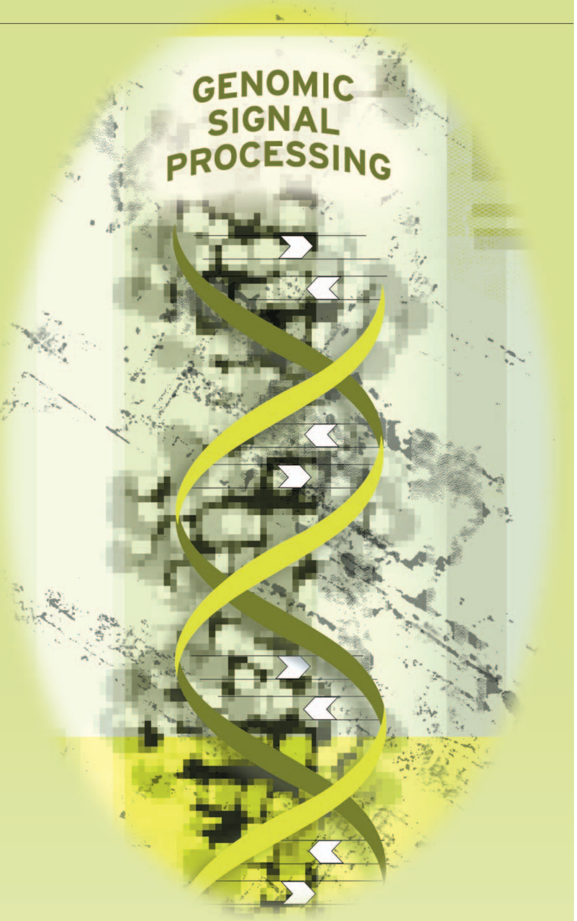
[Peng Qiu, Z. Jane Wang, and K.J. Ray Liu]



© EYEWIRE

Genomic Processing for Cancer Classification and Prediction

[A broad review of the recent advances in model-based genomic and proteomic signal processing for cancer detection]

Cancer is the fourth most common disease and the second leading cause of death in the United States. For instance, more than 500,000 people die from various forms of cancer each year in the United States. Cancer causes a significant financial burden to the health care system, in addition to the tremendous toll on patients and their families. Despite many advances derived from important innovations in technology during the last decades, in the field of cancer medicine, limited successes are still overshadowed by the tremendous morbidity and mortality incurred by this devastating disease. Therefore, the accurate detection, classification and early prediction of cancer is a research topic of significant importance. It has become increasingly important to integrate new technologies into cancer classification and prediction in hope to win the battle against cancer.

Life-science-based research has evolved rapidly during the past decade, driven largely by the sequencing of the complete genome of many organisms and high-throughput technological advances, such as microarray technique, with a shift from a reductionist approach towards an integrated approach. The new integrated approach investigating “complex” systems instead of individual components leads to the emerging field of systems biology aiming at a system-level understanding of biology systems. Since a thorough understanding of the DNA and protein related to cancer would eventually lead to breakthroughs in cancer study, we focus our attention on genomics and proteomics of cancer. Recently, microarray techniques, which allow measuring the expression level of thousands of genes simultaneously and thus present unique opportunities to investigate gene function on a genomic scale, are shown to provide insights into cancer study [1], [2] and have found promising applications in cancer classification by investigating molecular profiling based on gene expressions. Many different design formats of microarrays exist, and the types of gene expression assays include serial analysis of gene expression (SAGE), cDNA arrays (e.g., Stanford University), fiber optic arrays (e.g., Illumina), short oligonucleotide array (e.g., Affymetrix), and long oligonucleotide arrays (e.g., Agilent Inkjet). Since 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

1053-5888/07/$25.00©2007IEEE

cause of biological phenomenon, proteomics, the study of the proteins of a cell, is an emerging field in cancer research. The study of protein samples presents a new horizon for cancer classification and prediction. In recent years, protein separation methods coupled with various mass spectrometry (MS) technologies, considered as a major advance in the identification of polypeptides, have evolved as the dominant tools in the field of proteomics [3]. For protein samples, MS is a rapidly evolving methodology that converts proteins or peptides to charged pieces that can be separated on the basis of the mass-to-charge (m/z) ratio of the ionized proteins (or protein fragments). By measuring the intensity for different m/z ratio, the abundance of different peptides can be assessed. There are several types of MS ionization methods currently available, and interested readers are referred to [4]. In this article, we place our emphasis on signal processing and modeling of genomic and proteomic data from these two cutting edge technologies, namely microarray technology and MS technologies, as they are clearly among the leading frontiers that will rapidly reshape cancer study.

As the gene microarray and MS technologies become more accessible, microarray gene expression data and MS data analysis is finding applications in diverse areas of cancer study. Applied creatively, they can be used to test and generate new hypotheses. We now give some specific examples of microarray and MS's applications in cancer study to highlight the current advances and applications of these two technologies. The rationale behind these applications is based on the belief that the overall behavior of a cancer is determined by the expression profiles of genes/proteins. Some typical examples of the microarray's applications in cancer study include the following:

- Molecular classification of tumors. Serious limitations are associated with the traditional tumor classification method primarily based on morphological appearance, since tumors with similar histopathological appearance can follow significantly different clinical courses and show different responses to therapy. Many researches have been proposed in the literature for tumor classification by using gene expression data, including a hierarchical gene-clustering algorithm [5], a class discovery procedure based on self-organizing maps (SOMs) [6], an analysis procedure combining the ideas of partial least squares (PLS) and discriminant analysis (DA) in [7] for classifying (predicting) human tumor samples, and artificial neural networks (ANNs) approach in [8].
- Prediction of prognosis and tumor response to specific therapies. For example, in [1], gene expression profiling was used for predicting responders and nonresponders to chemotherapy.
- Drug development and identification of therapeutic targets. Expression microarray technology can be used to generate information rapidly for the identification and validation of novel therapeutic targets. For instance, analysis utilizing principle components analysis and hierarchical clustering revealed tissue-specific candidate targets in [9].

During the past few years, great attention has been focused on gene expression microarray data. Until very recently, MS has also been applied to cancer study. These approaches can be similarly categorized as in the case of microarray gene expression data classification. For instance, in [10] three biomarkers were selected using the linear combination based unified maximum separability analysis (UMSA) to best separate cancer and noncancer samples. A decision tree classification algorithm was discussed to differentiate prostate cancer from noncancer samples in [11]. The objective of these different examples can be generalized as to develop a multipurpose detector, classifier, and predictor based on the microarray data and MS data. The specific methods reviewed in later are applicable to address the above different concerns.

This tutorial article is organized as follows: we first review a few major design methodologies for cancer classification and prediction using genomic or proteomic data. We then present an ensemble dependence model (EDM)-based framework and discuss the concept of dependence network. The EDM framework is applied to both microarray gene expression and MS data sets in cancer study. Further, we present the performance-based idea and dependence network-based idea for biomarker identification. Our goal is to provide a broad review of the recent advances on model-based genomic and proteomic signal processing for cancer detection and prediction.

METHODS REVIEW FOR CANCER CLASSIFICATION

With the goal of understanding cancer development, assisting diagnosis, and treatment, many studies have investigated various methods for cancer classification using genomic and proteomic data. Current methods can be roughly divided into two broad categories: data-driven and model-driven methods.

