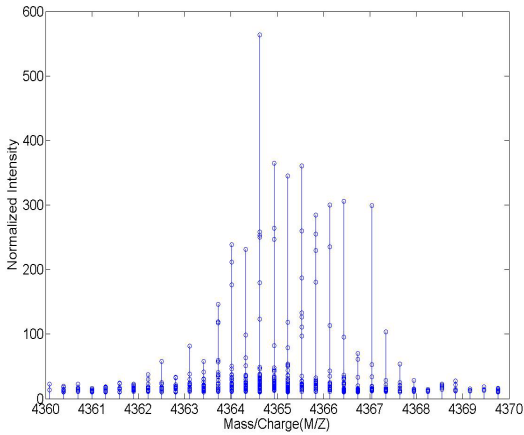
where N_{class} is the number of classes and $E(i)$ is the mean squared error for the i^{th} output. Here $t_p(i)$ denotes the i^{th} desired output for the p^{th} pattern, $y_p(i)$ denotes the i^{th} observed output for the p^{th} input pattern, and N_v denotes the total number of data patterns. In the PLC $y_p(i)$ is the output from the piece wise linear network..



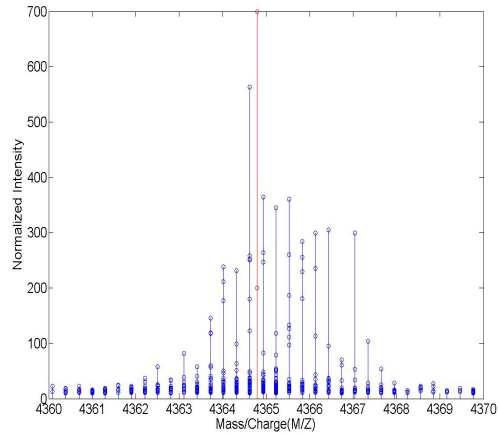
a)



b)



c)



d)

Fig 4. Results of a). Base line adjustment, b) Smoothing, normalization and peak detection, c). Back projection of peaks, d). Clustering and re-binning.

$$y_p(i) = \sum_{j=1}^{N+1} w^{(q)}(i, j) x_p^{(q)}(j) \quad (2)$$

where N is the number of features, $w^{(q)}(i, j)$ denotes the model weight to the i th output from the j th feature in the q th cluster, $x_p^{(q)}(j)$ is the j th feature in the q th cluster, and $x_p^{(q)}(N+1)$ is the bias term which equals to one. The PLC works by approximating the general Bayes discriminant, by dividing the available data into set of clusters, and by solving a set of linear equations and a local linear model is obtained for each cluster. We assume that $t_p(i_c) = 1$ and $t_p(i_d) = -1$, where i_c denotes the correct class number and i_d any incorrect class number for the current data pattern if $i_c = \underset{i}{\operatorname{argmax}} y_p(i)$. We say that PLC is correctly classified, otherwise an error is counted.

2.6 Floating search algorithm: Floating search method is designed through Piece wise Linear Orthonormal Least Square (PLOLS) procedure. The modified Schmidt procedure is utilized in PLOLS making each feature in the cluster orthonormal. When the data passes this procedure once, all the information which is required to search a good combination is stored in the auto- correlation and cross- correlation matrices. This feature selection algorithm is very efficient and only one data pass is required. The modified desired output may be represented in a matrix form as

$$t' = x^{(q)}w^{(q)} + \Xi^{(q)} \quad (3)$$

where each row in matrix $x^{(q)}$ represent one feature vector that was assigned to the q^{th} cluster, $w^{(q)}$ denotes weight matrix in the q^{th} cluster, and $\Xi^{(q)}$ are residual errors in the q^{th} cluster [6, 7].

2.7 Classification: Once the feature selection algorithm selects a compact set of good combination of features, we utilized a multi layer perceptron (MLP) to classify the spectra as normal or cancer [8, 9]. Unlike classical objective functions, the classifier uses a new objective function with more free parameters and to solve multiple sets of linear equations, the classifier used an iterative minimization technique. The error function with respect to the hidden weights are reduced by an enhanced feed forward network. In combination of the output reset (OR) enhanced MLP, we used an algorithm called Output Weight optimization-Hidden Weight Optimization (OWO-HWO) to train a classifier. In OWO-HWO, the hidden unit weights and the output weights are modified alternately to reduce the training error.

2.8 Ground truths generation: We have three ground truths in this study: AI1, AI2 and AI1&AI2 (Fig. 5a, Fig. 5b and Fig. 5c, respectively). The AI1 was generated by a quick looking at the adjacent slice and mapping back the most confident cancer region to MALDI data. The AI2 was achieved by a pathologist by carefully examining the Hirox scanned image and mapping back the identified cancer regions to MALDI data. We also intersected the AI1 and the AI2 to get their common areas (AI1&AI2).

