Journal of VLSI Signal Processing 35, 255–272, 2003

Discriminatory Mining of Gene Expression Microarray Data*

ZUYI WANG†
Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064, USA

YUE WANG
Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Alexandria, VA 22314, USA

JIANPING LU
Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064, USA

SUN-YUAN KUNG
Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

JUNYING ZHANG**
Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064, USA

RICHARD LEE
Lombardi Cancer Center, Georgetown University, Washington, DC 20007, USA

JIANHUA XUAN
Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064, USA

JAVED KHAN
National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

ROBERT CLARKE
Lombardi Cancer Center, Georgetown University, Washington, DC 20007, USA

Received November 2002, Revised March 2003

*This work was supported in part by the National Institutes of Health under Grants 5R21CA83231
†Present address: Center for Genetic Research, Children's National Medical Center, Washington, DC 20010, USA.
**Present address: Institute of Electrical Engineering and Institute of Computer Science, Xidian University, Xi'an, P.R. China 710071

Abstract. Recent advances in machine learning and pattern recognition methods provide new analytical tools to explore high dimensional gene expression microarray data. Our data mining software, VISual Data Analyzer for cluster discovery (VISDA), reveals many distinguishing patterns among gene expression profiles, which are responsible for the cell's phenotypes. The model-supported exploration of high-dimensional data space is achieved through two complementary schemes: dimensionality reduction by discriminatory data projection and cluster decomposition by soft data clustering. Reducing dimensionality generates the visualization of the complete data set at the top level. This data set is then partitioned into subclusters that can consequently be visualized at lower levels and if necessary partitioned again. In this paper, three different algorithms are evaluated in their abilities to reduce dimensionality and to visualize data sets: Principal Component Analysis (PCA), Discriminatory Component Analysis (DCA), and Projection Pursuit Method (PPM). The partitioning into subclusters uses the Expectation-Maximization (EM) algorithm and the hierarchical normal mixture model that is selected by the user and verified "optimally" by the Minimum Description Length (MDL) criterion. These approaches produce different visualizations that are compared against known phenotypes from the microarray experiments. Overall, these algorithms and user-selected models explore the high dimensional data where standard analyses may not be sufficient.

Keywords: computational bioinformatics, gene microarrays, finite normal mixture, cluster visualization and selection, machine learning

1. Introduction

With gene expression microarrays, the relative expression levels in two or more mRNA populations derived from tissue samples can be measured simultaneously for thousands of known genes [1, 2]. Thus, this technology is an efficient and cost-effective tool for large scale analysis of gene expression. Microarrays are composed of a platform (glass slide, nylon filter, or chip) to which are bound cDNAs or oligonucleotides that code for either known genes or Expressed Sequence Tags (ESTs). Data from gene expression microarrays have been used to classify clinical samples, to investigate the mechanism of drug action, to examine the effects of drugs on gene expression in yeast, and to identify and validate novel therapeutics for cancer patients [1, 3, 4]. Hidden links remain between genes and the biology of cancer; these links may be revealed through large scale gene expression analyses of normal and cancer cells. Specific gene expression patterns in malignant tissues determine their phenotypes, e.g., drug responses, growth proliferation rate, angiogenesis, and metastatic potential. Microarrays can measure concurrently the expression of individual genes but methods to analyze the complex, high dimensional data are not well developed [5]. Because the number of dimensions in a microarray data set frequently reaches several thousand, the development of accurate and robust analytical tools is crucial.

Advances in microarray technologies have enabled investigators to explore the dynamics of transcription on a molecular scale. The current challenge is to extract useful and reliable information from these large data sets. A common approach is cluster analysis. The primary objective of cluster analysis is to group samples that have comparable patterns of variation. This approach is useful for reducing the complexity of large data sets and for identifying predominant patterns within the data. However, additional methods are needed to extract information about individual samples from these large data sets.

A sample's gene expression profile (mRNA), is a snapshot of the transcriptome associated with that sample's phenotype. Each sample's profile is described as a point in d-dimensional gene expression space in which each axis represents the expression level of one gene. The presence of well-separated sample groups implies that the representations of samples within the same group are close to each other in this gene expression space but distant from those of other samples. Thus, the representations of phenotype-related samples form clusters.

To extract the most important information from gene expression profiles, our approach is divided into three major steps: cluster discovery, gene selection, and phenotype prediction. Cluster discovery detects previously unrecognized tumor subtypes [5]. Gene selection identifies the most relevant gene subset involving the biological process that generates the patterns. Phenotype prediction assigns each unknown tumor sample to a known tumor class [5]. The main challenge, however, is that the microarray data is high-dimensional, multimodal, and often lacking in complete prior knowledge.

Data clustering is a process of grouping together input data points with similar features in the

multi-dimensional space. The most common hierarchical clustering method often used by biologists for data clustering is the dendrogram [6]. Data points are arranged into a phylogenetic tree, the level of similarity of two pairs being represented by the length of the branch. While hierarchical clustering is simple and straight forward, it is designed to reflect true hierarchical tree structure. However, microarray data are not generated in this manner. It is very important to include more biological information rather than rigidly cluster data points. Hierarchical clustering may fail to group data points correctly because it is greatly influenced by local condition and cannot evaluate global data structure. In contrast, Self Organizing Maps (SOMs) attempt to search for relevant patterns by first imposing structure on the data with nodes that are expected to eventually move to the center of each cluster. The SOMs then updates the structure map at each iteration based on a data point randomly selected from the data set [7]. Thus, similar samples are grouped into the same cluster [7]. In unsupervised situations, the success of SOMs partially depends on the initialization of the map structure, e.g. number of nodes and different geometries. Without data modelling, SOMs lack criteria for validation of cluster structure, e.g. whether the number of clusters is optimal.

A gene clustering method based on graph theoretic techniques has been developed for the situation that the clusters are not assumed to be hierarchically structured [8]. Cluster information is mapped to an undirected graph where each clique in the graph indicates a cluster. It is assumed in the model that the input data contain underlying cluster structure contaminated by random errors. Through applying this clustering algorithm, with high probability, the cluster structure can be recovered by removing the random errors from the input data. However, the algorithm is developed for gene clustering in which the input data have much lower dimensionality (about 20 to 70) compared with our studies and those of many others (500 to 10000 or more dimensions). The capability of this method for handling very high dimensional data is uncertain. Furthermore, microarray data have significantly large overlaps among clusters, reflecting the nature of biological data. The ability of this method to cluster accurately data points where clusters overlap appears limited and may not be widely applicable to most very high dimensional gene expression data sets.