DATA-DRIVEN METHODS

CLUSTERING METHODS

As clustering is probably the most popular type of data-driven classification methods, many clustering methods have been proposed for classifying cancer and normal genomic or proteomic samples. Some example schemes include hierarchical clustering [5], K-means and its variations [12], SOM [13], and local maximum clustering [14], among which the hierarchical clustering is most commonly applied in the literature. In hierarchical clustering, a dendrogram binary tree is constructed to describe the similarity between all genes. A similarity measure is used to examine pairs of genes, e.g., in [5] the similarity is defined based on a form of thresholded correlation coefficient. For a set of n genes, each of which is represented by a node in the dendrogram tree, the similarities of their expression profiles are examined pair wisely. Then the gene pair showing the largest similarity is replaced by a newly created node, whose expression profile is set to be the average expression profile of the gene pair. Now with a set of $(n-1)$ nodes, the similarity scores are computed again, and the gene pair yielding the largest similarity score will be merged. In this way, after repeating the same process for $(n-1)$ times, a binary tree is constructed for display. In K-means method, genes are partitioned into K clusters, where K is a predetermined number. The algorithm initializes K

centers, one for each cluster, and each gene is assigned to one cluster based on its similarity to the K centers. Then the centers are updated as the average of genes within each cluster. This iterative process between assigning individual genes and updating the cluster centers continues until convergence. In the SOM method, the number of clusters is also predetermined. Different from K-means method, the clusters are connected according to some predetermined topology. After initializing the cluster centers, genes are assigned one by one. For each gene, it is assigned to the cluster having the most similar center, and the cluster center is updated toward the gene. At the same time, the cluster centers that are connected to the assigned one are also updated toward the similar direction but with a smaller step size. When the algorithm converges, SOM gives a map of cluster centers, where connected neighbor clusters have similar expression profiles. The clustering methods are able to group together genes with similar expression profiles. It is proposed that genes with common functions can be identified based on similar expression profiles. However, clustering can only provide qualitative analysis. In addition, determining the number of clusters is a challenging problem itself, and there is a lack of widely accepted measures to systematically perform classification and evaluate the clustering performance.

MACHINE LEARNING METHODS

Several machine learning schemes have been proposed in the literature, such as K-nearest neighbors (KNN) [15], perceptron method [16], support vector machine (SVM) [17], and neural network [18]. They have also been demonstrated effective in many signal processing applications, such as the face detection, speaker/speech recognition, handwriting recognition, etc., especially the SVM method. The KNN is a nonparametric method for density estimation. For the purpose of supervised classification, the KNN algorithm is easy to implement. Given labeled training samples and unlabeled testing samples, for each testing sample, a label is assigned based on a majority vote of the K most nearby training samples. In general, KNN's classification performance is affected by such factors as the number of training samples and the size and dimension of the sample space. The perceptron method is actually a simple form of neural network. For each sample, a weighted sum of its features (e.g., expression of different genes) is used to infer the class label. The weights are learned from the training set and used for the classification of testing samples. The SVM algorithm is a powerful supervised learning algorithm. Given a set of binary labeled training data (e.g., normal and cancer subjects' gene expression profiles), the SVM finds a hyperplane that best separates the two classes of training data. Such a hyperplane is the maximal margin hyperplane, which has the maximum distance from the two classes of training data. After learning the hyperplane, for each testing data, SVM assigns the class label based on which side of the hyperplane the testing data is in. In [17], SVM was compared with perceptron method, and the superior performance of SVM was reported, where SVM yields nearly perfect classification performance. In general, machine-learning methods yield better classification performance than that of the clustering methods, since the main difference between cancer and normal data in the data domain can be revealed in the machine learning methods. However, there still lacks of means to interpret the difference from the biological, DNA/protein functional point of view.

MODEL-DRIVEN METHODS

To our knowledge, a few model-driven methods have been pursued in the literature for this purpose. One example is the Bayesian network classifier (BNC) [19], where a Bayesian network is induced from the data and then the resulting model is used as a classifier. In Bayesian network, joint multivariate probability distributions are used to model the regulation relationships between genes. In [20], Bayesian network is constructed from *Saccharomyces cerevisiae* cell-cycle time series data, where 76 genes were analyzed. In [21], based on a microarray data set of human fibroblast response to a serum, a Bayesian network is constructed, and its potential in oral oncology study is outlined. Another approach is the EDM, recently developed by the authors, where the dependence relationship between genes/proteins is examined. The details of EDM model will be presented in the following sections. In model-driven methods, a model is induced to describe experiment data. A meaningful model not only normally yields better classification performance but, more importantly, can provide insights into the underlying biology systems. The primary appeal of an approach like BNC or EDM lies in its automated hypothesis generation ability, and thus it is capable of reverse-engineering the biological networks. Therefore, model-driven methods can potentially play a major role in understanding biology systems.

ENSEMBLE DEPENDENCE MODEL FOR CANCER CLASSIFICATION AND PREDICTION

The EDM was recently developed by the authors, where the dependence relationship between genes/proteins is examined [22], [23]. EDM is different from previous studies: in clustering methods, genes/proteins are compared pair wisely to find genes that have similar expression profiles; in machine learning methods, although genes form a feature vector and are processed jointly, they are still treated in a separate fashion. In either case, genes' group behaviors and interactions are not considered. The proposed EDM approach takes the genes' group behaviors and interactions into account and yields promising results in cancer classification and prediction.

EDM

Because of the large dimensionality and small sample size of current available data, it is not feasible to examine the dependence relationship among all genes. To avoid this "curse of dimensionality" and reduce the noise effect, genes are grouped into several clusters first. We predict, given well-sorted clustering results, that genes' group behaviors and ensemble dynamics can be revealed. The methods of clustering will be discussed in a later section. In this section, we assume that genes are clustered properly and focus on the EDM model.

After clustering, each cluster contains specific genes that have a well-defined relationship to one another. The average gene expression profile is used to represent each cluster, so that the experiment noise can be averaged out and the genes' common expression within each cluster can be enhanced. Without any prior knowledge, we assume that each cluster is, to some extent, dependent on all the other clusters. The dependence relationship between gene clusters is described by the following linear model, as shown in Figure 1, where each arrow represents an inter-cluster dependence relationship. The weight a_{ij} associated with each arrow indicates to what extent cluster i is dependent on cluster j. The so-called self-regulation is assumed to be zero, i.e., $a_{ii} = 0$. Because the cluster average is used to represent each cluster, the intracluster dependence relationship within each cluster is averaged out. Also, it is proved that, from a mathematic point of view, allowing nonzero a_{ii} terms will make the model-learning process trivial and un-reasonable, since the results will simply be $a_{ii} = 1$ for any i, and $a_{ij} = 0$, for any $i \neq j$.

The dependence relationship shown in Figure 1 can be expressed as the following linear equation:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} \\ a_{21} & 0 & a_{23} & a_{24} \\ a_{31} & a_{32} & 0 & a_{34} \\ a_{41} & a_{42} & a_{43} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \end{bmatrix} \quad (1)$$

or, equivalently,

$$X = AX + N \quad (2)$$

where x_i, $i = 1, 2, 3, 4$ are the expression profiles for each cluster. The noise-like terms n_i are contributed by model mismatch and measurement uncertainty from microarray experiments. The matrix A is called the dependence matrix. Each element in the dependence matrix A describes to what extent one gene cluster is dependent on another cluster. In the following section, we will show that, the dependence matrix and the statistics of the noise term could be used to perform classification on cancer and normal gene/protein samples.

