Prostate cancer region prediction by fusing results from MALDI spectra—processing and texture analysis
Shao-hui Chuang, Jiang Li, Xiaoyan Sun, Ayyappa Vadlamudi, Bo Sun, Lisa Cazares, Julius Nyalwidhe, Dean Troyer, John Semmes and Frederic D McKenzie
SIMULATION published online 21 March 2012
DOI: 10.1177/0037549712441522

The online version of this article can be found at:
http://sim.sagepub.com/content/early/2012/03/21/0037549712441522

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for Modeling and Simulation International (SCS)

Additional services and information for SIMULATION can be found at:

Email Alerts: http://sim.sagepub.com/cgi/alerts

Subscriptions: http://sim.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> OnlineFirst Version of Record - Mar 21, 2012

What is This?
Prostate cancer region prediction by fusing results from MALDI spectra–processing and texture analysis

Shao-Hui Chuang1, Jiang Li1, Xiaoyan Sun1, Ayyappa Vadlamudi1, Bo Sun2, Lisa Cazares3, Julius Nyalwidhe3, Dean Troyer3, John Semmes3 and Frederic D McKenzie4

Abstract
We present a three-step method to predict prostate cancer (PCa) regions on biopsy tissue samples based on high-confidence, low-resolution PCa regions marked by a pathologist. First, we will apply a texture-analysis technique on a high-magnification optical image to predict PCa regions on an adjacent tissue slice. Second, we will design a prediction model for the same purpose, using matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) tissue-imaging data from the adjacent slice. Finally, we will fuse those two results to obtain the PCa regions that will assist MALDI imaging biomarker identification. Experiment results show that the texture analysis–based prediction is sensitive (87.45%) but less specific (75%), and the prediction based on the MALDI spectra data processing is not sensitive (50.98%) but supremely specific (100%). By combining these two results, an optimized prediction for PCa regions on the adjacent slice can be achieved (sensitivity: 80.39%, specificity: 93.09%).

Keywords
prostate cancer, biomarker identification, imaging biomarker, MALDI mass spectra

1. Introduction
Prostate cancer is one of the leading cancers causing significant numbers of male deaths. It develops most frequently in men over 50 years old and is responsible for more male deaths than any other cancer except lung cancer. It is estimated that in 2008 around 186,320 men in the United States were diagnosed with prostate cancer and about 28,660 died of it.1 Prostate cancer can be cured by a number of therapies if it is detected early, when it is still localized in the prostate. Prostate-specific antigen (PSA) in serum is currently the most widely used approach for early detection. A high-PSA concentration implies a possible high risk of prostate cancer. However, PSA is prostate specific rather than prostate-cancer specific, and increased concentrations of PSA are also found in benign prostatic hyperplasia, acute and chronic prostatitis and prostatic intraepithelial neoplasia. Such misguided diagnoses result in high numbers (up to 75%) of unnecessary prostate biopsies.2 Furthermore, previous studies reported that 15% of men with low-PSA concentrations actually have prostate cancer.3 Therefore, there is a crucial need to develop more effective screening tools for early detection of prostate cancer.

There has been extensive research directed toward the utilization of mass spectrometry (MS) to generate molecular diagnostic and prognostic tools for cancers. Drake, et al. reviewed the current state-of-the-art MS techniques for finding the low molecular weight proteome of blood.4 These mass spectra can be represented as curves where the x-axis indicates the ratio of molecular weight to electrical...
markers identified to date that can be used for early diagnosis or if it enables real-time monitoring of drug response.

Unfortunately, only a few tumor types have had relevant biomarker is detected before clinical symptoms are evident. A trivial case would be to label everything as cancer, thereby achieving 100% sensitivity but 0% specificity. To be useful, biomarkers must have both high sensitivity and high specificity.

As a biological indicator of cancer status and for cellular changes associated with disease progression at specific times, biomarkers represent powerful tools for monitoring the course of cancer and gauging the efficacy and safety of novel therapeutic agents. They can have tremendous therapeutic impact in clinical oncology, especially if the biomarker is highly reliable in distinguishing normal tissue from cancerous tissue (quantified by the percentage of normal tissue identified correctly). A trivial case would be to label everything as cancer, thereby achieving 100% sensitivity but 0% specificity. To be useful, biomarkers must have both high sensitivity and high specificity.

Identification and quantification of biomarkers from those peaks can be done on the adjacent slice. To search for those biomarkers, the MALDI-MS data can be used to visualize protein/peptide concentrations across biopsy samples. Current practice is that a pathologist examines a prostate biopsy slice while the adjacent slice is subjected to the destroying MALDI imaging process (the biopsy slice examined by the pathologist is usually formalin fixed and hematoxylin and eosin (H&E) stained so protein structure is partially changed and, therefore, MALDI analysis has to be done on the adjacent slice). The highest confidence PCa region from the histopathological analysis is then mapped back to the MALDI spectra data to search for biomarkers. However, a detailed histopathological examination may vary with the training, experience and availability of time of the pathologist involved. For example, if an obvious PCa region is identified on the H&E stained biopsy sample, there might be a lack of motivation on the pathologist’s part to carefully examine the true boundary between normal tissue and PCa, which is not obvious to human eyes. Unfortunately, the information conveyed by the ambiguous boundary, especially at low resolutions, might be crucial for biomarker identification based on MALDI-MS data from the adjacent slice.