An interesting clustering approach using support vector concept is presented in [9], where data points are mapped to a high dimensional feature space and support vectors are used to define a sphere to enclose them. Data points are clustered hierarchically by adjusting parameters in the kernel function that mathematically represents the mapping from input space to feature space; outliers are allowed by setting appropriate penalty parameters. The method has advantages: it finds clusters with arbitrary shapes, dimensionality reduction is not needed, and the method can deal with outliers. However, the clusters with sizable overlaps cannot be correctly found by using this method, limiting its suitability for microarray data clustering.

We propose a model-supported hierarchical data exploration method that overcomes the limitations of other methods. Our method can evaluate the overall cluster structure, while hierarchical clustering and SOMs can only cluster blindly without knowing overall data structure. Our hierarchical data exploration scheme helps discover any hierarchical tree structure, if one exists, but is still valid if such a structure does not exist because the method is designed to discover the inner structure of any cluster. To find the best structure description, initialization of the clustering is supported by user interaction and verified by model selection criterion. Soft data decomposition allows modeling clusters with overlaps. The model selection procedure provides a theoretical and quantitative tool for cluster validation.

In this paper, we report our newly developed discriminatory data mining methods [10, 11]. There are three major components: (1) statistical modeling of gene expression microarray data with a Standard Finite Normal Mixture (SFNM) distribution; (2) development of a joint supervised and unsupervised data mining scheme to "discover" sample clusters in a discriminative visual pyramid; and (3) evaluation of the data clusters produced by such scheme with phenotype-known microarray experiments. Major differences are apparent between our work and the previous most related research [12–17]. First, since the high complexity of the data structure within a high dimensional space cannot be adequately explored by a single-level visualization [12], we developed a hierarchical visualization paradigm, involving mixture statistical sub-models and visualization subspaces. The resulting data mining tool is capable of capturing cluster distribution structure in high dimensional space and discover the relationships among clusters. Second, we propose three algorithms: (1) Discriminatory Component Analysis (DCA), (2) combined Projection Pursuit Method (PPM)/Independent Component Analysis

(ICA)—ICA-PPM, and (3) combined PPM/Principal Component Analysis (PCA)—PCA-PPM. They allow an effective separation of local clusters in dimensionally reduced visual spaces, which may represent the original data set well. Third, we implement an Incremental Expectation-Maximization (IEM) procedure to estimate SFNM distribution, and find the top two principal axes of each sub-cluster probabilistically for discovering its local structure. The computation is efficient when confronted with high dimensional data sets [18]. Finally, we impose a model selection procedure to determine the number of sub-clusters within each cluster using the minimum description length criterion. In addition, applying the MDL criterion also determines whether a further split or partition of a subspace should continue in completing the whole hierarchy [10, 16].

2. Theory

2.1. *Hierarchical Visual Data Exploration Scheme*

The purpose of cluster analysis is to determine whether there are certain number of well-defined data sets within the entire data distribution and/or derive most rational and optimal grouping scheme to partition data into a specified number of clusters.

Since a gene expression microarray data set is a mixture of samples of cancer and non-cancer, or a mixture of samples of various cancer phenotypes, and the main objective in this study is cluster discovery, we can use a model-supported approach to cluster the multimodal data set in the expression space through modeling the entire data set using a mixture model, i.e., SFNM model. There has been considerable success of applying SFNM model to model the distribution of the multi-modal data set [10, 16, 19]. The use of normal distribution to model each cluster is supported by (1) our observation on the microarray data histograms, (2) the biological interpretation and prediction, i.e., the samples with the same biological outcome (e.g., phenotype) are close to each other in the gene expression space but distant from those with different outcomes. The analogy between the characteristics of the normal distribution and those of the distribution of individual cluster implies that the clustering procedure supported by the mixture of normal distributions should produce sufficiently accurate cluster structure information. Moreover, because no prior knowledge is available for the true underlying distribution of individual cluster, a mixture model with unified distribution components is appropriate for this study. In the case that k clusters exist in the data set, a mixture model with k normal distributions can be used to describe the overall distribution of the data. We also estimate the density parameters of each cluster and the overall mixture.

Assume the sample points $\{\mathbf{t}_i\}$ in gene expression space form K_0 clusters $\{(\boldsymbol{\mu}_{\mathbf{t}1}, \mathbf{C}_{\mathbf{t}1}), \ldots, (\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k}), \ldots, (\boldsymbol{\mu}_{\mathbf{t}K_0}, \mathbf{C}_{\mathbf{t}K_0})\}$, where $\boldsymbol{\mu}_{\mathbf{t}k}$ and $\mathbf{C}_{\mathbf{t}k}$ are the mean vector and covariance matrix of cluster k, respectively. Using the SFNM to model the multi-modal distribution is usually successful [19]. Such a data distribution takes a sum of the following general form

$$p(\mathbf{t}_i) = \sum_{k=1}^{K_0} \pi_k g(\mathbf{t}_i \mid \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k}) \tag{1}$$

where π_k is the corresponding mixing proportion, with $0 \leq \pi_k \leq 1$ and $\sum \pi_k = 1$, and $g(\,)$ is the Gaussian kernel. Modeling SFNM on microarray data addresses a combination of the detection of the structural parameter K_0 (e.g., cluster discovery) and the estimation of regional parameters $(\pi_k, \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$ based on the observations $\mathbf{t}$. One natural criterion used for this modeling is the Maximum Likelihood (ML) estimation using the Expectation-Maximization (EM) algorithm [19].

The very high dimensionality of microarray data introduce difficulties in the revelation of data structure, which have been well studied in [10]. Cover's theorem on the separability of patterns tells us that one single projection on a dimension reduced space is not sufficient for revelation of the true data structure [18, 27]. Hierarchical visual exploration paradigm, involving hierarchical statistical models and visualization spaces/subspaces, may provide more opportunities for the user to understand the data distribution structure, and are essential for high dimensional microarray data study. We believe that introducing user interaction into the clustering algorithm is a more practical approach and greatly reduces both computational complexity and local optimum likelihood [10, 20]. A user-friendly graphical interface for data visualization is developed to allow the user to select initial centers of the data clusters. To visualize data, we further developed data projection methods based on the current methods used in [10], maximizing the discovery of cluster structures. Details about the various visualization techniques will be introduced in the next sub-section.