EDM-BASED MODEL LEARNING AND CLASSIFICATION

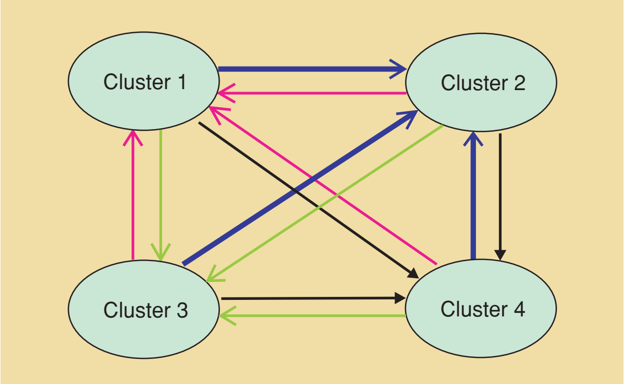
Since not all genes are informative in the classification of cancer and normal cases, feature selection is needed to exclude irrelevant genes. And, as required in the ensemble dependence model, gene clustering is performed to group together genes with similar expression. Then, the EDMs are used to describe the relationships among gene clusters, one model for the cancer case, and another for the normal case. With these two dependence models, a hypothesis-testing based method is applied to classify cancer and normal data. The main flow of the proposed classification method is shown in Figure 2. It includes four main components: feature selection, gene clustering, ensemble dependence model, and hypothesis testing.

FEATURE SELECTION

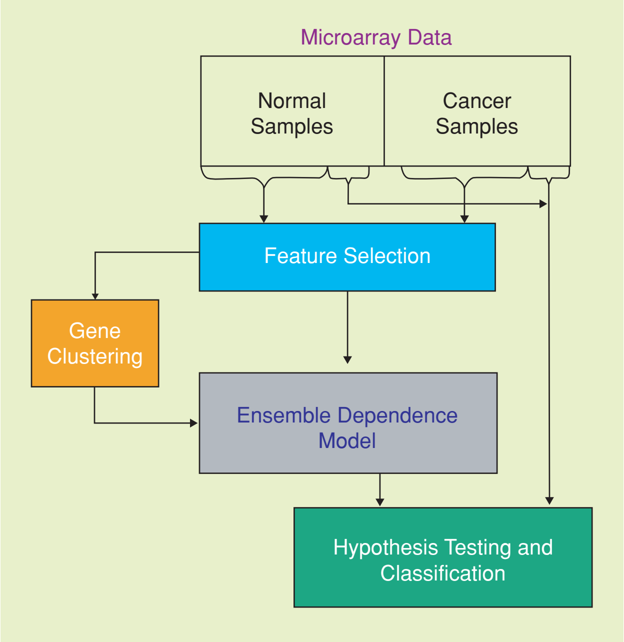
For the purpose of feature selection, the T-test is quite popular in microarray analysis. In the T-test, each gene is given a score, which evaluates the similarity between its expression profiles in cancer and normal samples. All genes are ranked according to their T-test scores. A p-value is chosen, and genes with scores lower than such a p-value are believed to behave most differently between cancer and normal samples. Another feature selection criterion was proposed in [6], using (3) to calculate a score for each gene

$$F(x_j) = \left| \frac{\mu_j^+ - \mu_j^-}{\sigma_j^+ + \sigma_j^-} \right| \quad (3)$$

where μ_j^+, σ_j^+ and μ_j^-, σ_j^- are the mean and standard deviation of gene j's expression level in cancer and normal samples, respectively. Similarly, genes are ranked and selected according to $F(x_i)$ scores. In [24], a more sophisticated feature selection criterion is discussed, where the features are ranked by the classification performance associated with certain classifier. It is reported that, for the purpose of classification, the optimal feature selection depends on the genes' expression profiles, the specific classifier, and the size



[FIG1] **Ensemble dependence model with number of clusters being four.**



[FIG2] **The EDM-based classification framework.**

of the training set. In the following section, clustering of genes is also a form of feature selection and dimension reduction.

In feature selection, one concern is the selection bias, regarding cross-validation methods. In [25], it is reported that, regarding cross validation, if the feature selection step is based on both training and testing data, the classification performance will tend to be higher. It is suggested that the feature selection step and the classification validation step should be considered jointly to avoid such bias. In other words, the feature selection step should be based on the training data only.

CLUSTERING OF GENES

As mentioned above, a proper way of gene clustering is required by the ensemble dependence model. Three standard clustering algorithms are considered in [22]: K-means [12], SOM [13], and Gaussian mixture model (GMM) [26]. K-means clustering is an unstructured method, and it depends more on algorithm initials. SOM is a soft-clustering method, but it blurs the difference between adjacent clusters, which is what we want to examine. Therefore, GMM is chosen to cluster genes, since it is a soft-clustering method, it can capture cluster difference, and it is much more stable than K-means clustering.

Before clustering, the number of clusters needs to be decided. The optimal number of clusters is difficult to determine, because it may depend on different diseases and different sets of genes under investigation. In literature, the MDL, AIC, and BIC criteria have been studied to solve such an order selection problem [27]. In our study, we took a simple approach to determine this parameter. We examined different choices, applied the proposed classification method, and suggested the best one by comparing the overall classification performance. In this study, the number of clusters is chosen to be four. Although the appropriate number of clusters is hard to determine, in general, the more clusters, the more the dependence relationship is examined, and the more differences between cancer and normal samples could be revealed. In practice, however, since the number of model parameters grows quadratically with the number of clusters, how many clusters to be analyzed is limited by the available samples size.

MODEL LEARNING AND CLASSIFICATION

The EDM-based classification scheme is a supervised learning method. Given the gene-clustering result, cluster average is calculated to represent each cluster. Based on cancer training data and normal training data with reduced dimensions, the ensemble dependence models for normal case (A_n and N_n) and cancer case (A_c and N_c) are estimated, respectively. The dependence matrix can be estimated row by row, based on the least squares (LS) criterion. For example, for the first row of the dependence matrix, $x_1 = a_{12}x_2 + a_{13}x_3 + a_{14}x_4 + n_2$, coefficients $a_{1i},\ i = 2, 3, 4$ that minimize noise term n_1 are estimated by projecting x_1 on to the subspace $\mathit{span}\{x_2, x_3, x_4\}$. The statistics of the noise-like term n_1 is estimated at the same time. The two estimated dependence models can form the two hypotheses in a binary hypothesis-testing problem, and the maximum likelihood rule can be applied to classify unknown testing samples based on the two hypotheses:

$$\begin{aligned} H_1 &: X = A_c X + N_c \\ H_0 &: X = A_n X + N_n. \end{aligned} \tag{4}$$

For each unknown testing sample X (samples not used in model learning), the ML decision rule is applied to predict whether it is cancer or normal. That is, we check whether the testing sample fits the cancer model better or fits the normal model better, by comparing the following two log-likelihoods:

$$\begin{aligned} \Pr(X\,|H_1) = &- 0.5\log((2\pi)^k|V_c|) - 0.5(X - A_c X - M_c)^T \\ &\times V_c^{-1}(X - A_c X - M_c) \\ \Pr(X\,|H_0) = &- 0.5\log((2\pi)^k|V_n|) - 0.5(X - A_n X - M_n)^T \\ &\times V_n^{-1}(X - A_n X - M_n) \end{aligned} \tag{5}$$

where k is the number of clusters and M_c, V_c and M_n, V_n are the first- and second-order statistics of the Gaussian noise-like terms in cancer and normal cases, respectively. Some details of the dependence model are available in our research Web site (http://dsplab.eng.umd.edu/~genomics/dependencenetwork/).