Since human grading for prostate cancer based on histological images is always time-consuming and subjective, a number of computer-aided diagnosis (CAD) systems have been proposed to automatically analyze pathological images. Tan et al. present an automatic system for cancer diagnosis and Gleason grading of histological images of prostate tissue. They aggregated color, texture and morphometric cues from tissue images to either diagnose prostate cancer or to estimate the Gleason grades, and a better accuracy was achieved for diagnosis than for Gleason grading. Fractal dimension–based features have also been applied to grade prostate pathological images in which feature selection techniques, different supervised classifiers and a leave-one-out validation method had been applied to 205 pathological prostate images, and an accuracy of 94.6% was achieved for the classification. The Gray Level Co-occurrence Matrix (GLCM) is one of the best-known tools for texture analysis and has also found applications in pathological image analysis. Accurate Gleason grading, however, does not lead one down the path to better non-invasive prostate cancer screening tests, and thus there is a need for optimized proteomic biomarker analysis.

We describe a hybrid method to predict the true PCa boundary by fusing the prediction given by a texture-analysis technique based on the adjacent histological biopsy slice image and the prediction given by analyzing the MALDI spectra data. The overall objective of this...
study is to achieve a highly accurate true PCa boundary estimate that is optimized to assist biomarker identification from MALDI spectra data.

2. Methods

The overall system can be divided into two parts: texture analysis and MALDI-MS data processing. For texture analysis, a CAD system is developed on a high-magnification optical image to predict PCa regions on an adjacent tissue slice. The system consists of three steps—image preprocessing, feature extraction and diagnosis. The aim of the preprocessing step is to eliminate the background noise and to improve the image quality for determining the focal areas in the image. In this step, we scanned a prostate tissue sample using a Hirox HI-SCOPE KH-1300 scanner (Figure 1), which is a high-magnification digital microscope. The goal of the feature extraction step is to extract useful texture features that can distinguish normal cells from PCa cells. Those features include the Gray-Level Run Length Matrix (GLRLM)–based features, which have proved to be excellent for texture analysis, and mean and variance of the pixel values. We divided the whole optical image into a set of subimages, and those features were extracted for each subimage. The purpose of the diagnosis step is to classify those subimages into normal or PCa cells based on the extracted features. We utilized a feature-selection algorithm for the identification of the most important features and a multilayer perceptron (MLP) classifier for the classification.

In MALDI-MS data processing, we utilize a three-step method to predict PCa regions using MALDI-MS data from the adjacent slice, based on a high-confidence, low-resolution PCa region marked by a pathologist using the adjacent histological slice as a manual guide. First, we implemented techniques to preprocess the MS data, including background adjustment, smoothing, normalization, peak detection and peak re-binning. This preprocessing step is critical for making the system robust. Second, we applied a feature selection algorithm to identify peaks that are most useful for discriminating normal MS spectra from cancer ones. Finally, MLP classifiers were trained based on the pathologist-defined ground truth to predict PCa regions. Once we obtained PCa region predictions from the texture analysis and the MALDI-MS data processing, we fused those two results as the final optimized PCa regions that will assist MALDI imaging biomarker identification. Performance of the system was evaluated through k-fold cross-validation in terms of sensitivity and specificity.

3. Data preparation

The prostate tissue slides were collected at the Eastern Virginia Medical School (EVMS) from a formalin-fixed and paraffin-embedded prostate biopsy sample. Two adjacent slices approximately six microns thick were then obtained. One slice was prepared for MALDI processing, and the other was assessed by a pathologist and served as

![Figure 1](https://example.com/figure1.jpg)

**Figure 1.** (a) Normal image, (b) Cancer image, (c) A cropped portion of the overall stitched image with highlighted cancer regions.
the ground truth for the normal and PCa regions on the slice. The pathologist also mapped a high-confidence PCa region to the MALDI processed tissue.

3.1 High-magnification image data preparation

The H&E stained sample was scanned using a Hirox HI-SCOPE KH-1300 at a magnification of 402x, yielding a total of 184 pictures (see image examples in Figure 1). Two adjacent pictures had at least 20% overlap. The overall tissue sample image was then reconstructed by stitching the 184 pictures together (Figure 2a) to create a high-magnification optical image. The sample had been rotated and images had been captured as a vertical strip in order to minimize the image size. The image was saved in TIFF format so no information would be lost by file compression. The histopathology tissue image was reconstructed using E-tiling and Adobe Photoshop CS3. The resolution of the histopathology tissue image is 7428×4333×3 pixels. To reduce computational complexity, a subimage covering the cancer regions and some of the normal regions was cropped (Figure 1c, with cancer regions highlighted) for further processing.