In this approach, the techniques involved are: statistical modeling of microarray data with SFNM distribution, discriminative data projections jointly

accomplished by supervised and unsupervised learning processes, soft cluster decomposition based on the EEM procedure, and evaluation of the data clusters with phenotype-known microarray experiments.

The hierarchical version of the SFNM model can be extended to include more levels based on the same principle as above. The more hierarchical levels the tree has, the more sub-models are used and the finer are the sub-models. The formation of the hierarchical visualization tree is guided and verified by model selection over $\mathbf{x}$-spaces/subspaces. Model selection refers to the detection of the structural parameter K (the number of clusters or sub-clusters). In addition to the user's visual inspection, we propose to use an information theoretic criterion, i.e., the MDL [21]. The MDL calculation is a model fitting procedure, in which an optimal model is selected such that the selected model best fits the observed data. Thus, the value of K is selected by minimizing

$$\mathrm{MDL}(K_a) = -\log(l_{\mathrm{ML}}) + 0.5K_a \log N \qquad (2)$$

where K_a is the number of free adjustable parameters, and l_{ML} is the joint maximum likelihood of the model.

2.2. *Discriminatory Data Projection*

The purpose of developing discriminatory data projection tools is to maximally discover hidden cluster structure in the data space. The consideration of using multiple data projection tools is primarily based on the fact that the performance of the individual projection scheme tends to be case-dependent due to limited number of data samples in nearly all existing microarray data. Therefore, it is insufficient to use only one projection tool, which may increase risk of losing chances to discover cluster structure. The four discriminatory projection tools presented in this paper are: PCA, DCA, PCA-PPM, and ICA-PPM. The details of each method are discussed in the following sub-sections.

2.2.1. Principal Component Analysis. PCA is an effective unsupervised method for achieving dimensionality reduction [14, 18, 22]. For a set of observed d-dimensional data vectors $\{\mathbf{t}_i\}, i \in \{1, \ldots, N\}$, the q principal axes $\mathbf{w}_m, m \in \{1, \ldots, q(\leq d)\}$, are those orthogonal axes onto which the retained variances under projection are maximal. It can be shown that the principal axes $\mathbf{w}_m$ are given by the q dominant eigenvectors (i.e., q maximal eigenvalues) of the sample covariance matrix

$$\mathbf{C}_{\mathbf{t}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}})(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}})^{\mathrm{T}} \qquad (3)$$

such that

$$\mathbf{C}_{\mathbf{t}}\mathbf{w}_m = \lambda_m \mathbf{w}_m \qquad (4)$$

where $\boldsymbol{\mu}_{\mathbf{t}}$ is the sample mean and λ_m is the eigenvalue. The vector $\mathbf{x}_i = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}})$, where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q)$, is thus a q dimensional new representation of the observed vector $\mathbf{t}_i$. Two issues contribute to the limitations of conventional PCA: its global linearity without considering data structure; and its optimality based on reconstruction error rather than pattern separability.

2.2.2. Discriminatory Component Analysis. If class information is known, the search of directions in data space for discovering cluster structure is under better guidance. There are two types of class information we may be able to obtain: known phenotypes from biological experimental setting, and cluster/sub-cluster information resulting from cluster decomposition under unsupervised condition. Therefore, DCA can be applied under both supervised and unsupervised conditions. For the top level projection, DCA can be a supervised process by using the known phenotype (class) information. However, DCA can also be used in an unsupervised situation on the top level and sub-levels by using the second type of class information discussed above. Demonstrations of different applications of DCA are shown in the Results Section.

The application of DCA under unsupervised situation is discussed first. Suppose that a data set is initially partitioned into $K_D(\leq K_0)$ clusters and density parameters $(\pi_k, \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$, are estimated for $k = 1, \ldots, K_D$, so that the probability density functions (pdf) of the distribution of each cluster, $g(\mathbf{t}_i \mid \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$ and the mixture model $p(\mathbf{t}_i)$ can be calculated. With all these information available, we can emphasize the inter-cluster separation for the multi-modal data set by replacing the total covariance matrix with the Fisher's scatter matrix [13, 22], i.e., to find the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{w}_m = \lambda_m \mathbf{w}_m \qquad (5)$$

where the *within-cluster scatter matrix* ($\mathbf{S}_w$) is the *joint* scatter of data point $\mathbf{t}_i$ around the conditional mean

vector $\boldsymbol{\mu}_{tk}$ of K_D classes (on the top level) or sub-clusters (on the sub-levels)

$$\mathbf{S}_w = \sum_{k=1}^{K_D} \pi_k \mathbf{C}_{tk} \tag{6}$$

with cluster conditioned covariance matrix

$$\mathbf{C}_{tk} = \frac{\sum_{i=1}^{N} z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{tk})(\mathbf{t}_i - \boldsymbol{\mu}_{tk})^{\mathrm{T}}}{\sum_{i=1}^{N} z_{ik}} \tag{7}$$

where

$$z_{ik} = \frac{\pi_k g(\mathbf{t}_i \mid \boldsymbol{\mu}_{tk}, \mathbf{C}_{tk})}{p(\mathbf{t}_i)}, \tag{8}$$

is the posterior probability of sample $\mathbf{t}_i$ belonging to cluster (model) k, and the *between-cluster scatter matrix* ($\mathbf{S}_b$) is the scatter of the cluster conditional mean vector $\boldsymbol{\mu}_{tk}$ around the overall data center $\boldsymbol{\mu}_t$

$$\mathbf{S}_b = \sum_{k=1}^{K_D} \pi_k (\boldsymbol{\mu}_{tk} - \boldsymbol{\mu}_t)(\boldsymbol{\mu}_{tk} - \boldsymbol{\mu}_t)^{\mathrm{T}} \tag{9}$$

where $\pi_k = \frac{1}{N}\sum_{i=1}^{N} z_{ik}$ and $\boldsymbol{\mu}_{tk} = (\sum_{i=1}^{N} z_{ik}\mathbf{t}_i)/(\sum_{i=1}^{N} z_{ik})$, such that the separability of patterns is maximized, that is

$$\mathbf{W} = \arg\max_{\mathbf{W}_0} \left\{ Trace\left(\mathbf{W}_0^{\mathrm{T}} \mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{W}_0\right) \right\} \tag{10}$$

This is termed as Discriminatory Component Analysis. Under a supervised condition, the density parameters, $\boldsymbol{\mu}_{tk}$ and $\mathbf{C}_{tk}$, in Eqs. (6) and (9) can be directly estimated based on the first type of class information