EDM ANALYSIS USING GENOMIC DATA

RESULTS ON MICROARRAY GENE EXPRESSION DATA

In [22], EDM is applied on five public-available cDNA microarray data sets and three affymetrix microarray data sets. In Tables 1 and 2, the results are the classification performance of leave-one-out cross validation. The results indicate that the EDM method is highly effective in distinguishing cancer and normal samples. Based on a comparison with a widely applied classifier, SVM, results show that both methods have similar classification performance. However the EDM algorithm presents a fundamental departure from the traditional SVM approach to classification because of its plausible hypothesis generation ability.

EIGENVALUE PATTERN AND EDM-BASED PREDICTION

The EDM yields excellent classification performance. Now, we want to explore its working principle. From the comparison of the estimated cancer dependence matrix A_c and the normal dependence matrix A_n, no clear difference is observed entry wisely. However, when exploring the eigenvalue domain, we observe two clearly different patterns. In Figure 3, 200 different subsets of the gastric data set are used to estimate cancer and normal dependence matrices and their eigenvalues are calculated and plotted. It is noted that, in general, the eigenvalues for the normal dependence matrix have larger absolute values than those of the cancer case. The difference is most distinct at the smallest eigenvalue. We believe that the different patterns in the eigenvalue domain could play an important role in cancer classification.

To explain the eigenvalue patterns, an ideal case is defined where there is no noise-like term in (2), meaning that the four cluster expression profiles are completely linearly dependent. In this case, the dependence matrix will have a special structure as follows:

$$A_{\text{ideal}} = \begin{bmatrix} 0 & \alpha_1 & \alpha_2 & \alpha_3 \\ \frac{1}{\alpha_1} & 0 & -\frac{\alpha_2}{\alpha_1} & -\frac{\alpha_3}{\alpha_1} \\ \frac{1}{\alpha_2} & -\frac{\alpha_1}{\alpha_2} & 0 & -\frac{\alpha_3}{\alpha_2} \\ \frac{1}{\alpha_3} & -\frac{\alpha_1}{\alpha_3} & -\frac{\alpha_2}{\alpha_3} & 0 \end{bmatrix}. \qquad (6)$$

It is proved that the eigenvalues of the above matrix are $1, 1, 1, -3$, no matter what are the values of α_i, $i = 1, 2, 3$. For a more general case where we have M clusters, we note that the eigenvalues of the M-by-M matrix A_{ideal} are $\{1, 1, \ldots, 1, -(M-1)\}$, no matter what are the values of α_i, $i = 1, 2, \ldots, M-1$ [23].

Based on the ideal case model, we gradually introduce larger and larger random variation to make the four cluster expression profiles more and more independent. At each variation level, a dependence matrix is estimated, and the corresponding eigenvalues are calculated. Compared with the ideal case, as the cluster expression profiles suffer more and more noisy variations, the eigenvalues of their dependence matrix will change and follow the trends shown in Figure 4(a). Compared with Figure 3, it can be suggested that the cluster expression profiles in cancer samples correspond to a much larger variation level than those of the normal samples. Here we try to explain this intuitively. In the normal samples, the gene clusters' dependence relationship is clearer, and gene clusters work more cooperatively. On the other hand, in the cancer case, the dependence relationship between gene clusters is overwhelmed by a large variation caused by diseases, which thus makes gene clusters work less cooperatively and makes the cell system become worse and worse. Moreover, the transition stage between normal and cancer patterns suggests that the resulting eigenvalue pattern from the proposed models can be used as a feature to predict the early stage of cancer development, i.e., whether a sample is in transition from healthy to cancer. The authors speculated that the patterns in eigenvalue spectrum have the promise of early-stage cancer prediction. To support the above argument, we use the prostate cancer data set as an example. The prostate data set contains four stages of data, NAP, BPH, PCA and MET, that can be simply regarded as being from normal (NAP and BPH), to early cancer stage (PCA), to late stage cancer (MET). The dependence matrix and eigenvalues of each stage are calculated. As shown in Figure 4(b), the overall trend of eigenvalues from normal to cancer follows the trend in Figure 4(a), which supports the above argument.

EDM ANALYSIS USING PROTEOMIC DATA

RESULTS ON MS PROTEOMIC DATA

Encouraged by the promising performance, the authors extended the original EDM concept to the protein MS data by taking into consideration the special properties of the proteomic MS data [23]. Due to the different properties of microarray and MS data, a few preprocessing step and different detail treatments is needed before EDM is applicable. However, the general framework is similar.

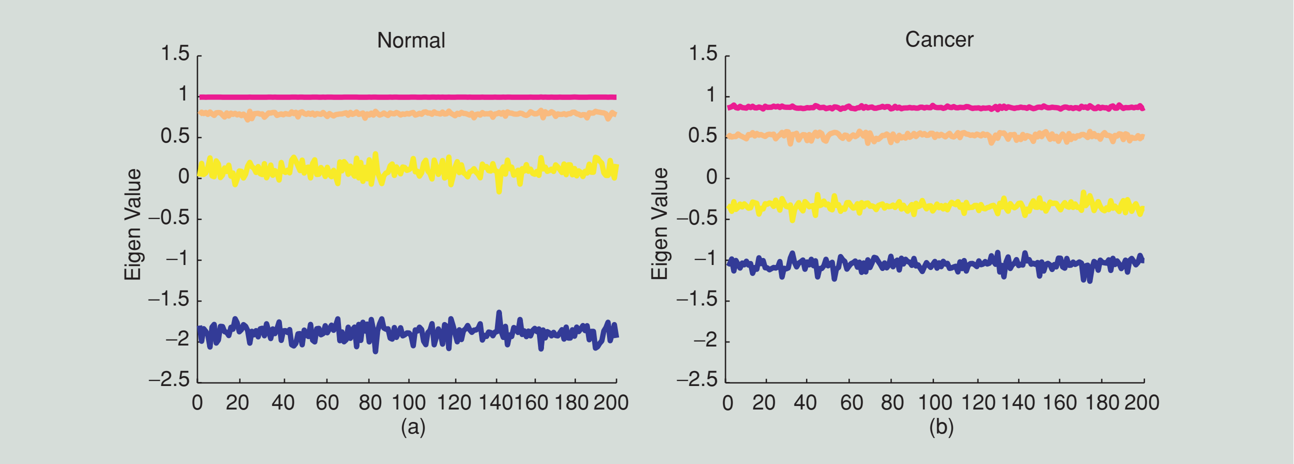
In the classification of microarray genes expression data, to effectively represent each gene cluster, the cluster expression profile is generated by the average expression profile of

[TABLE 1] CLASSIFICATION PERFORMANCE COMPARISON BETWEEN EDM AND SVM IN CDNA MICROARRAY DATA SETS.

DATA SET	EDM CLASSIFICATION RATE	SVM CLASSIFICATION RATE
GASTRIC CANCER	100%	99.1%
LIVER CANCER	98.72%	98.72%
PROSTATE CANCER	97.5%	100%
CERVICAL CANCER	93.9%	93.9%
LUNG CANCER	95.35%	97.67%

[TABLE 2] CLASSIFICATION PERFORMANCE COMPARISON BETWEEN EDM AND SVM IN AFFYMETRIX MICROARRAY DATA SETS.