3.2 MALDI-MS data preparation

A technique known as MALDI Imaging was used to generate two dimensional molecular maps of hundreds of peptides and proteins directly from the surface of the tissue sample.21,22 A matrix solution, sinapinic acid (SPA), was deposited onto the tissue section, and the soluble proteins were extracted from the tissue and co-crystallized with the matrix. Spectra were collected across the entire tissue area using the Ultraflex III MALDI-TOF/TOF instrument (Bruker Daltonics) with a SmartBeam laser operating at 200 Hz in linear mode over a mass range of 2,000 to 45,000 Daltons. A laser spot diameter of 100 µm and a raster width of 100 µm were employed. Using the FlexImaging software (Bruker Daltonics), teaching points were generated to ensure the correct positioning of the laser for spectral acquisition. The software exports the specific geometry of the tissue to be analyzed, and an instrument-specific automated method is created and generates a grid across the tissue of spots where the laser will acquire data. A total of 200 laser shots were accumulated and averaged from each laser spot rastered across the tissue section. Calibration was performed externally using a peptide standard in the mass range of 700–4500 Da. The laser scanned over the entire tissue, yielding a total of 974 MALDI spectra. Each spectrum corresponds to a single spot about 100 micrometers in diameter on the tissue sample over which the laser was pointed. A mass spectrum can be represented as a curve where the x-axis indicates the m/z values of proteins and the y-axis represents their concentrations. Figure 3a shows one MALDI spectrum from a normal tissue spot, and Figure 3b illustrates one MALDI spectrum from one cancer tissue spot. Since the MALDI imaging was processed on the whole tissue, Figure 3c demonstrates the intensity distribution of one MS peak across the whole tissue using the BioMap software,23 where high concentration is colored as red.

3.3 Ground truth definitions

We have defined two different ground truths: High-Magnification Ground Truth (HMGT) and Low-Magnification Ground Truth (LMGT). The high-resolution optical image from the H&E stained sample was examined by a pathologist to classify the tissue regions into either normal or PCa, yielding the HMGT as highlighted in Figure 2f. Subsequently, an image mask was generated (Figure 2g). The LMGT was obtained by the pathologist after he looked at the H&E stained tissue sample and marked the most confident PCa regions as LMGT. We also formed an artificial image by mapping the data coordinates from the MALDI data into a two dimensional image, as in Figure 2c, where the bright region is the LMGT mapped from the most confident PCa region on the H&E stained sample by the pathologist. Note that the artificial image has a much lower resolution compared to the scanned H&E stained tissue image.

Our purpose is to predict the HMGT based on the LMGT marked by the pathologist in order to provide an optimized PCa region to support biomarker identification. A combined ground truth (Figure 2h) was also generated by intersecting of the LMGT and the HMGT. Before the intersection, a landmark-based registration algorithm (which is described in Section 5.2.1) was utilized to register the high-resolution optical image (Figure 2b) and the low-resolution artificial image (Figure 2c). The histopathology image (Figure 2b) of the tissue was used as the base image, and the MALDI coordinates visualized image was taken as the input image. The selected control points (landmarks) can be seen in Figure 2b and 2c. After the image was registered, the new image was overlaid on the reference image to check for offsets (Figure 2e). As they are not from the same slice, they do show some differences.
among them. The registered image can be seen in Figure 2d, which was used as the LMGT for texture analysis.

4. PCa region prediction based on texture analysis

In texture analysis, we first divided the whole optical image into a set of small subimages from which the run length matrix (RLM)- and co-occurrence matrix (CM)-based features were extracted. Those features then went through a feature selection step, and the selected features were utilized to classify the subimages as normal or cancer. Figure 4 shows the system diagram for texture analysis. Texture is a connected set of pixels that occurs repeatedly in an image. It provides information about the variation in the intensity of a surface by quantifying those properties. Gray Level RLM (GL-RLM) and Gray Level CM (GL-CM) are the two most widely used models in textural features. In our application, GL-RLM outperformed GL-CM so we report on only those RLM results.

4.1 Feature extraction

The GL-RLM method was proposed by Galloway in 1975.20 GL-RLM extracts higher order statistical texture features and provides a quantitative description of the texture features in a given image. By definition, consecutive pixels with the same gray level in a given direction constitute a gray level run, and the number of pixels in the run is called run length. For a given image, we can compute its GL-RLM along a given direction. Each element \( p(i, j) \) in the GL-RLM represents the total number of runs with a pixel gray level of \( i \) and run length of \( j \) in the whole images. For an image with \( n_r \) pixels, the size of the run-length matrix will be \( M \) by \( N \), where \( M \) and \( N \) are the maximum gray level and the possible maximum run length in the corresponding image, respectively.

In this research, a RLM feature matrix was generated for each subimage by summing the four angles (0°, 45°, 90° and 135°) of the directional RLM feature matrices together for a global view. To correctly classify each tissue area, the size of the subimages \( n_r \) should be chosen carefully. Too large an area may contain both cancerous and normal cells, while too small an area may contain insufficient information. The size of subimages used for this work is 200*200 pixels. For each subimage, 13 features are extracted, including 11 GL-RLM features and mean and standard deviation.