The original vectors $\{\mathbf{t}_i\}$ are linearly transformed by $\mathbf{W}$, a $d \times 2$ matrix, through $\mathbf{x}_i = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_t)$ into a two-dimensional projection space. For a normal distribution $g(\mathbf{t}_i \mid \boldsymbol{\mu}_{tk}, \mathbf{C}_{tk})$ over the data space, a similar dimensionally reduced probability distribution $g(\mathbf{x}_i \mid \boldsymbol{\mu}_{tk}, \mathbf{C}_{tk})$ of the new variable $\mathbf{x}$ in the projection space is simply defined by the *Radon* transform [29] of $g(\mathbf{t}_i \mid \boldsymbol{\mu}_{tk}, \mathbf{C}_{tk})$

$$g(\mathbf{x}_i \mid \boldsymbol{\mu}_{xk}, \mathbf{C}_{xk}) = \int g(\mathbf{t}_i \mid \boldsymbol{\mu}_{tk}, \mathbf{C}_{tk}) \delta(\mathbf{x}_i - \mathbf{W}^{\mathrm{T}}\mathbf{t}_i + \mathbf{W}^{\mathrm{T}}\boldsymbol{\mu}_t)\, d\mathbf{t} \tag{11}$$

where $\delta(\,)$ is the delta function. According to the linear superposition property of *Radon* transform and the projection invariant property of normal distribution, we have

$$f(\mathbf{x}_i) = \sum_{k=1}^{K_0} \pi_k g(\mathbf{x}_i \mid \boldsymbol{\mu}_{xk}, \mathbf{C}_{xk}) \tag{12}$$

as the counterpart of Eq. (1) in the $\mathbf{x}$-space defined by projection matrix $\mathbf{W}$.

However, when the data set is projected onto a single lower dimensional subspace, its inherent multi-modal nature may be partially or completely obscured according to Cover's theorem on the separability of patterns [18]. In other words, while the cluster structure of a data set may be evident from the higher dimensional space, it is possible to have the finer cluster patterns concealed after a single linear projection, leading to an unidentifiable correspondence between Eqs. (1) and (12) [10]. A novel approach is to model high-dimensional multi-modal data set with a hierarchical mixture model and accordingly with a collection of probabilistic principal discriminatory subspaces [10, 14–16], namely the exploratory cluster discovery.

Assume a top-level model consisting of a single *Radon* transform $\mathbf{W}$ and a mixture of $K_1 (< K_0)$ normal distributions $p(\mathbf{t}_i) = \sum_{k=1}^{K_1} \pi_k g(\mathbf{t}_i \mid \boldsymbol{\mu}_{tk}, \mathbf{C}_{tk})$ which is identifiable in $\mathbf{x}$-space, i.e., $f(\mathbf{x}_i) = \sum_{k=1}^{K_1} \pi_k g(\mathbf{x}_i \mid \boldsymbol{\mu}_{xk}, \mathbf{C}_{xk})$, we can form a two-level hierarchy by associating a group of SFNM sub-models with each model k at top-level

$$p(\mathbf{t}_i) = \sum_{k=1}^{K_1} \pi_k \sum_{j=1}^{K_{2,k}} \pi_{j|k} g\left(\mathbf{t}_i \mid \boldsymbol{\mu}_{t(k,j)}, \mathbf{C}_{t(k,j)}\right) \tag{13}$$

where $\pi_{j|k}$ again corresponds to a set of mixing proportions, one for each k, with $0 \le \pi_{j|k} \le 1$ and $\sum_j \pi_{j|k} = 1$, and $\sum_{k=1}^{K_1} K_{2,k} = K_0$. To reveal the hidden cluster pattern within each model k at top-level, i.e., $g(\mathbf{t}_i \mid \boldsymbol{\mu}_{tk}, \mathbf{C}_{tk}) = \sum_{j=1}^{K_{2,k}} \pi_{j|k} g(\mathbf{t}_i \mid \boldsymbol{\mu}_{t(k,j)}, \mathbf{C}_{t(k,j)})$, an associated probabilistic principal discriminative subspace is constructed that focuses on the separability of patterns within the data portion defined by model k, where the opaque degree of a data point in the subspace plot is proportional to its posterior probability of belonging to this model, i.e., z_{ik} determined at top-level.

The further cluster discovery is a two-stage procedure: a soft partitioning of each model k into $K_{2,k}$ sub-clusters followed by a construction of corresponding subspace. Instead of assigning each given data point exclusively to one subspace, the contribution to its generation is shared among all the subspaces. As discussed

above, with pre-estimated sub-cluster density parameters, the subspaces of the sub-models at second-level can generated by the probabilistic DCA such that

$$\mathbf{S}_{k\_w}^{-1}\mathbf{S}_{k\_b}\mathbf{w}_{k\_m} = \lambda_{k\_m}\mathbf{w}_{k\_m} \tag{14}$$

where $\mathbf{S}_{k\_b} = \sum_{j=1}^{K_{2k}} \pi_{j|k}(\boldsymbol{\mu}_{\mathbf{t}(k\_j)} - \boldsymbol{\mu}_{\mathbf{t}k})(\boldsymbol{\mu}_{\mathbf{t}(k\_j)} - \boldsymbol{\mu}_{\mathbf{t}k})^{\mathrm{T}}$, $\boldsymbol{\mu}_{\mathbf{t}(k\_j)} = \sum_{i=1}^{N} z_{i(k\_j)}\mathbf{t}_i / \sum_{i=1}^{N} z_{i(k\_j)}$, $z_{i(k\_j)} = z_{ik}\pi_{j|k}g(\mathbf{t}_i \mid \boldsymbol{\mu}_{\mathbf{t}(k\_j)}, \mathbf{C}_{\mathbf{t}(k\_j)})/g(\mathbf{t}_i \mid \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$, $\pi_{j|k} = \sum_{i=1}^{N} z_{i(k\_j)} / \sum_{i=1}^{N} z_{ik}$, and $\mathbf{S}_{k\_m} = \sum_{j=1}^{K_{2k}} \pi_{j|k}\mathbf{C}_{\mathbf{t}(k\_j)}$, $\mathbf{C}_{\mathbf{t}(k\_j)} = \sum_{i=1}^{N} z_{i(k\_j)}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}(k\_j)})(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}(k\_j)})^{\mathrm{T}} / \sum_{i=1}^{N} z_{i(k\_j)}$. The probability distribution of model k in $\mathbf{x}$-space at second-level is now defined by the model k focused *Radon* transform of $g(\mathbf{t}_i \mid \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$, i.e., $g(\mathbf{x}_i \mid \boldsymbol{\mu}_{\mathbf{x}k}, \mathbf{C}_{\mathbf{x}k}) = \int g(\mathbf{t}_i \mid \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})\delta(\mathbf{x}_i - \mathbf{W}_k^{\mathrm{T}}\mathbf{t}_i \mid \mathbf{W}_k^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{t}k})\, d\mathbf{t}$. It should be noted that each component in Eq. (13) now corresponds to an independent sub-model with projection matrix $\mathbf{W}_k$ that is obtained through PCA on the pre-estimated $\mathbf{C}_{\mathbf{t}k}$. To interpret the corresponding set of visualization subspaces, all data points $\mathbf{x}_{ik} = \mathbf{W}_k^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})$ will appear in every plot of the K_1 subspaces at the second-level, with their opaque degree proportional to z_{ik}.