DATA SET	EDM CLASSIFICATION RATE	SVM CLASSIFICATION RATE
COLON CANCER	88.71%	85.48%
PROSTATE CANCER	85.29%	91.18%
LUNG CANCER	97.79%	99.45%



[FIG3] Eigenvalue pattern of gastric data set: (a) the normal case and (b) the eigenvalues for the cancer case. Eighty percent normal samples of the gastric cancer microarray data set are randomly picked 200 times to learn 200 dependence matrixes. Eigenvalues are calculated and plotted in (a).

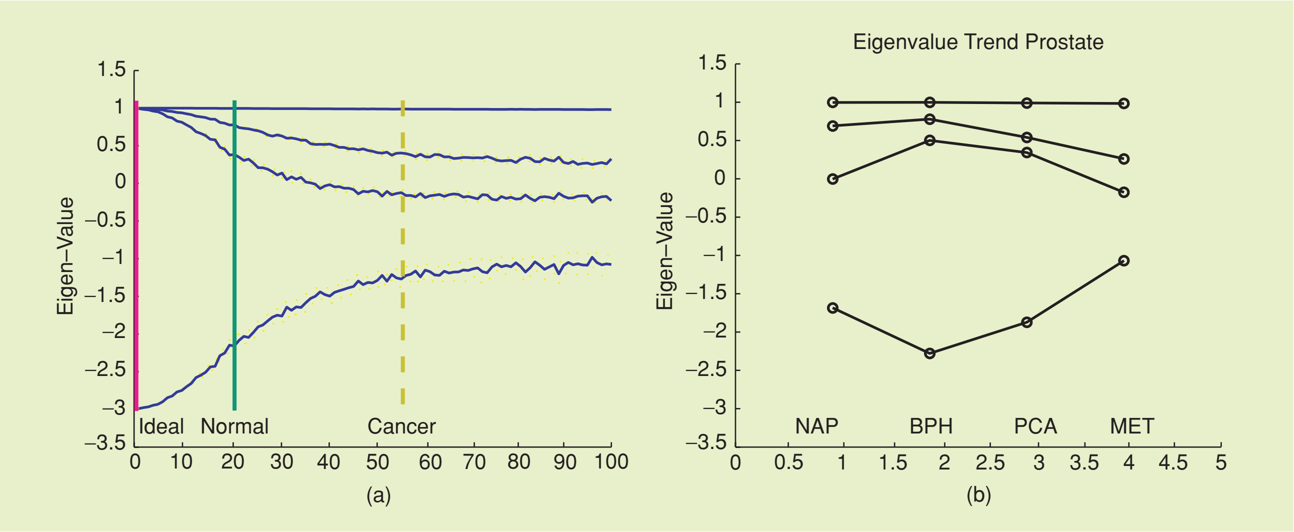
all genes within the cluster. However, due to the specific properties of the protein MS data, we propose the concept of virtual protein as cluster representatives, where virtual protein is generated by a weighted combination of different MS features within a cluster. The virtual protein representation is favored for two main reasons. First, in mass spectrum data, some features correspond to high intensity peaks, while some features correspond to low intensity peaks. To avoid high-intensity features dominating its cluster, the virtual protein is generated by a weighted average of the cluster members. Second, mass spectrometry measures the m/z ratio of the ionized peptides and their abundance 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 that better represents the underlying cancer-related protein. In our approach, the weights are determined through linear discriminant analysis (LDA). Since we are interested in virtual proteins that are cancer related and thus best represent the difference between a cancer and noncancer sample, LDA provides an efficient way to construct a virtual protein.

Except the virtual protein cluster representative, the classification framework for protein MS data is the same as that of the microarray gene expression data. EDM is applied on two protein MS experiment data sets. In Table 3, the results are the classification performance of leave-one-out cross validation. It is shown that EDM has high discriminate power in cancer and normal MS data. Moreover, it is noticed during comparison that, in one MS data set with normal samples, early-stage cancer samples and late-stage cancer samples, EDM and SVM have the same classification performance in distinguishing normal and late-stage cancer samples. However, when classifying normal and early-stage cancer samples, a task that is more difficult, EDM outperforms the SVM, as shown in Table 3.

EIGENVALUE PATTERN AND EDM-BASED DEPENDENCE NETWORK

The functionality of a molecular component (e.g., gene or protein) is not solely characterized by its own structure. Its surroundings and interacting/dependent components also play important roles in determining its function. In short, the interaction/dependence network can provide detailed functional insights of the whole system. Moreover, such a network is also the basis for finding biological signal pathways for diseases, which is important in understanding the diseases mechanism. These motivate to further explore the EDM concept and learn a dependence network for exploring the functionalities of the underlying biological system.

When studying gene clusters, two different eigenvalue patterns in normal and cancer samples are observed. For the gene microarray data, the analysis is based on gene clusters, not individual genes. 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 noise. This enables us to examine individual mass features and their dependence relationship. In recent study by the authors, when examining several individual protein MS features, the eigenvalue pattern also exhibit difference between cancer and normal cases. Recall Figure 4(a), where the eigenvalue pattern is closely related to dependence relationship. From the ideal case, as the features' expression suffer from more and more random variations, the eigenvalue pattern will change monotonically, especially the smallest eigenvalue. Therefore, the eigenvalue pattern can indicate how closely these individuals are dependent on one another. Thus, through the dependence model and eigenvalue pattern, closely dependent genes/proteins can be identified. From these dependence relationships, a network can be assembled, which is called a dependence network.



[FIG4] (a) The theoretical curve of the eigenvalue change caused by noise variation. The horizontal axis is variation level, which indicates how noisy the four cluster expression profiles are. As the cluster expressions become more independent, the eigenvalues of the corresponding dependence matrix will change and follow the curves. (b) The trend of eigenvalue change in the four-stage prostate data set, which matches the theoretical curve.

Since the eigenvalue pattern can serve as an indicator of how closely related the examined features 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 with one another. Because of the computational complexity of exhaustive search, we choose to examine three MS features at one time. Future analysis will be conducted to examine other model orders, four, five, etc. Take the ovarian cancer data set as an example. For the cancer and normal cases, respectively, all possible feature triples are examined. 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. We call this kind of triples the "binding triples." In the normal case, 520 triples pass the threshold; while in the cancer case, 269 triples pass the threshold. Moreover, there are only 80 overlapping triples. The small overlap indicates that, from healthy to cancerous, the overall dependence relationship goes through a major change.