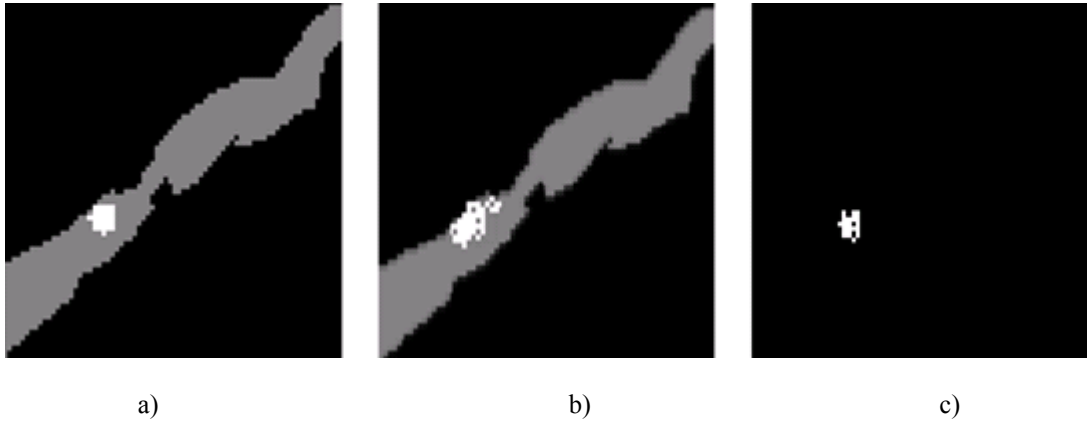


Figure 5: Ground truths generation. a). High confident ground truth identified by a quick looking at the adjacent tissue slice and mapping back the most confident region to MALDI data denoted as AI1, b). The complete cancer region identified by the pathologist on the adjacent tissue slice and mapped back to MALDI data denoted as AI2, c). Intersection of AI1 and AI2 denoted as AI1&AI2.

3. EXPERIMENTS:

We contacted three experiments in this study. In experiment A, cancer MS spectra for training were taken from the intersection of the cancer areas in AI1 and AI2 (Fig. 5c), and normal MS spectra for training were taken from a normal area that is not close to the cancer spots, yielding 19 cancer and 19 normal spectra in the training data. We then performed feature selection and classification on those preprocessed training spectra, yielding a classifier model. The trained model was subsequently applied to AI2 to obtain testing classification accuracy.

In experiment B, cancer MS spectra for training were taken from the cancer areas in AI1, and normal MS spectra for training were obtained in a similar way as that in experiment A. The training data contained 27 cancer and 27 normal spectra. After feature selection and classification, a trained model was applied to AI2 to get testing performances. This is the most important experiment because the overall goal of this study is to predict a more accurate cancer boundary based on a low-resolution ground truth and to help biomarker identification task.

In experiment C, we performed a 3-fold cross validation (CV) using the pathologist annotated cancer and normal regions in AI2 as ground truth. There were 51 cancer and 923 normal spectra in the data set. In the 3-fold CV procedure, data was partitioned into 3 parts, 2 parts were used for training and the remaining part was used for testing. This procedure was repeated three times such that each part was used for testing once. The tested parts were then pooled together to compute sensitivity and specificity for the classification.

4. RESULTS:

Sensitivities and specificities were calculated for each of the experiments and are shown in Table 1. The feature selection algorithm selected 5, 5 and 3 peaks for classification in experiment A, B and C, respectively. In experiment A, we achieved a sensitivity of 50.92% and a specificity of 98.65%. In experiment B, we obtained a sensitivity of 54.90% and a specificity of 98.48%. In experiment C, a sensitivity of 62.75% and a specificity of 98.37% were achieved. Those prediction results were mapped back to the MALDI data and are shown in Fig. 6.

Table 1. The overall result of Experiments A, B & C

| Experiments | Sensitivity | Specificity | Accuracy |
|-------------|-------------|-------------|----------|
| Exp A | 50.92% | 98.65% | 97.13% |
| Exp B | 54.90% | 98.48% | 96.20% |
| Exp C | 62.75% | 98.37% | 96.52% |

