DEPENDENCE MODEL AND NETWORK FOR BIOMARKER IDENTIFICATION AND CANCER CLASSIFICATION

Peng Qiu[1]*, Z. Jane Wang*[2]*, and K.J. Ray Liu*[1]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, USA
[2]Department of Electrical and Computer Engineering, University of British Columbia, Canada
email: qiupeng@umd.edu, zjanew@ece.ubc.edu, kjrliu@umd.edu

ABSTRACT

Of particular interest in this paper is to develop statistical and modeling approaches for protein biomarker discovery to provide new insights into the early detection and diagnosis of cancer, based on mass spectrometry (MS) data. In this paper, we propose to employ an ensemble dependence model (EDM)-based framework for cancer classification, protein dependence network reconstruction, and further for biomarker identification. The dependency revealed by the EDM reflects the functional relationships between MS peaks and thus provides some insights into the underlying cancer development mechanism. The EDM-based classification scheme is applied to real cancer MS datasets, and provides superior performance for cancer classification when compared with the popular Support Vector Machine algorithm. From the eigenvalue pattern of the dependence model, the dependence networks are constructed to identify cancer biomarkers. Furthermore, for the purpose of comparison, a classification-performance-based biomarker identification criterion is examined. The dependence-network-based biomarkers show much greater consistency in cross validation. Therefore, the proposed dependence-network-based scheme is promising for use as a cancer diagnostic classifier and predictor.

1. INTRODUCTION

In genomics studies, great efforts have been made to develop the gene regulatory network using microarray gene expression data [17]. Recently, it is believed that it is the proteomic data and the collective functions of proteins that directly dictate the phenotype of the cell and, thus, are more accurate in interpreting the cause of biological phenomenon. Many changes in gene expression might not be reflected at the level of protein expression or function [1]. Therefore, *proteomics*, is an emerging field for the discovery and characterization of regulated proteins or biomarkers in different diseases in the post-genome era. During cancer development, the cancerous cells may release unique proteins and other molecules, which may be regarded as early biomarkers. These biomarkers normally serve as the indicators of diseases. Correctly identification of protein biomarkers for cancer holds enormous potential for the early detection of cancer and effective treatments. However, due to the complicate nature of protein functions, it is a research topic with significant challenge.

For the analysis of protein samples, mass spectrometry (MS) technologies have become increasingly important tools [2]. MS is able to convert proteins or peptides to charged pieces that can be separated on the basis of the mass-to-charge ratio (m/z) and their abundances. There are several types of MS ionization methods currently available, including surface enhanced laser desorption ionization (SELDI), electrospray ionization (ESI), and matrix-assisted laser desorption ionization (MALDI) [8]. The produced protein or peptide spectra are then analyzed for different purposes, such as identifying proteins via peptide mass fingerprints, cancer classification, etc. Until very recently, it has also been applied for cancer biomarker identification, but only simple classification-based approaches were studied. For instance, in [4], a panel of three biomarkers were selected using the linear combination based Unified Maximum Separability Analysis (UMSA) to best separate cancer and non-cancer samples.

In [9], we developed an ensemble dependence model (EDM)-based approach for cancer classification based on microarray gene expression data. The proposed method yields promising performance in gene expression data. To further explore the EDM concept, we apply it on proteomic data and then present the idea of building dependence networks based on MS data. The dependency revealed by the dependence model provides some insight into the functional interaction relationships between proteins. This paper is organized as follows. In Section 2, we present the classification results for protein MS data based on EDM, and present the proposed dependence network idea. Then, in Section 3, the classification-performance-based biomarkers and dependence-network-based biomarkers are examined. Finally, the conclusions are presented in Section 4.

2. DEPENDENCE MODEL AND DEPENDENCE NETWORK

As mentioned earlier, the concept of ensemble dependence model (EDM) for cancer classification is proposed in [9]. In Section 2.1, we modify and apply the EDM concept for classification of proteomic data. The classification performance on two apply public-available protein MS datasets is reported. In Section 2.2, we will focus on the idea of dependence network.