As indicated by its definition, relatively long gray-level runs would occur more often in a coarse texture, while primarily short runs would describe fine textures. From the run-length matrix, various texture features can be derived.

![Figure 3](image-url)
The following are 11 descriptors typically extracted from the run-length matrices: short run emphasis (SRE), long run emphasis (LRE), high gray-level run emphasis (HGRE), low gray-level run emphasis (LGRE), pair-wise combinations of the length and gray-level emphasis (SRLGE, SRHGE, LRLGE and LRHGE), run-length non-uniformity (RLNU), gray-level non-uniformity (GLNU) and run percentage (RPC). Formulas for calculating these feature descriptors are listed as follows:

**Short Run Emphasis**  
$$SRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{j^2}$$

**Long Run Emphasis**  
$$LRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \ast j^2$$

**Low Gray-Level Run Emphasis**  
$$LGRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2}$$

**High Gray-Level Run Emphasis**  
$$HGRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \ast i^2 \ast j^2$$

**Short Run Low Gray-Level Emphasis**  
$$SRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \ast i^2}{j^2}$$

**Short Run High Gray-Level Emphasis**  
$$SRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \ast j^2}{i^2}$$

**Long Run Low Gray-Level Emphasis**  
$$LRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \ast i^2}{j^2}$$

**Long Run High Gray-Level Emphasis**  
$$LRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \ast i^2 \ast j^2$$

**Gray-Level Non-uniformity**  
$$GLNU = \frac{1}{n_r} \sum_{i=1}^{M} \left( \sum_{j=1}^{N} p(i,j) \right)^2$$

**Run Length Non-uniformity**  
$$RLNU = \frac{1}{n_r} \sum_{j=1}^{N} \left( \sum_{i=1}^{M} p(i,j) \right)^2$$

**Run Percentage**  
$$RPC = \frac{n_r}{M+N}$$

### 4.2 Feature selection

Feature selection was performed by an advanced feature-selection algorithm that was chosen because of its computational efficiency and effectiveness. The algorithm performs feature selection by generating an appropriate piecewise linear network (PLN) model for the given data set, applying the orthonormal least squares (OLS) procedure to the PLN model, then selecting useful feature subsets through a floating search algorithm.

This algorithm consists of three important components: a piecewise linear classifier, an OLS procedure and a floating search algorithm. It selects $N_s$ features from the available $N$ ones by performing the following steps:

1. Determines an appropriate number of clusters, $N_c$, to be used by the PLN model using the trial-and-error method.
2. Solves a set of linear equations and designs an $N_c$-cluster PLN classifier for each cluster.
3. Performs the OLS procedure for each cluster.
4. Searches for good feature combinations using the floating search algorithm.
4.3 Multi-layer perceptron and support vector machine classifiers

A MLP, one of the most commonly used neural network classifiers, was used for classification in this paper. The three-layer MLP (input-hidden-output) has proved to be capable of learning a rich variety of nonlinear decision surfaces and was chosen in the texture-analysis experiments. The sigmoid and linear functions were used as activation functions in the hidden layer and the output layer, respectively.

Another common classifier used, a support vector machine (SVM), defines a classification based on the structural risk minimization (SRM) concept. Given a set of training data pairs \( \{x_p, i_p\}_{p=1}^{N} \), where \( x_p \in R^N \) is the feature vector extracted from one subimage, \( i_p \in \{+1, -1\} \) is a class label (PCa, normal) associated with \( x_p \). An SVM defines a hyperplane, \( f(x) = w^{T} \phi(x) + b = 0 \), to separate the data points onto two classes, where \( w \) and \( b \) are the plane parameters, and \( \phi(x) \) is a function mapping the vector \( x \) to a higher dimensional space. After the hyperplane is determined, a subimage is declared as PCa if \( f(x_p) > 0 \), otherwise a normal subimage is declared. In order to combine the outputs from texture analysis and MALDI-MS data processing, we utilize a sigmoid function for both MLP and SVM to transfer their outputs to a posteriori probabilities.

The purpose of this study is to predict the HMGT regions (not always available) based on the LMGT regions (usually easy to define) by fusing the predictions from texture analysis and MALDI-MS data processing. Eventually, we hope the predicted HMGT will assist prostate cancer biomarker identification.

4.4 Cross-validation and performance metrics

To evaluate the performance of the proposed system, a three-fold cross-validation (CV) was performed on the available data in some of our experiments. We divided the available dataset into three parts. We left one part for testing, and the remaining parts were used for training. This procedure was repeated three times so that each part was used for testing once. Sensitivity and specificity were then computed based on the testing results.

Performance of the proposed system was evaluated in terms of sensitivities and specificities. The sensitivity is a measure of the ratio of correctly classified PCa subimages or MALDI spectra, while the specificity denotes the ratio of correctly identified normal subimages or MALDI spectra.