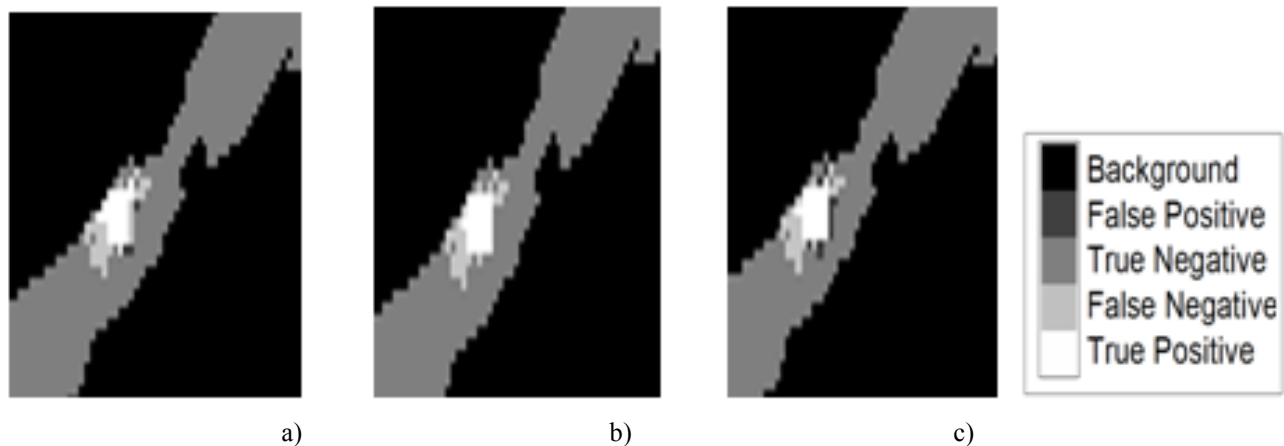


Figure 6. a) Result for experiment A. b) Result for experiment B. c) Result for experiment C

5. CONCLUSION

The overall goal of this study is to predict true cancer regions in MALDI data based on a high confident cancer region identified on an adjacent prostate tissue slice. If successful, the predicted true cancer regions will greatly assist the prostate cancer biomarker identification task, and will save the time consuming cancer identification process. It is observed that the predictions in our study are very specific with specificities close to 100% but sensitivities could be improved. In a company paper [10], we utilized a texture analysis technique based on the Hirox scanned image for the same purpose and achieved high sensitivities but relatively low specificities. We are currently trying to fuse those two complementary results and expecting better results, which are both sensitive and specific.

REFERENCES

- [1] Prostatecancerfoundation.org.
- [2] J. G. F. Francis, et al., "Measurement of prostate-specific antigen in serum as a screening test for prostate cancer," *New England Journal of Medicine*, vol. 324, 1991, pp 1156-1161, 1991.
- [3] N. Dossat, A. Mang, J. Solassol, W. Jacot, Ludovic Lhermitte, T. Maudelonde, J. P. Dauris, and N. Molinari, "Comparison of supervised classification methods for protein profiling in cancer diagnosis," *Cancer Informatics*, Vol. 3, pp. 295–305, 2007.
- [4] Dale McLerran, et al., "SELDI-TOF MS whole serum proteomic profiling with IMAC surface does not reliably detect prostate cancer," *Clinical Chemistry*, vol. 54, pp 53-60, 2008
- [5] Vamsi Mantena, Wenjuan Jiang, Jiang Li and Rick McKenzie, "Prostate cancer biomarker identification using MALDI-MS data: initial results", *IEEE/NIH LiSSA'09*, pp. 116-119, 2009.
- [6] Jiang. Li, J. Yao, R. M. Summers, N. Petrick, M. T. Manry, and A. K. Hara, "An efficient feature selection algorithm for computer-aided polyp detection," *International Journal on Artificial Intelligence Tools*, vol. 15, pp. 893-915, 2006.
- [7] Jiang. Li, M. T. Manry, P. L. Narasimha, and C. Yu, "Feature selection using a piecewise linear network," *IEEE Transaction on Neural Network*, vol. 17, pp. 1101-1105, 2006.
- [8] Jiang Li, Michael T. Manry, Li-Min Liu, Changhua Yu, and John Wei, "Iterative improvement of neural classifiers," *Proceedings of the Seventeenth International Conference of the Florida AI Research Society*, pp. 700-705, 2004.
- [9] R. G. Gore, Jiang Li, M. T. Manry, L. M. Liu, and Changhua Yu, "Iterative design of neural network classifiers through regression," *Special Issue of International Journal on Artificial Intelligence Tools*, vol. 14, pp. 281-302, 2005.
- [10] Shao-Hui Chuang, Xiaoyan Sun, Lisa Cazares, Julius Nyalwidhe, Dean Troyer, O. John Semmes, Jiang Li and Frederic D. McKenzie, "Adjacent Slice Prostate Cancer Prediction to Inform MALDI Imaging Biomarker Analyses", accepted by *SPIE Medical Imaging*, 2010.