2.2.3. Principal Component Analysis-Projection Pursuit Method. To search projections for cluster separability, we take an alternative unsupervised approach, the Projection Pursuit Method (PPM). PPM searches for "interesting" projections, although, it is not universally agreed upon what constitutes an "interesting" projection in PPM research community. We require a projection where the data points separate into distinct and meaningful clusters [23]. We have two particular goals of using PPM in this project: (1) find low dimensional (equal or less than three) projections that provide the most revealing view of the overall data distribution; (2) use PPM for dimensionality reduction so that we will focus directly on the discriminatory projections rather than indirectly searching through covariance. The advantage of PPM is that it finds the directions that are not affected by the linear scale and correlational structure of the data, which is the disadvantage of PCA.

To include a user input for pattern discovery in high dimensional data, we first look at the projections onto the spaces spanned by two or three of the dimensions. However, any arbitrary direction could be the right one for cluster structure discovery and we have to search the space in all possible directions. In our approach, we have simplified the PPM by using non-Gaussianity as a criterion and using PCA and Independent Component Analysis (ICA) as vehicles for finding discriminatory projections.

The direction where the distribution of the projected data set is a Gaussian or super-Gaussian distribution is the one containing the least data structural information; on the other hand, the least Gaussian distribution indicates plentiful structural information. If the data distribute as one Gaussian or super-Gaussian distribution (a "spiky" pdf with long and heavy tails) in a direction, it implies that the data points are most likely forming a one-cluster structure instead of the multi-modal structure necessary for adequate cluster separation. On the contrary, the data may construct two or more clusters in the directions where distributions are non-Gaussian, e.g., sub-Gaussian (a "flat" pdf and more like a uniform distribution). We used kurtosis as the non-Gaussianity measure given by

$$kurt(m) = E((Y_m - \mu_{Ym})^4) - 3(E((Y_m - \mu_{Ym})^2))^2 \tag{15}$$

where $Y_m = \{\mathbf{w}_m^{\mathrm{T}}\mathbf{t}_1, \ldots, \mathbf{w}_m^{\mathrm{T}}\mathbf{t}_i, \ldots, \mathbf{w}_m^{\mathrm{T}}\mathbf{t}_N\}$ is a random variable consisting of the projections of all data points onto eigenvector $\mathbf{w}_m$ derived from PCA via Eq. (4), and $\mu_{Ym} = E(Y_m)$. Kurtosis can be positive or negative. Positive and negative values of kurtosis indicate the super-Gaussian distribution and sub-Gaussian distribution respectively [24, 25].

We try to fully utilize our PCA results to test PPM and reduce computational load. A prototype computer algorithm, termed as PCA-PPM, is implemented to calculate the kurtosis of the projection on each principal axis resulting from PCA and rank them so that the identified optimal directions are those that show the strongest sub-Gaussianity.

2.2.4. Independent Component Analysis-Projection Pursuit Method. Independent component analysis (ICA) is a recently developed method for finding linear representations of non-Gaussian data such that the components are statistically independent, or as independent as possible. Since PPM is designed to search the directions with the least Gaussian distribution and the least Gaussian distribution is the criterion for estimating ICA model, ICA and PPM are closely related. The non-Gaussianity measures can be adopted as projection pursuit functional indices. The ICA-PPM algorithm, which is PPM assisted with ICA, is then implemented to directly search for directions with

non-Gaussian distribution in data space [22, 24]. Similar to PCA-PPM, kurtosis is calculated on each independent component resulting from ICA and the two components with the most negative kurtosis are chosen for 2-D projection.

3. Algorithm

We now present the description of our algorithms that progressively proceeds by fitting a series of sub-models to the clusters of the data set interactively and incrementally.

To address the problem of high dimensionality with microarray data, we try to design a frame-work consisting of complementary model selection and EM clustering for estimating mixture model parameters. First, using the dimensionality reduction tools discussed above, we can obtain a low dimensional space perceptible by human for visual inspection to the cluster structure and moreover cluster initialization. We project the data set into a 2-D space, **x**-space, through a linear transformation, $\mathbf{x}_i = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}})$, where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2)$ is a set of top two most significant principal axes resulting from one of the dimensionality reduction methods, i.e., PCA, PCA-PPM, ICA-PPM and DCA based on their individual specific principles. In the 2-D data representation, the user can initialize multiple cluster structures by pinpointing cluster centers using a computer pointer device [10], which results in initial cluster centers $\boldsymbol{\mu}_{\mathbf{x}k}^{(0)}$ and assign $\pi_k^{(0)} = 1/K_1$ and $\mathbf{C}_{\mathbf{x}k}^{(0)} = \mathbf{W}^{\mathrm{T}}\mathbf{C}_{\mathbf{t}}\mathbf{W}$ in the case that a cluster structure with K_1 clusters are initialized. The model parameters $(\pi_k, \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$ are first estimated in **x**-space, and then fine tuned in **t**-space. The Incremental EM (IEM) procedure [26, 28] provides "soft" partitioning of the data points, hence allowing each data point to contribute simultaneously to multiple clusters, which results in

E-Step

$$z_{(i+1)k}^{(i)} = \frac{\pi_k^{(i)} g\left(\mathbf{x}_{i+1} \mid \boldsymbol{\mu}_{\mathbf{x}k}^{(i)}, \mathbf{C}_{\mathbf{x}k}^{(i)}\right)}{f\left(\mathbf{x}_{i+1} \mid \pi_k^{(i)}, \boldsymbol{\mu}_{\mathbf{x}k}^{(i)}, \mathbf{C}_{\mathbf{x}k}^{(i)}\right)}, \tag{16}$$