2.1 Dependence Model for Cancer Classification

The dependence model focuses on exploring and modeling the group dependence relationship. Because of the limited sample size of current MS data, it is not feasible to examine the dependence relationship among all mass features at one time. In the proposed ensemble dependence model, features are clustered into several clusters. Given appropriate and well-sorted clustering results, we predict that proteins' group behaviors and ensemble dynamics can be revealed. In this study, the Gaussian Mixture Model [15] is applied for

	Classification ovarian dataset	Classification prostate dataset normal vs early stage cancer	Classification prostate dataset normal vs late stage cancer
EDM-2	100%	99.39%	100%
EDM-3	100%	100%	100%
EDM-4	100%	98.18%	100%
EDM-5	96.60%	98.79%	99.39%
SVM	96.83%	78.79%	98.79%

Table 1: Correct classification rates for two MS datasets when applying the proposed ensemble dependence model with different choices of cluster number. The number after "EDM-" refers to the number of clusters

feature clustering. After clustering, each cluster contains specific features that have a well-defined mathematical relationship to one another. For each cluster, an interesting problem is how to effectively represent each cluster's profile. A most straightforward way could be using the average of all features within one cluster to represent the cluster. In this paper, due to the specific properties of the protein MS data, we propose a concept of *virtual protein*, where virtual protein is illustrated by a linear weighted combination of different MS features within a cluster. In order to represent each cluster, a virtual protein is generated as the cluster representative.

We argue that a virtual protein representation makes more sense than a straightforward averaging, for two main reasons. First, in mass spectrum data, some features correspond to high intensity peaks, while some features correspond to low intensity peaks. In order to avoid high intensity features dominating its cluster, the virtual protein generated by the weighted average expression of cluster members can provide better information to the entire cluster. Secondly, for protein samples, mass spectrometry measures the mass-to-charge ratio of the ionized peptides and their abundances in the sample. Due to the measurement process of MS, one particular cancer-related protein can be represented by several peptides. A linear combination of MS features may lead to a virtual protein which better represents the underlying cancer-related protein. Another important question is how to represent a virtual protein, i.e. determining the weights. In our approach, the weights are determined through linear discriminant analysis (LDA) [11]. Since we are interested in the virtual proteins which are cancer-related and thus best represent the difference between a cancer and non-cancer sample, LDA provides an efficient way to construct a virtual protein. Given the virtual proteins as cluster representatives, the ensemble dependence model in [9] is applied for classification.

In this study, there are two mass spectrum datasets under investigation, one ovarian cancer dataset, with 91 normal samples and 161 cancer samples [6], and one prostate cancer dataset, with 81 normal samples, 84 early stage cancer samples and 84 late stage cancer samples [12]. Raw mass spectra are downloaded from the National Cancer Institute, and Eastern Virginia Medical School. Preprocessing is performed, including smoothing, baseline correction, peak alignment and peak detection, similar as in [13]. 50 top mass peaks are obtained by the criterion proposed in [14]. All analysis are based on the 50 selected peak features.

The ensemble dependence model is applied to classify cancer and normal data through leave-one-out cross-validation [16]. In the proposed model, an unmentioned problem is how to choose the number of clusters. The optimal number of clusters is difficult to determine, because it may depend on different diseases and different sets of examined features. To examine this parameter, we apply different choices to the proposed model and compare the overall classification performance, as shown in Table 1. Theoretically, as the number of clusters increase, more dependence relationship is examined, thus better classification performance will be achieved. However, in Table 1, as the number of clusters increase, the classification performance first increases and then decreases. Because, when more dependence relationship is examined, the number of model parameters also increases quadratically. The maximum number of clusters is limited by the size of the training dataset. From Table 1, we can see that the proposed model yields good classification performance. In order to examine the performance of the proposed model, we compare it with the widely-applied linear support vector machine (SVM) approach. It has been applied in bioinformatics studies [10], where it is illustrated that SVM provides excellent classification performance. To ensure a fair comparison, both SVM and the proposed model are applied to the top 50 features. From Table 1, we can see that in the ovarian cancer dataset, the proposed model and SVM have comparable performance. In the prostate cancer dataset, when we classify normal samples against late stage cancer samples, the two schemes also performs comparably. However, in the prostate cancer dataset, when we classify normal samples against early stage cancer samples, where the classification task appears to be more difficult, the proposed ensemble dependence model out performs SVM.

2.2 Dependence Network

The functionality of a protein is not solely characterized by its own structure. Its surroundings and interacting proteins also play important roles in determining the protein's function. In this study, we propose to apply the dependence model for protein dependence network construction. In the previous subsection, the concept of ensemble dependence model was applied on virtual proteins, the representatives of clusters. Now, we use the dependence model to examine individual protein mass features, zooming in to build a network that captures interactions among proteins.