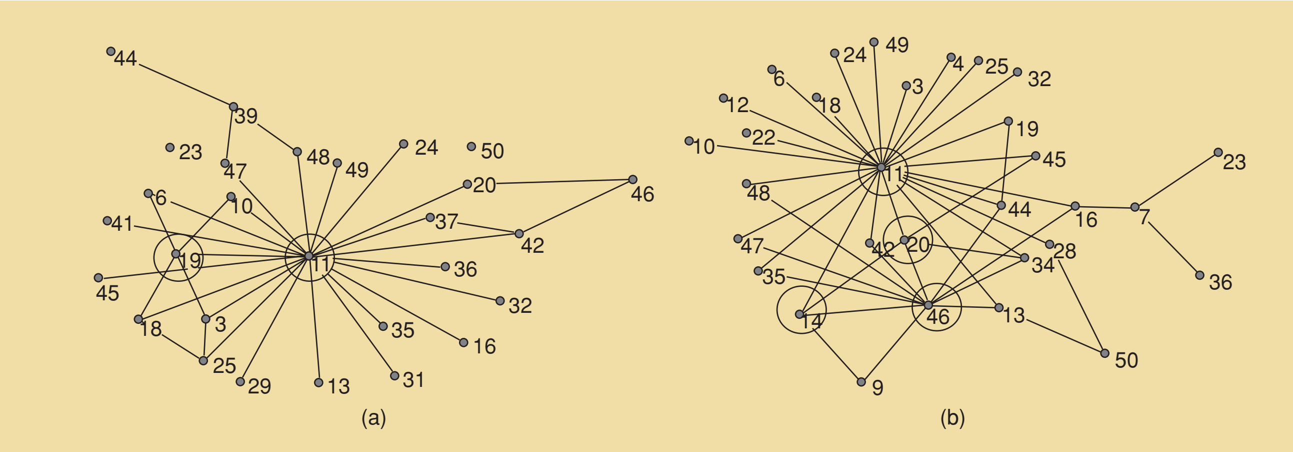
The dependence network is constructed from binding triples. As in graph theory, the topology of an n-node network can be represented by an $n \times n$ adjacency matrix D. If feature i and feature j both appear in a binding triple, it is suggested by the dependence model that feature i and feature j are closely related. So, we will count once for D_{ij}. Then, the adjacency matrix D is normalized by the total number of binding triples. Each element D_{ij} is a confidence value, describing the importance and strength of the connection between feature i and feature j. We call network D the dependence network. Since a confidence value D_{ij} associated with each connection indicates the strength of the dependence relationship, making use of this information, the dependence networks can be presented as shown in Figure 5, 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, $1/D_{ij}$. From Figure 5, we are able to see the importance of each node and identify potential biomarkers. 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. Similar to a previous study [22], it can be suggested that in cancer case, there are large variations which mess up the normal dependence relationships. These core features are strongly suggested to be cancer related biomarkers.

[TABLE 3] CLASSIFICATION PERFORMANCE OF EDM WITH DIFFERENT MODEL PARAMETERS AND SVM IN TWO MS DATA SETS.

	CLASSIFICATION RATE IN OVARIAN CANCER DATA SET	CLASSIFICATION RATE IN PROSTATE CANCER DATA SET, NORMAL VS. EARLY STAGE CANCER	CLASSIFICATION RATE IN PROSTATE CANCER DATA SET, NORMAL VS. LATE STAGE CANCER
3-CLUSTER EDM	100%	100%	100%
4-CLUSTER EDM	100%	98.18%	100%
5-CLUSTER EDM	96.60%	98.79%	99.39%
LINEAR SVM	96.83%	78.79%	98.79%

BIOMARKER IDENTIFICATION

In a certain disease, biomarkers are defined as the alternations of patterns at the cellular, molecular, or genetic level. These biomarkers normally serve as the indicators of diseases. Biomarker identification is a direction of great importance because it provides new insights into the early detection, diagnosis of cancer, and treatments. In [23], the authors have studied and compared two biomarker identification criteria derived from the dependence model and network: the classification performance-based criterion and the dependence network-based criterion. The two criteria are applied to MS data to find biomarkers. For the classification performance-based criterion, protein features are examined three at one time. A dependence model-based classifier is



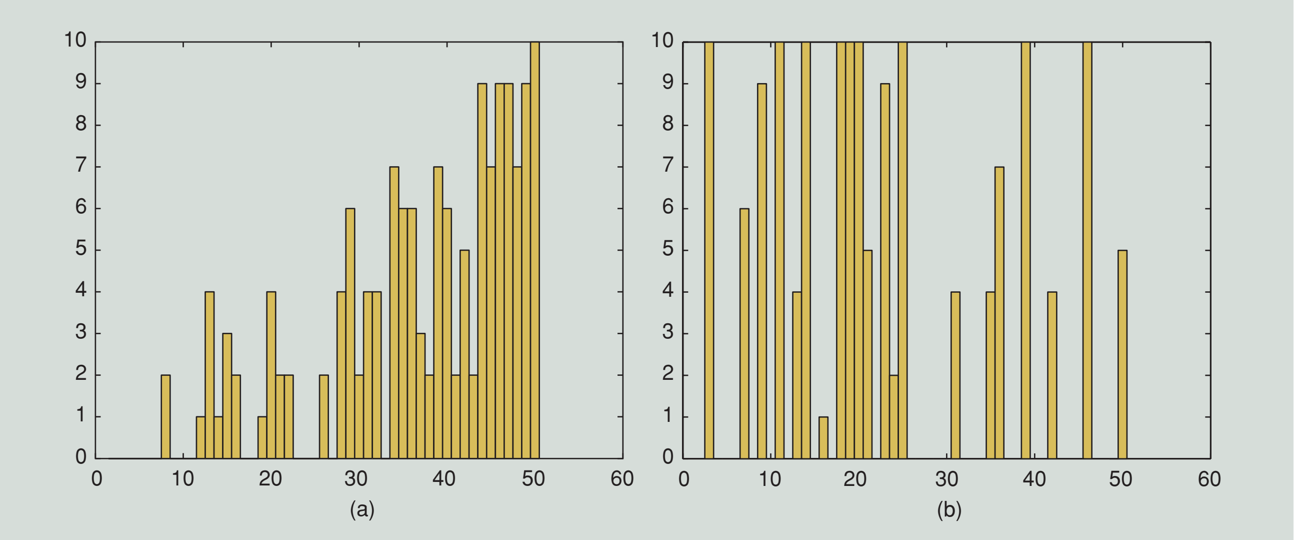
[FIG5] **Dependence networks for normal and cancer cases in the ovarian cancer MS data set. (Isolated nodes are omitted.) For the purpose of illustration, the circles are used to indicate the core features.**

build upon the three features to examine their classification power. Through an exhaustive search, the classification performances of all possible feature triples are examined. Triples with classification accuracy higher than 95% are considered to be informative triples. And, features frequently appear in the informative triples are regarded as important cancer biomarkers. These biomarkers are called the classification performance-based biomarkers. As described earlier, given a data set containing cancer and normal samples, dependence networks can be constructed for both cases. By comparing the dependence networks, more specifically by comparing the two adjacency matrixes for cancer and normal cases, the features with large topology changes are identified as biomarkers. These biomarkers are called the dependence network-based biomarkers.

Take the ovarian cancer MS data set as an example. To examine the reproducibility and consistency of the two criteria, both criteria starts with 50 features preselected from T-test, and ten-fold cross-validation strategy [28] is employed. During each of the ten iterations in ten-fold cross-validation, 15 biomarkers are identified from the classification performance-based criterion; 15 biomarkers are identified from the dependence network-based criterion. In Figure 6(a), the histograms of all classification performance-based biomarkers (during ten iterations) are shown. The horizontal axis shows the feature indexes, and vertical axis shows how many times out of the ten iterations a feature is identified. Since the histogram is not concentrated, from the widely spread histogram, we can conclude that the result is not quite consistent. In Figure 6(b), the histogram of all dependence network-based biomarkers is shown. It is observed that biomarkers identified by dependence networks are much more consistent than the biomarkers identified by classification performance. Another observation is that if we apply a simple differential method, such as T-test, for biomarker identification, the identified biomarkers will be features with indexes $40 \sim 50$ (since the pre-election 50 features are based on T-test). From Figure 6, we can see that the histogram of performance-based biomarkers have a high correlation with the simple differential method. However, the network-based criterion identifies many biomarkers that are not simply the most differentially expressed features. The results indicate that the network-based biomarker identification criterion yields much more information than the simple T-test and the performance-based criterion.