M-Step

$$\boldsymbol{\mu}_{\mathbf{x}k}^{(i+1)} = \boldsymbol{\mu}_{\mathbf{x}k}^{(i)} + a(i)\left(\mathbf{x}_{i+1} - \boldsymbol{\mu}_{\mathbf{x}k}^{(i)}\right) z_{(i+1)k}^{(i)}, \tag{17}$$

$$\mathbf{C}_{\mathbf{x}k}^{(i+1)} = \mathbf{C}_{\mathbf{x}k}^{(i)} + b(i)\left[\left(\mathbf{x}_{i+1} - \boldsymbol{\mu}_{\mathbf{x}k}^{(i)}\right)\left(\mathbf{x}_{i+1} - \boldsymbol{\mu}_{\mathbf{x}k}^{(i)}\right)^{\mathrm{t}} - \mathbf{C}_{\mathbf{x}k}^{(i)}\right] z_{(i+1)k}^{(i)}, \tag{18}$$

$$\pi_k^{(i+1)} = \frac{i}{i+1}\pi_k^{(i)} + \frac{1}{i+1} z_{(i+1)k}^{(i)} \tag{19}$$

for $k = 1, \ldots, K_1$, where $a(i)$ and $b(i)$ are introduced as the learning rates, two sequences converging to zero, ensuring unbiased estimates after convergence. In the ith iteration (i starts from 0), in the E-step, the "contribution" of an input sample point $(i+1)$ to the component (cluster) k, i.e., $z_{(i+1)k}$, the posterior probability of belonging to component k, is computed with the current density parameter estimates; in the M-step, new density parameters are re-estimated with the inclusion of data point $(i+1)$ and the current updated value of $z_{(i+1)k}$. The optimum value of K_1 is determined based on MDL (e.g., Eq. (2)) where $K_a = 6K_1 - 1$.

The corresponding sub-level mixture model $\sum_{j=1}^{K_{2,k}} \pi_{j|k} g(\mathbf{t}_i \mid \boldsymbol{\mu}_{\mathbf{t}(k,j)}, \mathbf{C}_{\mathbf{t}(k,j)})$ can again be estimated using IEM algorithm to allow a SFNM distribution with $K_{2,k}$ sub-models to be fitted to cluster k. Again, cluster k will be projected into a 2-D space represented by the top two most significant principal axes $\mathbf{W}_k$ determined by performing PCA on $\mathbf{C}_{\mathbf{t}k}$ (readily available via top level **t**-space clustering). The user will again pinpoint the initial sub-cluster centers $\boldsymbol{\mu}_{\mathbf{x}(k,j)}^{(0)}$ and assign $\pi_{j|k}^{(0)} = 1/K_{2,k}$ and $\mathbf{C}_{\mathbf{x}(k,j)}^{(0)} = \mathbf{W}_k^{\mathrm{T}}\mathbf{C}_{\mathbf{t}k}\mathbf{W}_k$ to initialize $\sum_{j=1}^{K_{2,k}} \pi_{j|k} g(\mathbf{x}_i \mid \boldsymbol{\mu}_{\mathbf{x}(k,j)}, \mathbf{C}_{\mathbf{x}(k,j)})$, and an optimal $K_{2,k}$ is obtained through the model selection procedure for cluster k. The estimation of the principal axes $\mathbf{W}_k$ and model density parameters $(\pi_{j|k}, \boldsymbol{\mu}_{\mathbf{t}(k,j)}, \mathbf{C}_{\mathbf{t}(k,j)})$ of the cluster k at the second level can be viewed as a two-step estimation problem, in which further splitting of the sub-models is determined within each of the clusters identified at the top-level such that its internal structures can be further explored over cluster-focused **x**-space.

The construction of the entire hierarchical tree is completed when no further data splitting is recommended in all of the current subspaces, followed by the generation of the bottom level subspaces (for example, the third level). Every data point $\mathbf{x}_{i(k,j)} = \mathbf{W}_{k,j}^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}(k,j)})$ will appear in every plot of the total K_0 subspaces at the bottom level with its color depth proportional to the conditional probability of belonging to sub-cluster (k, j) given by $z_i(k, j)$, where $\mathbf{W}_{k,j}$ for sub-cluster (k, j) is calculated through PCA on $\mathbf{C}_{\mathbf{t}(k,j)}$.

4. Results

A demonstration of the capability of finding cluster structure by PCA, PCA-PPM, DCA, and ICA-PPM is first done on a simulated data set that consists of three dimensions and four clusters ($N = 200$ for each cluster). The results are illustrated in Fig. 1, where we can see that the maximal information about the cluster

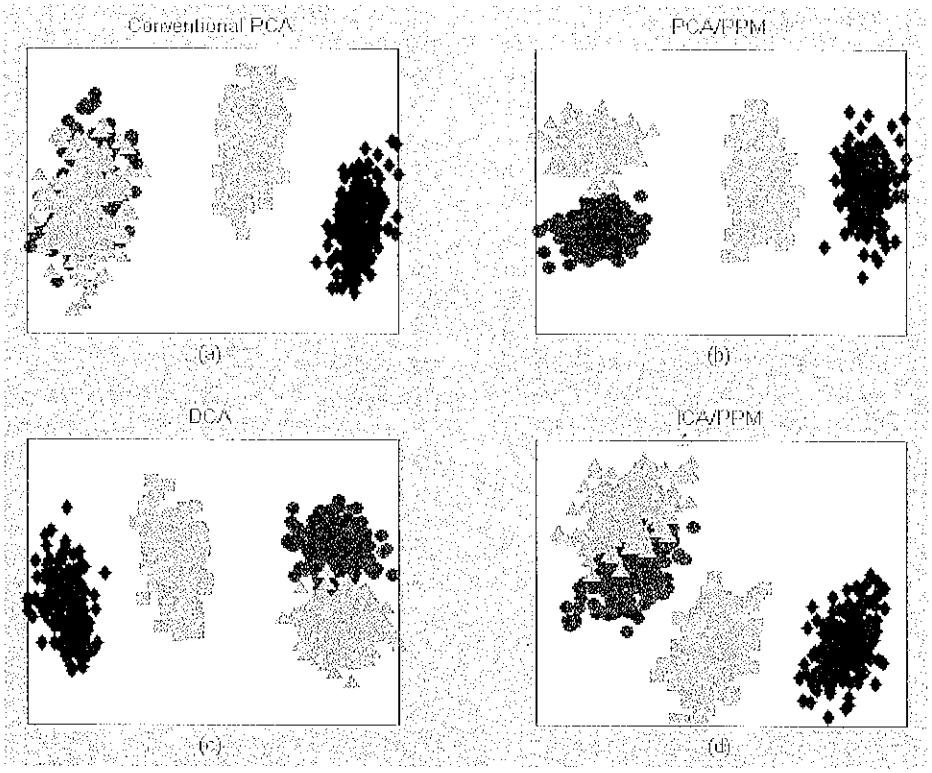


Figure 1. The 2-D projections of the simulation data resulting from conventional PCA, PCA PPM, DCA, and ICA-PPM