A dependence network is a set of components, such as MS peaks in our case, and linear dependence interactions among them that collectively carry out specific functions, where each arrow represents an inter-component dependence relationship with an associated weight a_{ij} indicating to what extent component i depends on component j. In the following, we describe how a dependence network is constructed.

In [9] it is shown that the eigenvalue pattern is closely related to dependence relationship, especially the smallest eigenvalue. Take a three-feature case for example. From the noise-free ideal case, as the three features' expression profiles more and more independent, the eigenvalues of their dependence matrix will change and follow the trends, as shown in Fig.1. In the three-features example, when the expression profiles are dependent, the smallest eigenvalue is -2. When the dependence relationship become weaker and weaker, the smallest eigenvalue increases, and eventually saturate to around -0.7. Thus, for any feature triple, by examining the eigenvalue pattern of their dependence matrix, we are able to tell how dependent they are, how closely related they are.

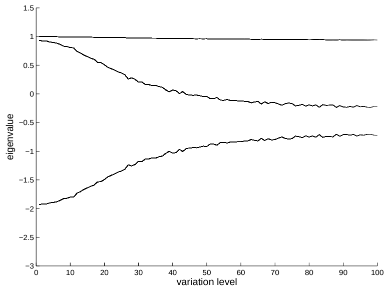
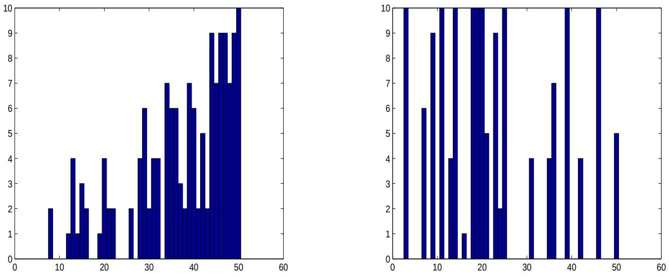


Figure 1: The horizontal axis is variation level, which indicates how noisy the three cluster expression profiles are. As the cluster expression profiles become more noisy, the eigenvalues of the corresponding dependence matrix will change, following the above curves.

Since, the eigenvalue pattern can serve as an indicator of how closely related they are, if we examine three individual MS features at one time, through an exhaustive search, we can find all closely related feature triples. The elements in each triple share a strong dependence relationship, which indicates that they have a strong influence on each other in the protein interaction network. Take the ovarian cancer dataset as an example. For the normal case, we exhaustively examine the eigenvalue pattern for all possible feature triples. A threshold -1.5 is applied. If the smallest eigenvalue of a feature triple is lower than the threshold, there exists a strong dependence relationship within the triple, which is called the "binding triple". Similar analysis is applied to cancer samples. In the normal case, 520 triples pass the threshold; while in the cancer case, 269 triples pass the threshold. Moreover, there are only 80 triples in the overlap between normal and cancer cases. The results suggest that, from healthy to cancerous, some dependence relationships among proteins are disabled; while some other dependence relationships are activated. The small overlap indicates that, from healthy to cancerous, the overall dependence relationship goes through a major change.

The dependence network is constructed from binding triples. As in graph theory, the topology of an n-node network can be represented by an $n \times n$ adjacency matrix D. If feature i and feature j both appear in a binding triple, it is suggested by the dependence model that feature i and feature j are closely related. And we will count once for D_{ij}, the connection between feature i and feature j. Basically, we count the appearance of all feature pairs, and form an adjacency matrix D. Then, the adjacency matrix D is normalized by the total number of binding triples. Each element D_{ij} is a confidence value, which indicates the importance and strength of the connection between feature i and feature j. We call network D the dependence network. Making use of this information, the dependence networks can be presented as shown in Fig.3, where strong dependence relationship is reflected in small distance between connected nodes. The length of each connection is defined to be inverse proportional to the confidence value. Because the confidence values are normalized, through $1/D_{ij}$, features with strong dependence relationship will locate close to each other, while features with weak dependence relationship will be far apart. From Fig.3, we are able to see the importance of each node and identify potential biomarkers.



(a)

(b)

Figure 2: Fig (a) is the histogram of performance based potential biomarkers in ovarian cancer dataset. Fig (b) is the histogram of network based potential biomarkers of the ovarian cancer dataset. In both figures, the horizontal axis is the feature indexes, and the vertical axis shows how many times one feature is identified during the 10-fold iterations. From this figure, we can see that the network based criterion yields more consistent results than the performance based criterion.