Another example is based on the prostate cancer MS data set, which contains three stages of samples: normal, early-stage cancer, and late-stage cancer. Two tasks are performed: one is to find biomarkers for the early cancer stage based on normal samples and early-stage cancer samples; and the other is to find a biomarker for the late cancer stage based on normal samples and late-stage cancer samples. Both the performance-based and the network-based criteria are examined. Similar to the previous example, the histograms of both methods are shown for the two tasks. In Figure 7(a) and (c), the histograms of performance-based biomarkers for the two tasks are both widespread, indicating a lack of consistency, while in Figure 7(b) and (d), the histograms of network-based biomarkers show a much higher consistency than the performance-based biomarkers. In comparison to the simple differential method, T-test, it is again observed that the performance-based criterion has high correlation with T-test while the network-based criterion yields much more information.

More other examples can be found in [29], where the two biomarker identification schemes have been applied to three protein MS data sets and two gene microarray data sets. Similar results are observed. In all investigated data sets, the network-based biomarkers consistently show much higher consistency and reproducibility than the performance-based biomarkers.

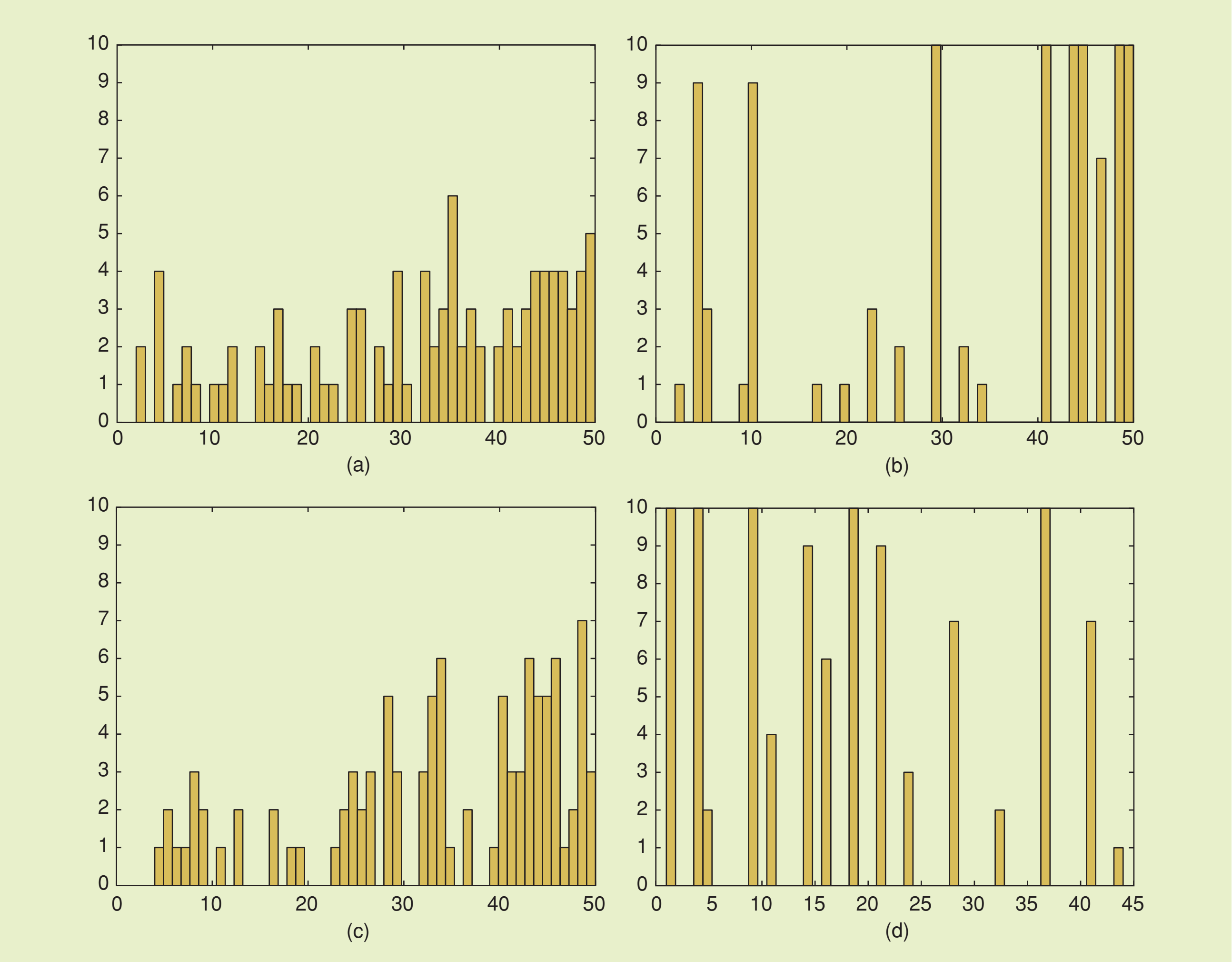


[FIG6] (a) Histogram of performance-based biomarkers in the ovarian cancer data set. (b) Histogram of network-based biomarkers of the ovarian cancer data set. In both figures, the horizontal axis is the feature indexes, and the vertical axis shows how many times one feature is identified during the ten-fold iterations. Since in ten-fold cross validation, there are ten iterations, one feature can be identified at most ten times. From these figures, we can see that the network-based criterion yields much more consistent biomarkers than the performance-based criterion, which shows the superiority of network-based criterion over the performance-based criterion.

(http://dsplab.eng.umd.edu/~genomics/dependencenetwork/results.htm). Moreover, the identified biomarkers are also examined from a biology point of view in [29], where the relationship between the functions of the identified biomarkers and cancer development is discussed.

CONCLUSIONS AND FUTURE DIRECTIONS

In conclusion, we surveyed a few major design methodologies for cancer classification and prediction using genomic or proteomic data and reviewed a dependence modeling and network framework for cancer classification and biomarker identification. The results on real data sets clearly show that the EDM method yields high accuracy, outperforming SVM, a widely applied supervised machine learning algorithm. The advantages of the EDM-based scheme lie in its nature as a model-driven approach. It takes features' group behaviors and interactions into account. This model-driven approach can reveal the relationship between the global gene/protein profiles and the subjects' health status. Moreover, the eigenvalue pattern observed in the gene microarray data shows promise in early cancer detection and prediction. The EDM concept is then extended to construct dependence networks between protein MS features, and to identify cancer biomarkers. The EDM framework provides two schemes (i.e., performance based and network based) to identify biomarkers. Based on real MS data examination, we found that the network-based approach provides much more consistent results in identifying biomarkers. This interesting consistency motivates us to further explore the idea of a dependence network. The encouraging results reported here demonstrate that the protein MS combined with the dependence modeling and network framework can both facilitate discovery of better biomarkers for different types of cancer and is promising to provide an efficient cancer diagnostic platform that can improve the early cancer detection and prediction. In the future, we plan to further explore the effects of each component within the proposed EDM-based framework. A mathematical analysis will be desired to quantify the effects of small sample sizes and to examine the effects of a possible model mismatch on the proposed scheme. Also, the eigenvalue spectrum seems



[FIG7] (a) and (b) Histograms of performance-based biomarkers and network-based biomarkers of the prostate cancer data set, respectively, when classifying normal samples against early stage cancer samples. (c) and (d) Histograms of performance-based biomarkers and network-based biomarkers of the prostate cancer data set, respectively, when classifying normal samples against late stage cancer samples.

promising for predicting the early stage of cancer development. Furthermore, the analyses of the proposed EDM are all based on static data, one gene/protein sample for every subject. However, it is believed that the analysis of time series data may reveal more information about the processes of cancer development. The authors conducted some preliminary studies about the quality control and synchronization of time series data [30]. Further efforts will be made to extend EDM analysis to dynamic time series data, for potential cancer diagnosis usage.