5. PCa region prediction based on MALDI-MS data analysis

We first registered the ground truth information (HMGT) back to MALDI-MS data samples resulting in a ground truth mask, on which a similar experiment, i.e., predicting the HMGT from the LMGT using MALDI-MS data, was based on and performed. We then had the HMGT PCa regions prediction from two distinct sources, and finally those two predictions were fused to produce a more accurate PCa region estimate.

5.1 Image registration

The Hirox scanned image with annotated PCa regions (HMGT) 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 (Figure 5). The images obtained from the MALDI co-ordinates visualization and histopathological examinations are very much different from one another in size and orientation. For any further studies to be done on these images, they must be of the same size and orientation. Therefore, it is necessary to adopt a registration process. In landmark-based registration, a control point selection tool can be launched for selecting a set of control points (landmarks) from both input images. Those control points will determine a transformation matrix based on their geometrical relationship. There are various transformation matrix types available, such as piecewise linear, affine and projective. In this paper, we used the affine transformation for the registration. Figure 5 shows the results of registering the optical image back to the artificial image constructed from the MALDI coordinates. Note that the mapped ground truth in MALDI data space in Figure 5e is usually not available for proteomic analyses. The purpose of this project is to predict the ground truth shown in Figure 5e using the available ground truth shown in Figure 5b that is usually easy to obtain.

5.2 MALDI spectra data processing

The MALDI-MS data preprocessing consists of several steps, including (1) baseline adjustment, (2) smoothing, (3) normalization, (4) peak detection and (5) re-binning. After the re-binning step, there are 821 peaks resulting from the 974 MALDI-MS spectra available for this study. In the following subsections, we will briefly introduce each of the processing steps.
5.2.1 Baseline adjustment. This step is required to remove the ion overload and chemical noise that are usually higher at smaller m/z values. There is no general solution to this problem because baseline characteristics vary from one experiment to another and each spectrum has to be assessed individually. The baseline was estimated within multiple shifted windows of width 200 m/z, and a spline approximation was used to regress the varying baseline. The result of baseline adjustment is shown in Figure 6a, where the blue lines show the original spectra, and the regressed baseline is shown in red.

5.2.2 Smoothing. We used a wavelet-based algorithm with the Haar basis function\textsuperscript{26} to denoise and thus enhance the signal-to-noise ratio of each spectrum. We also discarded m/z values lower than 3000 due to large noises and m/z values greater than 10,000 due to low intensities. The result from smoothing is shown in Figure 6b.

5.2.3 Normalization. There are often systematic variations among spectra due to variations in the amounts of protein or variations in detector sensitivities. A global normalization procedure\textsuperscript{27} was used, where mass intensities for the same peak from different spectra were scaled (divided) by a common factor. For a given peak, the area under the peak was computed as the sum of all intensities for this peak from all spectra. The common factor for each peak was then defined as the ratio of the area under this peak to the median of areas of all the other peaks in a single spectrum.

5.2.4 Peak detection. A crucial step for the identification and quantification of proteins in mass spectra is to find m/z values that correspond to high peak intensities. We used an algorithm\textsuperscript{28} to detect peaks corresponding to m/z value points with reasonable intensity and signal-to-noise (S/N) ratio. To eliminate interference from noise, peaks with intensity values below 10 were zeroed. After smoothing and peak detection were performed, a total of 75,719 peaks from the available 974 spectra were obtained. Each spectrum represents the protein profile of one tissue spot in the prostate biopsy sample. The result after smoothing, normalization and peak detection for one spectrum is shown in Figure 7a, and its zoomed-in version is shown in Figure 7b. The original spectrum in blue is unclear because of noise. The denoised spectrum is shown in green, and the detected peaks are shown in red cross-marks.

5.2.5 Re-binning. The same protein/peak may have slightly different m/z values in different spectra due to the fact that the mass spectra exhibit shifts along the horizontal axis between multiple spectra and instruments have a small error on the m/z scale. Thus, detected peaks that have m/z values within a certain range should be considered as identical. To this end, we aligned or clustered the same protein/peak detected from different spectra. We first projected peaks obtained from the available 974 spectra onto a single axis. We then merged peaks that have m/z values within 0.13% of each other and assigned the new peak the average m/z value in the range. After this step, we obtained a total of 821 aligned peaks representing different proteins. Finally, we projected back these peaks onto individual spectra and used them to identify biomarkers. If the individual spectrum consists of a peak that corresponds to a cluster, the peak’s intensity was maintained. If the spectrum contains no such peak, then the intensity for the peak was replaced by zero. Figure 8 shows one re-binned peak where all blue peaks were re-binned to the red peak’s m/z value.

Figure 6 (a). Baseline adjustment, (b). Result of smoothing.
After all the raw data were preprocessed, we obtained a set of peaks that were denoted as \( \{x_p, i_p\}_{p=1}^{N} \), where \( x_p \in \mathbb{R}^N \) and \( i_p \in \mathbb{R} \), \( x_p \) corresponds to a vector consisting of peaks detected from \( p \)th spectrum and \( i_p \) is the class ID (1. normal, 2. cancer) associated with this spectrum. The total number of peaks from each spectrum is \( N \) (821 in our study), and the total number of spectra is \( N_v \) (974 in this paper). The results from the pathological analysis determined the class ID and were considered as the ground truth. Note that we have different resolution ground truths so that a spectrum might have a different class ID based on which ground truth we consider.