3. BIOMARKER IDENTIFICATION

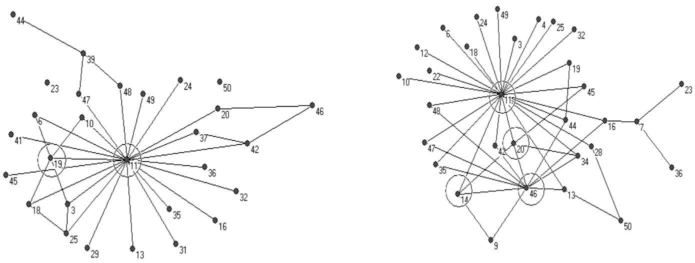
3.1 Classification-Performance-Based Biomarkers

In our early work [9], the concept of ensemble dependence model was applied to classify microarray gene expression data, but it was not used for biomarker identification because gene expression data is quite noisy. If individual genes are examined, large noises may overwhelm the underlying dependence relationship. However, in proteomic MS data, the peaks are relatively strong compared with noises. This enables us to examine individual mass features and their dependence relationship.

We examine three features at one time, and apply the proposed model for classification. Through an exhaustive search, all possible feature triples are examined, and the classification performance is recorded as a metric. Triples with classification accuracy higher than 95% are considered to be informative triples. Features with high appearance frequency in informative triples are regarded as important cancer biomarkers. These are biomarkers identified based on the criterion of classification performance. We call them the classification-performance-based biomarkers.

First, we examine the ovarian cancer MS dataset. To ensure reproducibility of the identified biomarkers, we apply a similar strategy with 10-fold cross validation. where, the ovarian cancer dataset is divided into 10 parts; 9 parts are used for model learning (training) and the one left is used for validation (testing). We search for biomarkers every iteration based on each different choice of training and testing samples. For each iteration, through an exhaustive search, the top 15 potential biomarkers are kept for reference. 9 features are commonly identified as biomarkers by 7 or more out of 10 iterations. Fig.2(a) shows the histogram of potential biomarkers, where horizontal axis is the feature indexes, and the vertical axis shows how many times one feature is identified during the 10-fold iterations. We can see that the result is not quite consistent.

We further examine the prostate MS dataset for the cases



(a) Normal case (b) Cancer case

Figure 3: Dependence networks for normal and cancer cases in ovarian cancer dataset. (Isolated nodes are omitted.) For the purpose of illustration, the circles are used to indicate the core features.

of both early stage and late stage. Our main purpose in analyzing this dataset is to demonstrate the consistency and possible difference between dominant biomarkers in early cancer stage and late stage. Similar with above analysis, 10-fold cross validation is applied. Again, every iteration, top 15 potential biomarkers are kept for reference, and features picked up by 7 or more out of 10 iterations are identified as biomarkers. Based on normal samples and early stage cancer samples, we identified 2 biomarkers. They are features $41, 48$. Based on normal samples and late stage cancer samples, we identified 5 biomarkers, $39, 42, 46, 48, 50$. When examining the histograms, we observed similar figures as Fig.2(a), which indicates that the performance-based criterion is not consistent under 10-fold cross validation. It is noted that, the number of biomarkers in late cancer stage is more than that in the early stage. And the identified biomarkers are not as consistent as those in the ovarian cancer dataset. Our intuitive explanation to the above observations is as follows. In 10-fold cross validation, we use different training samples to find potential candidates and regard the overlapping candidates as biomarkers. In early cancer stage, the protein expression pattern of prostate cancer may not be the dominant factor, thus, is more sensitive to different choices of training samples. Another reason might be, the prostate cancer dataset is round half the size of the ovarian dataset. Less number of samples may cause less consistency in 10-fold cross validation.

3.2 Dependence-network-based Biomarkers in Ovarian Cancer Dataset

In the ovarian cancer dataset, from binding triples of normal samples, we build a dependence network for normal case D_{normal}. From the binding triples of cancer samples, we build a dependence network for cancer case D_{cancer}. By comparing D_{normal} and D_{cancer}, we are able to see that which features go through a large topology change from normal to cancer and, thus, are potentially biomarkers. Similar to the previous subsection, 10-fold cross validation is applied. For each iteration, D_{normal} and D_{cancer} are calculated, and 15 features with large topology changes are kept for reference. 12 features are commonly identified as potential biomarkers by 7 out of 10 iterations. We call them the dependence-network-based biomarkers. Fig.2(b) shows histogram of identified biomarkers. From this figure, we can see that the network-based criterion yields much more consistent result, compared with the performance-based criterion.