AUTHORS

Peng Qiu (qiupeng@umd.edu) received the B.Sc. degree from the University of Science and Technology of China in electrical engineering. He is currently a Ph.D. candidate and research assistant in the Electrical and Computer Engineering Department and the Institute for Systems Research at the University of Maryland, College Park. His researches are in the areas of genomic signal processing and biomedical imaging.

Z. Jane Wang (zjanew@ece.ubc.ca) received the Ph.D. degree from the University of Connecticut in 2002 in electrical engineering. She has been research associate of the Institute for Systems Research at the University of Maryland. Since August 2004, she has been with the ECE Department at the University of British Columbia, Canada, as an assistant professor. Her research interests are in the broad areas of statistical signal processing, with applications to information security, biomedical imaging, genomic and bioinformatics, and wireless communications. She co-received the 2004 EURASIP Best Paper Award and 2005 IEEE Signal Processing Society Best Paper Award . She coedited the book *Genomic Signal Processing and Statistics* and coauthored the book *Multimedia Fingerprinting Forensics for Traitor*. She is an associate editor for the *EURASIP Journal on Bioinformatics and Systems Biology*.

K.J. Ray Liu (kjrliu@ece.umd.edu) is a professor and associate chair of Graduate Studies and Research of Electrical and Computer Engineering Department, University of Maryland, College Park. He is the recipient of numerous honors and awards including best paper awards from IEEE Signal Processing Society (twice), IEEE Vehicular Technology Society, and EURASIP; IEEE Signal Processing Society Distinguished Lecturer, EURASIP Meritorious Service Award, and National Science Foundation Young Investigator Award. He also received various teaching and research awards. He is vice president, Publications, of the IEEE Signal Processing Society. He was editor-in-chief of *IEEE Signal Processing Magazine* and the *EURASIP Journal on Applied Signal Processing*.

REFERENCES

[1] J. Chang, E. Wooten, A. Tsimelzon, S. Hilsenbeck, M. Gutierrez, R. Elledg, S. Mohsin, C. Osborne, G. Chamness, D. Allred, and P. O'Connell, "Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer," *Mechan. Disease*, vol. 362, no. 9831, pp. 362–369, 2003.

[2] L. Van't Veer, H. Dai, and M. Van De Vijver, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[3] E. Diamandis, "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: Opportunities and potential limitations," *Mol. Cell Proteomics*, vol. 3, no. 4, pp. 367–378, 2004.

[4] X. Fu, C. Hu, J. Chen, Z.J. Wang, and K.J. Ray Liu, "Cancer genomics, proteomics, and clinic applications," in *Genomic Signal Processing and Statistics,* Edward R. Dougherty, Ilya Shmulevich, Jie Chen, and Z. Jane Wang, Eds. New York: Hindawi, 2005.

[5] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Nat. Academy Sci.*, vol. 95, no. 25, pp. 14, 683–14, 688, 1998.

[6] 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*, vol. 286, No. 5439, pp. 531–537, 1999.

[7] D. Nguyen and D. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, 2002.

[8] G. Bloom, I. Yang, D. Boulware, K. Kwong, D. Coppola, S. Eschrich, J. Quackenbush, and T. Yeatman, "Multi-platform, multisite, microarray-based human tumor classification," *Amer. J. Pathology*, vol. 164, no. 9, p. 16, 2004.

[9] M. Orr, A. Williams, L. Vogt, J. Boland, H. Yang, J. Cossman, and U. Scherf, "Discovery of 830 candidate therapeutic targets and diagnostic markers for breast cancer using oligonucleotide microarray technology," Nature Publishing Group, *Nature Genetics*, vol. 27, pp. 77, no. supp, 2001**,** http://www.nature.com/ng/journal/v27/n4s/index.html and http://www.nature.com/ng/journal/v27/n4s/full/ng0401supp_77b.html

[10] 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, no. 8, pp. 1296–1304, 2002.

[11] J. Liu and M. Li, "Finding cancer biomarkers from mass spectrometry data by decision lists," in *Proc. 2004 IEEE Computational Syst. Bioinformatics Conf.*, pp. 622-625.

[12] S. Tavazoie, D. Hughes, M. Campbell, R. Cho, and G. Church, "Systematic determination of genetic network architecture," *Nat. Genet.*, vol. 22, no. 3, pp. 218–285, 1999.

[13] T. Kohonen, *Self-Organizing Maps*. Berlin, Germany: Springer-Verlag, 1997.

[14] W. Wu, Y. Chen, R. Bernard, and A. Yan, "The local maximum clustering method and its application in microarray gene expression data analysis," *EURASIP J. Appl. Signal Processing*, vol. 2004, no. 1, pp. 53–63, 2004.

[15] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," *J. Computat. Biol.*, vol. 7, no. 3-4, pp. 559–583, Aug. 2000.

[16] F. Rosenblatt, "The preceptron: A probabilistic model for information storage and organization in the brain," *Psych. Rev.*, vol. 65, pp. 386–407, 1958.

[17] 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, no. 10, pp. 906–914, 2000.

[18] M. O'Neill, and L. Song, "Neural network analysis of lymphoma microarray data: Prognosis and diagnosis near-perfect," *BMC Bioinformatics*, vol. 4, no. 13, pp. 28–41, 2003.

[19] P. Helman, R. Veroff, S. Atlas, and C. Willman, "A Bayesian network classification methodology for gene expression data," *J. Computat. Biol.*, vol. 11, no. 4, pp. 581–615, 2004.

[20] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *J. Computat. Biol.*, vol. 7, no. 3-4, pp. 601–620, 2000.

[21] P. Sebastiani, Y. Yu, and M. Ramoni, "Bayesian machine learning and its potential applications to the genomic study of oral oncology," *Advances Dental Res.*, vol. 17, no. 1, pp. 104–108, 2003.

[22] 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*, vol. 21, no. 14, pp. 3114–3121, 2005.

[23] P. Qiu, Z.J. Wang, and K.J.R Liu, "Dependence modeling and network for biomarker identification and cancer classification," in *Proc. EUSIPCO*, 2006.

[24] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E.R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, 2005.

[25] C. Ambroise and G.J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," in *Proc. Nat. Academy Sci.*, vol. 99, no. 10, pp. 6562–6566, 2002.

[26] C. Steinhoff, T. Muller, U. Nuber, and M. Vingron, "Gaussian mixture density estimation applied to microarray data," *Lecture Notes in Computer Sciences (LNCS),* vol. 2810, pp. 418–429, 2003.

[27] X. Zhou, X. Wang, E. Dougherty, and S. Wong, "Gene selection using logistic regressions based on AIC, BIC, and MDL criteria," *New Mathematics and Natural Computation,* vol. 1, no. 1 129--145, 2005.

[28] A. Statnikov1, C. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multi-category classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[29] P. Qiu, Z.J. Wang, and K.J.R. Liu, "Dependence network modeling for biomarker identification," accepted by *Bioinformatics*, Oct, 2006.

[30] P. Qiu, Z.J. Wang, and K.J.R. Liu, "Polynomial model approach for resynchronization analysis of cell-cycle gene expression data," *Bioinformatics*, vol. 22, no. 8, pp. 959–966, 2006.