5.3 Peak selection and spectra classification

Using all of the 821 peaks for spectra classification will have the curse of dimensionality problem. We again utilized a feature-selection algorithm\(^5,6\) to select a compact set of peaks for classification in the three experiments described later. The feature-selection algorithm chose the best feature or the best combination of features that correlates to cancer. To further improve computational efficiency, the Fisher criterion was used to reduce the 821 peaks to 30 before feature selection. All classification tasks in the experiments using MALDI spectra data were performed by using the SVM implemented by Matlab with default parameters. SVM tries to maximize the margin that is defined to be the smallest distance between the decision boundary and any of the support vectors.\(^24\)

6. Prediction fusion

We propose a technique to optimize the predictions by fusing the results from the texture analysis with the predictions based on MALDI spectra data. Both methods will produce a probability map for PCa regions. After registering the probability map from the texture-analysis method to that obtained from the MALDI spectra data, we averaged the two probability maps to obtain the overall PCa prediction. The 200*200 block used in the texture analysis corresponds to a single spot (spectrum) in the MALDI. The classification probabilities from the two techniques were averaged to get a new probability. If the fused probability was greater than 0.5, it classified the region as PCa (normal).

7. Experiment design

We performed three experiments on our data set. For the MALDI-MS spectra–based method, the spectra data went through the preprocessing steps before performing any of the experiments. In all of our experiments, the classifiers were firstly trained using a portion of the selected PCa
ground truth (either HMG, LMGT or intersection) and then tested on the rest of the data for classification using a cross-validation technique.

Experiment A served as a benchmark for our prediction results. In Experiment A, PCa MALDI spectra for training were taken from the intersection of the ground truth as shown in Figure 5f, and normal 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 the HMG in Figure 5e to obtain testing classification accuracy. For the texture analysis–based method, we used the same intersected ground truth and the normal samples from the same location for training. However, all the processing steps were based on the optical image as shown in Figure 5a. Once the predictions from both methods were available, we fused the results to form an overall PCa prediction. Experiment A was performed to ensure that our model was first trained with our most confident PCa ground truth dataset, which combined the HMG from the optical data and the LMGT from the MALDI data and then determined the intersection of the two.

In Experiment B, PCa MALDI spectra for training were taken from the PCa areas in the LMGT as shown in Figure 5b, and normal MALDI spectra for training were obtained in a similar way as in Experiment A. The training data contained 27 cancer and 27 normal spectra. After feature selection and classification, a trained model was applied to the HMG in Figure 5e to get testing performances. We also applied the texture analysis method in the same manner (trained on LMGT and tested on HMG) using the same training and testing data samples but from the optical image and fused the two predictions as our final prediction for PCa regions. Experiment B served to represent what can be realistically put into practical use. To be successful, we need to predict the HMG region solely from the LMGT region that is routinely identified by pathologists.

In Experiment C, we performed a three-fold CV evaluation using the pathologist annotated PCa and normal regions in the HMG as ground truth (Figure 5e). There were 51 PCa and 923 normal spectra in the data set. In the three-fold CV procedure, data was partitioned into three parts: two 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. The texture analysis–based method and the fusion step were again performed in a similar way. It is reasonable for Experiment C to have the highest sensitivity in the optical data since the model would be trained on the HMG that was identified by the pathologist using the optical image and then tested on the same image. However, since the MALDI data comes from an adjacent slice, the highest sensitivity is not assured. Experiment C serves to provide information about what may be occurring on the adjacent MALDI slice. For example, higher sensitivities would indicate that the true PCa region is likely to be bigger than the LMGT, while lower specificities would indicate that we were misidentifying normal spots as cancer spots.

8. Results

For the texture-analysis method as shown in Table 1, Experiment A yielded a sensitivity of 83.92% and a specificity of 81.25%. Since using only the intersection of ground truths in the optical data means leaving out significant portions of PCa in the training, one would expect the sensitivity to be lower here and highest in Experiment C where we are training and testing on the same datasets. As expected, Experiment C gave the highest sensitivity of the optical texture analysis of 95.29% and a specificity of 72.03%. Experiment B, where we tried to find the high-magnification ground truth from the given low-magnification ground truth, gave a sensitivity of 87.45% but a very low specificity of 75%. Since there are many more normal points than PCa points, this gives a poor accuracy of 77.65%.

Table 2 lists the MALDI spectra analysis results in the three experiments. The feature selection algorithm selected five, five and three peaks for classification in Experiments A, B and C, respectively. In contrast to the texture analysis, the MALDI results exhibited poor sensitivity but very high specificity. This showed that the picked ground truth MALDI regions were not enough to distinguish PCa perhaps from misalignment of high-confidence regions and changing morphology of the tumor on the adjacent slice. In Experiment A, we achieved a sensitivity of 51.02% and a specificity of 99.67%. In Experiment B, we obtained a sensitivity of 50.98% and a specificity of 100% in Experiment.
Table 3. Results from fusion.