From Fig.3, we can see the important features in the normal and cancer dependence networks. In the normal case, features 11 and 19 are important core features. They have rich dependence relationships with lots of other features. However, in the cancer case, there are more core features 11, 14, 20, 46. From normal case to cancer case, the number of dependence relationships increases, and the number of core features increases. Some unimportant features in normal case become core features in cancer case, especially feature 46. Similar with [9], it can be suggested that in cancer case, there are large noise variations which mess up the normal dependence relationships. These core features are strongly suggested to be biomarkers in ovarian cancer. It is our intention to investigate the origin and identity of these features.

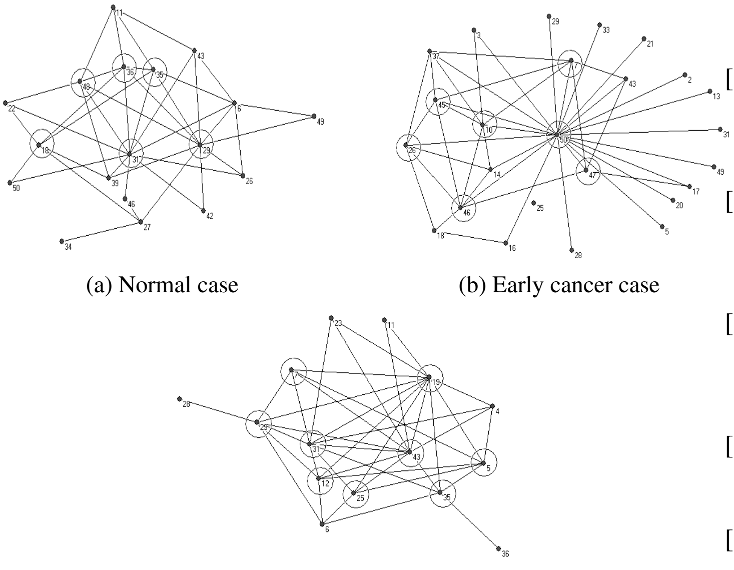
3.3 Dependence-network-based Biomarkers in Prostate Cancer Dataset

We further examine the prostate MS dataset. One objective here is to identify biomarkers active in early cancer and in late cancer stage. From binding triples of samples from normal, early cancer stage, and late cancer stage, we build dependence networks D_{normal}, D_{early} and D_{late}, respectively. Based on similar analysis to the case of ovarian cancer, we identify biomarkers for early stage cancer samples and late stage cancer samples, respectively. Consistent with the ovarian cancer dataset, similar observation with Fig.2 is observed, which indicates that the network-based criterion gives more consistent results under 10-fold validation than the performance based criterion.

The dependence networks for normal, early cancer stage and late cancer stage are shown in Fig.4. It is informative to examine the difference revealed by different dependence networks which are believed to provide insights into the major underlying phenotype under different situations. From this figure, we can see some interesting behaviors of the identified network-based biomarkers. For example, features 7, 12, 19, 43 are isolated or peripheral in normal stage. However, in late cancer stage, these features plays more important roles in the dependence network. These features may correspond to the key proteins on the pathways activated by the prostate cancer. On the contrary, features 18, 26, 36 are important network nodes in normal and early cancer stage. However, they become isolated or peripheral in late cancer stage. These feature may correspond to pathways disabled by the prostate cancer. The most interesting examples are features 11 and 31. They are important network nodes in both normal stage and late cancer stage. However, they seems to be deactivated in early cancer stage. These features might be the key to early stage cancer development, and deserve to be further investigated.

4. CONCLUSION

In this study, we extend a dependence modeling framework for cancer classification using MS data, propose to construct dependence networks between protein MS features, and take advantage of the differences revealed between dependence networks under different cancer and non-cancer situations to identify cancer biomarkers. With advantages lying in its nature as a model-driven approach, the proposed EDM-based classification scheme outperforms SVM, a widely applied supervised machine learning algorithm.