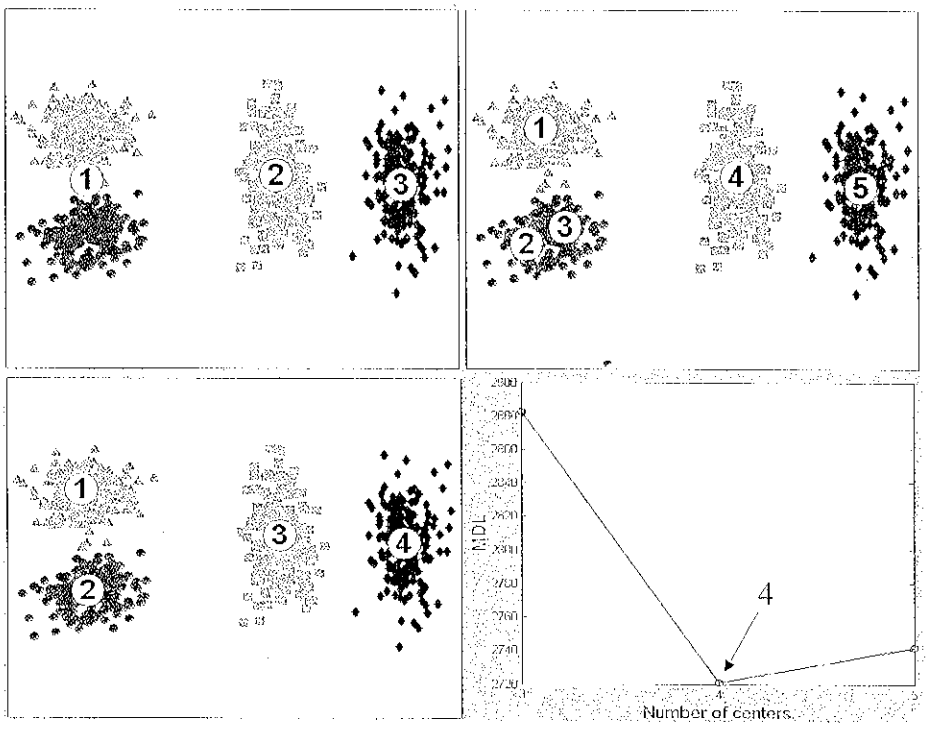


Figure 2 The model selection and the MDL curve of the simulated data at the top level projection.

structure is revealed by using DCA, PCA-PPM, and ICA-PPM comparing to the conventional PCA. The four cluster structure is clearly shown in (b), (c), and (d) in Fig. 1, but only three clusters are seen in the PCA plot (a) without incorporating symbol and color information in all plots. PCA, PCA-PPM and ICA-PPM are unsupervised processes and do not rely on the known class information; only the DCA method is supervised here. The class information is used only to show the four distinct classes with four different colors and symbols.

The model selection of simulated data at the top level projection generated by DCA is performed, and the results showed that a four-cluster structure may best describe the data distribution on this level. In the Fig. 2, three different model selection patterns are tested and the MDL curve is plotted; the MDL suggests that the four-cluster structure is optimal. Note that the numbers in the figures are the results of the cluster initialization by the user.

A hierarchical visualization trees, as shown in Fig. 3(a), is generated on the simulated data set. The top figure is a top level projection of the complete data set, where we can only see three clusters, the middle figure is a second level projection that provides individual different views of the three sub-clusters selected in the top level projection. In the second level, we can