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>84.31%</td>
<td>87.83%</td>
<td>87.32%</td>
</tr>
<tr>
<td>B</td>
<td>80.39%</td>
<td>93.09%</td>
<td>91.26%</td>
</tr>
<tr>
<td>C</td>
<td>88.29%</td>
<td>85.13%</td>
<td>85.64%</td>
</tr>
</tbody>
</table>

Figure 9. Results from left to right: (a) Result from MALDI in Experiment A, (b) result from MALDI in Experiment B, (c) result from MALDI in Experiment C, (d) result from fusion in Experiment A, (e) result from fusion in Experiment B, (f) result from fusion in Experiment C.

C, a sensitivity of 58.82% and a specificity of 96.82% were achieved. The highest sensitivity in Experiment C seems to indicate that expanding the ground truth region by mapping the HMGIT area enhanced the reliability of the model by including more PCA information.

As mentioned, we observed that the texture-analysis method gave relatively high sensitivities but lower specificities, while the MALDI spectra-based method obtained higher specificities with lower sensitivities. Those results are highly desired because we can achieve both high sensitivities and specificities if the results from the two methods are fused. Combining the probabilities from each area of the same spots on the optical image and the MALDI data proved this to be true, as shown in Table 3, where we achieved a sensitivity of 84.31% and a specificity of 87.83% in Experiment A. Experiment B yielded a sensitivity of 80.39% and a specificity of 93.09%. A sensitivity of 88.29% and a specificity of 85.13% were obtained for Experiment C. Figure 9 shows the fused prediction results on the MALDI tissue image constructed from the spectra coordinates. As we had hoped, the fusion of both MALDI and optical texture analysis resulted in the best combined sensitivity and specificity, and of the three fused results, Experiment B achieved the highest accuracy.

9. Discussion

Although biomarker research has progressed considerably in recent years, the practical impact of this research on cancer screening, diagnosis and prognosis remains limited.9,10,29,30 Recent projects initialized by the Early Detection Research Network (EDRN) for validating previously identified biomarkers showed that some serum biomarkers do not reliably detect prostate cancer.31,32 The projects consisted of two stages to evaluate previously published EDRN studies for prostate cancer detection based on the surface-enhanced laser desorption/ionization (SELDI) time-of-flight (TOF) MS data.8,31,32 The validation experiments failed to separate patients with prostate cancer from biopsy-negative controls. Nor did the experiments separate patients with prostate cancer with Gleason scores < 7 from those with Gleason scores ≥ 7. The overall accuracy was close to what it would have been if all samples were classified as prostate cancer. Possible reasons for the validation failure could be due to fact that the SELDI-TOF MS whole serum proteomic profiling might not be powerful enough to reliably detect prostate cancer32 or that the precision of the SELDI-TOF MS data might not be high enough.33 To address those challenges, prostate biomarker identification based on a recently developed technique, MALDI-MSI of prostate biopsy tissue samples, attracted a significant amount of attention.13-16

Identifying prostate biomarkers based on MALDI-MS tissue data requires an accurate definition of PCs on MALDI-MS coordinates that is a time-consuming and challenging task. To precisely classify cells into normal or PCs, pathologists need to put the adjacent biopsy slice under a high-magnification microscope and carefully mark the PCs regions. Those regions are then manually mapped back to the MALDI-MS coordinates for subsequent biomarker identification. This detailed examination, however, may vary with pathologists’ experiences, level of fatigue and availability of time. Instead, pathologists often examine the biopsy sample and simply map the highest confidence PCA region back to the coordinates for biomarker identification. The highest confidence PCA region usually is much smaller than the true PCA region and information conveyed near the true boundary will be lost. Therefore, a systematic, automatic system that is capable of accurately defining true PCA region on biopsy samples and is able to map them back to MALDI-MS coordinates for biomarker identification is highly desirable.

We have developed a system for predicting an optimized true PCA boundary on prostate biopsy samples based on the high-confidence PCA region defined by a pathologist. Experiments show that the prediction task cannot be satisfactorily fulfilled by either texture analysis or MALDI-MS data processing. The texture-analysis results are good in sensitivities and relatively poor in the specificities, while The MALDI-MS processing results are low in sensitivities and high in specificities. The natures of MALDI-MS data and optical images are quite different and complementary. Optical images perform well in recognizing PCa regions, while MALDI-MS data works perfectly in identifying normal regions on biopsy samples. To design an efficient model that can predict the true PCA boundary, we fused results from both the texture analysis and the MALDI-MS processing. The specimen tissues used in the two methods are adjacent slices but were not identical, and landmark registrations had been performed...
to ensure the similarity. Our experiments show that it is feasible to reasonably predict true PCa regions from the highest confidence, low-resolution PCa regions.