(a) Normal case

(b) Early cancer case

(c) Late cancer case

Figure 4: Dependence networks for the prostate cancer dataset: normal, early and late cancer cases. Nodes isolated in all cases are omitted for simplicity. For the purpose of illustration, the circles are used to indicate the core features.

In building the dependence network, the dependence relationship among several features can be indicated by the corresponding eigenvalue spectrum. From binding triples found via the desired eigenvalue spectrum, the dependence networks for both cancer and normal cases are built. We developed a dependence modeling and network framework to identify cancer biomarkers using protein MS data. The proposed framework provides two efficient schemes (i.e. performance-based and dependence-network-based) to identify MS features as biomarkers collectively. Based on real MS data examination, it is observed that the dependence-network-based approach provides much more consistent results in identifying biomarkers, as shown in Fig.2. This interesting consistency motivates us to further explore the idea of dependence network. We plan to further investigate this for potential cancer diagnosis usage.

REFERENCES

[1] M. Washburn, A. Koller, G. Oshiro, R. Ulaszek, D. Plouffe, C. Deciu, E. Winzeler, and J. Yates, "Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae", *Proc Natl Acad Sci USA*, vol. 100, pp. 3107-3112, 2003.

[2] E. Diamandis, "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: Opportunities and potential limitations", *Mol. Cell Proteomics*, vol. 3, pp. 367-378, 2004.

[3] H. Budzikiewicz, "Selected reviews on mass spectrometric topics", *Mass Spectrometry Reviews*, Vol. 24, Issue 4, pp. 611-612, 2005.

[4] J. Li, Z. Zhang, J. Rosenzweig, Y. Wang, and D. Chan, " Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer", *Clin Chem*, vol. 48, pp. 1296-1304, 2002.

[5] J. Liu and M. Li, "Finding Cancer Biomarkers from Mass Spectrometry Data by Decision Lists", *Proc. of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)*, 2004.

[6] E. Petricoin, A. Ardekani, B. Hitt, P. Jevine, V. Fusaro, S. Steinberg, G. Mills, C. Simone, D. Fishman, E. Kohn, and L. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer", *Lancet*, pp. 572-577, 2002.

[7] X. Fu, C. Hu, J. Chen, Z.J. Wang, and K.J.R. Liu, "Cancer genomics, proteomics, and clinic applications", Genomic Signal Processing and Statistics, Hindawi Publishing Corporation, 2005.

[8] H. Budzikiewicz, "Selected reviews on mass spectrometric topics", *Mass Spectrometry Reviews*, Vol. 24, Issue 4, pp. 611-612, 2005.

[9] P. Qiu, Z.J. Wang, and K.J.R. Liu, "Ensemble Dependence Model for Classification and Predication of Cancer and Normal Gene Expression Data", *Bioinformatics*, 21(14):3114-3121, 2005.

[10] T. Furey, N. Cristinanini, N. Duffy, D. Bednarski, M. Schmmer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data", *Bioinformatics*, vol. 16, pp. 906-914, 2000.

[11] R. Fisher, "The Use of Multiple Measurements in Taxonomic Problems" *Annals of Eugenics* 7, 179-188, 1936.

[12] B. Adam, Y. Qu, J. Davis, M. Ward, M. Clements, L. Cazares, O. Semmes, P. Schellhammer, Y. Yasui, Z. Feng, and G. Wright "Serum Protein Fingerprinting Coupled with a Pattern-matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men" *Cancer Research*, vol. 62, 3609-3614, 2002.

[13] Q. Liu, B. Krishnapuram, P. Pratapa, X. Liao, A. Hartemink, and L. Carin "Identification of Differentially Expressed Proteins Using MALDI-TOF Mass Spectra", *ASILOMAR Conference: Biological Aspects of Signal Processing*, November 2003.

[14] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligirui, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring" *Science*, 286, 531-537, 1999.

[15] C. Steinhoff, T. Muller, U. Nuber, and M. Vingron, "Gaussian mixture density estimation applied to microarray data" *RECOMB*, 147, 2003.

[16] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data" *Bioinformatics*, vol. 19, 563-570, 2003.

[17] E. van Someren, L. Wessels, E. Backer, and M. Reinders, "Genetic Network Modeling", *Pharmacogenomics*, 3, 507-525, 2002

[18] A. Walhout, and M. Vidal, "Protein interaction maps for model organisms" *Nat Rev, Mol Cell Biol.*, 2, 55-62, 2001.