Our future work includes applying the proposed system to more datasets with true PCa regions identified by the pathologist and further validating the method for biomarker identification. We will compare prostate biomarkers identified based on true PCa regions marked by the pathologist with those found based on the predicted PCa regions. Statistical analysis will be performed to test whether statistical differences exist between the two set of biomarkers.

10. Conclusions

We have developed PCa region prediction models by fusing results from a texture-analysis technique and predictions from MALDI spectra processing. The designed models are able to predict a high-resolution PCa region based on a low-resolution PCa area defined on the adjacent biopsy slice. This prediction methodology can benefit prostate cancer biomarker identification using MALDI imaging techniques by providing a more optimized and realistic indication of where PCa regions are located, including those regions too difficult to identify without painstaking and time-intensive histological analysis and impossible to identify on the non-histological processed adjacent slice subjected to MALDI proteomic processing. Such regions may also contain useful biomarker information, especially at the PCa borders.

Funding

This work is supported by the Office of Research, Old Dominion University, Norfolk, VA 23529, USA.

Acknowledgment

A short version of this paper appeared in Li et al.84

References


**Author biographies**

**Shao-Hui Chuang** received BS and MS degrees in electrical engineering from Old Dominion University (ODU), Virginia, in 2008 and 2009, respectively. She is currently a PhD student with the Medical Imaging, Diagnosis and Analysis (MIDA) Laboratory at ODU. Her current research interests include medical image analysis and texture recognition.

**Jiang Li** received his BS degree in electrical engineering from Shanghai Jiaotong University, China, in 1992, his MS degree in automation from Tsinghua University, China, in 2000, and his PhD degree in electrical engineering from the University of Texas at Arlington, Texas, in 2004. His research interests include machine learning, computer-aided medical diagnosis systems, medical signal/image processing and modeling and simulation. He is also affiliated with Virginia Modeling, Analysis, and Simulation Center (VMASC). Dr. Li worked as a postdoctoral fellow at the department of radiology, National Institutes of Health, from 2004 to 2006. He joined ODU as an assistant professor in 2007. He is a member of the IEEE and Sigma Xi.

**Xiaoyan Sun** received her BS degree in biomedical engineering from Zhejiang University, China, in 2004, and MS degree in biomedical engineering from Peking University, China, in 2007. She is currently a PhD candidate with the MIDA laboratory at ODU. Her research interests focus on computer-aided and robot-assisted technologies in medical applications, including image-guided robotic dental implantation, medical image processing and augmented reality.

**Ayyappa Vadlamudi** received his BTech degree in electrical engineering from the Acharya Nagarjuna University, India, in 2008, and his MS degree in electrical engineering from ODU in 2010, respectively. His research interests include machine learning and medical image analysis.

**Bo Sun** received her BS degree in computer science from Wuhan University, China, in 2001, her MS degree in computer science from Lamar University, Texas, in 2004, and her PhD degree in modeling and simulation from ODU, Virginia, in 2008. Her research interests include medical simulation and medical image processing. She was an assistant professor in computer science at University of the Sciences in Philadelphia from 2008 to 2011. Dr. Sun joined Lincoln University as an assistant professor in 2011.

**Lisa Cazares** received her PhD degree in microbiology and molecular cell biology from the Eastern Virginia Medical School (EVMS) in 2009. She is now a post-doctoral fellow in the Leroy T. Canoles Jr. Cancer Research Center at EVMS. Her research interests focus on the early detection of cancer.

**Julius Nyalwidhe** received his BS degree in biology from Kenyatta University, Nairobi, Kenya, in 1987, his MS degree in medical parasitology from University of Nairobi, Nairobi, Kenya, in 1992 and his PhD degree in molecular parasitology & proteomics from Philipps University, Marburg, Germany, in 2003. His research interests focus on the early detection of cancer.

**Dean Troyer** received his BA degree in chemistry from the Earlham College, Virginia, in 1973 and his MD degree in medicine from Indiana University School of Medicine, in 1978. He completed his residency training from 1978 to 1982 in pathology at the University of Texas Health Science Center at San Antonio. Dr. Troyer is a professor in the department of pathology and the department of molecular biology at EVMS. His research interest focuses on the translational research in prostate and genitourinary cancers.

**John Semmes** received his BS degree in biology from Auburn University in 1980, his MS degree in marine biology from University of West Florida in 1984 and his PhD degree in biochemistry from George Washington University in 1990. He is a professor in the department of microbiology and molecular cell biology at EVMS. Dr. Semmes directs a multi-disciplinary team of investigators with the long-term goal of the development and application of clinical proteomics techniques to the early detection of cancer. This program brings together a unique group of individuals with the experience and expertise needed to achieve the objectives of biomarker discovery within existing national and international collaborative efforts, most notably the National Cancer Institute Early Detection Research Network.

**Frederic (Rick) McKenzie** received his PhD in computer engineering from the University of Central Florida in 1994. He is a professor in the department of modeling, simulation, & visualization engineering and a joint faculty member in the electrical and computer engineering department at ODU. Prior to joining ODU, Dr. McKenzie held a senior scientist position at Science Applications International Corporation (SAIC).