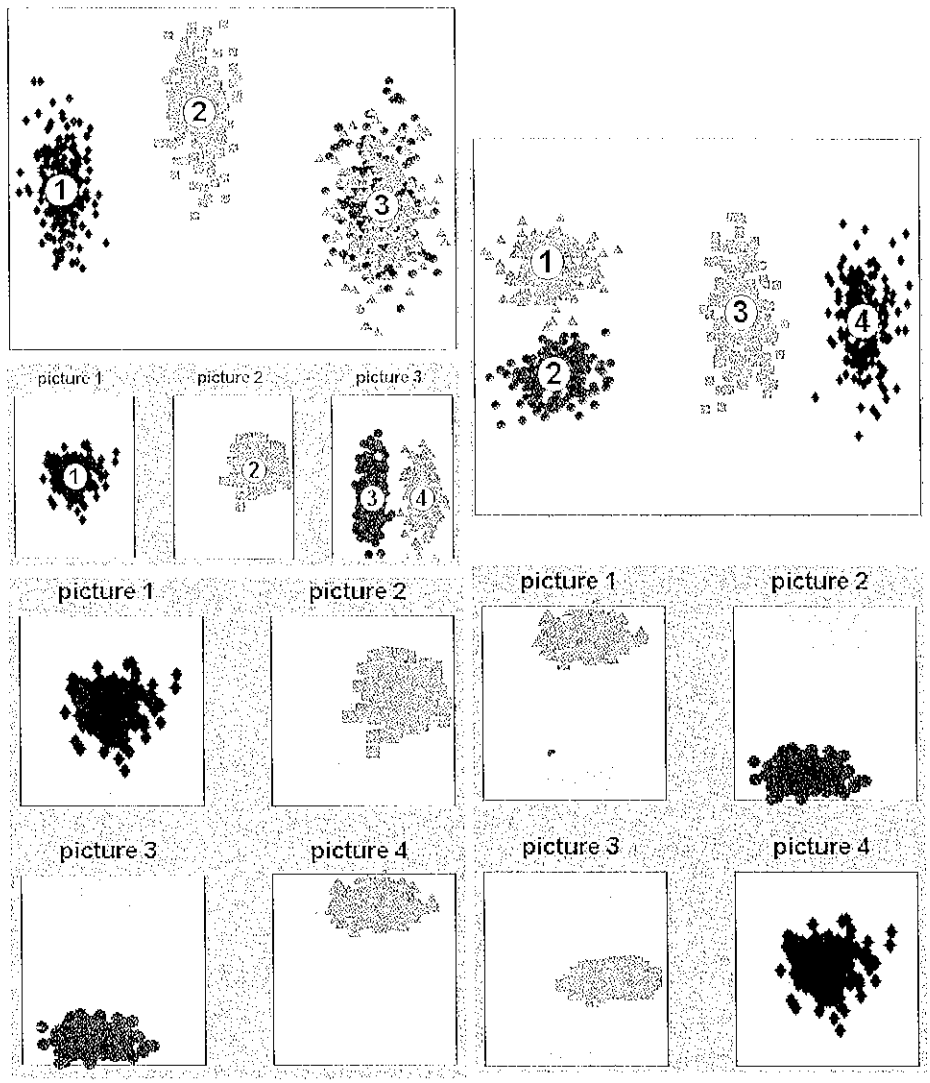


Figure 3 The hierarchical data exploration on the simulated data. The left figure (a) is generated by using PCA, and right figure (b) are got by using DCA.

see two hidden clusters in sub-cluster #3, this gives the user opportunities to discover more information about data structure, and it makes further partitioning possible. Clusters are partitioned and shown in their own windows in the bottom figure. In this experiment, only conventional PCA is used to generate all projections.

To illustrate a joint way of the discovery data structure by combining PCA, DCA, PCA-PPM and ICA-PPM, we also tested PCA-PPM, DCA, and ICA-PPM in the generation of a hierarchical visualization tree. In this case, since conventional PCA is unable to locate the directions in the space where real cluster structure can be displayed, we can use PCA-PPM, DCA, and ICA-PPM as alternatives. When DCA is used to plot the top level projection in Fig. 3(b), more information about the cluster structure is revealed after the first step, i.e., the directions to show the four-cluster structure are found in the space by DCA. For the simulated data set, the hierarchical visualization by DCA is used as an illustration because DCA, PCA-PPM and ICA-PPM are similar to one another. The two hierarchical trees in Fig. 3 are produced independently. The consistency of the clustering results and the known class grouping can be seen from the symbols and colors of the data points in each individual window on the bottom level. The fact that the data points within one cluster have the same symbol and color implies that the data points belonging to the same phenotype group are grouped into one cluster.

Besides testing the methodology on the simulated data, we also evaluate actual microarray data sets from National Cancer Institute (NCI) and Massachusetts Institute of Technology (MIT). In the 2308 dimensional (genes) microarray data sets of round blue cell tumors from NCI, there are four classes: neuroblastoma (NB; $N = 12$), rhabdomyosarcoma (RMS; $N = 21$), non-Hodgkin lymphoma (NHL; $N = 8$) and the Ewing family of tumors (EWS; $N = 23$). Figure 4 illustrates how all projection methods (PCA, PCA-PPM, DCA, and ICA-PPM) on NCI data explore cluster structure. According to our experience, these four methods may show inconsistent capabilities of finding directions for discovering cluster structure in various cases when the number of data points is limited. Thus, using all applications to examine the data structure are more informative than one or two methods combined. As discussed above, the known phenotype information has been only used for representing data points

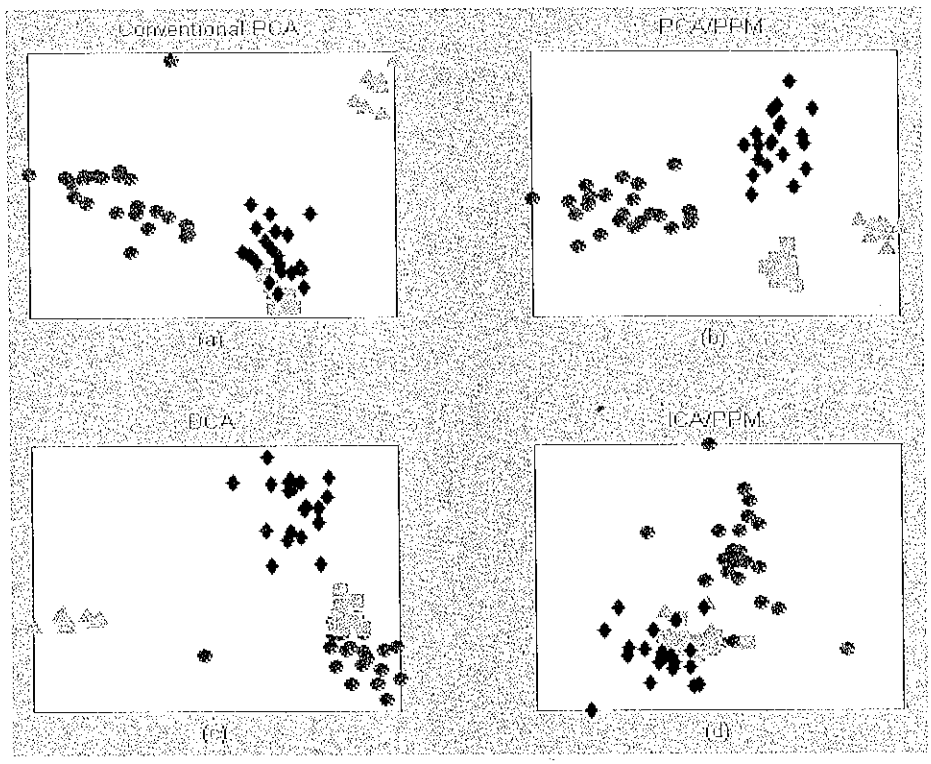


Figure 4 The 2-D projections of NCI data resulting from PCA, PCA-PPM, DCA, and ICA-PPM

from different classes in different symbols and colors in PCA, PCA-PPM and ICA-PPM experiments, rather than in projection searching; only the DCA experiment is supervised.

PCA and PCA-PPM on the NCI data generate hierarchical visualization trees (Fig. 5). Using PCA to generate 2-D projections produces the left hierarchical tree Fig. 5(a), and applying the PCA-PPM generates the right tree Fig. 5(b). MDL curves are also plotted for the two different top level projections. The curves indicate that the three-cluster from PCA and the four-cluster structures from PCA-PPM are best in describing the data distribution at the top level projections. The number of clusters determined by MDL agrees with the user visual inspection. As discussed in the hierarchical exploration experiment on the simulated data, the consistency of the clustering and the known biological phenotypes defined above is shown by the unified color and symbol of data points within each cluster. The clustering scheme

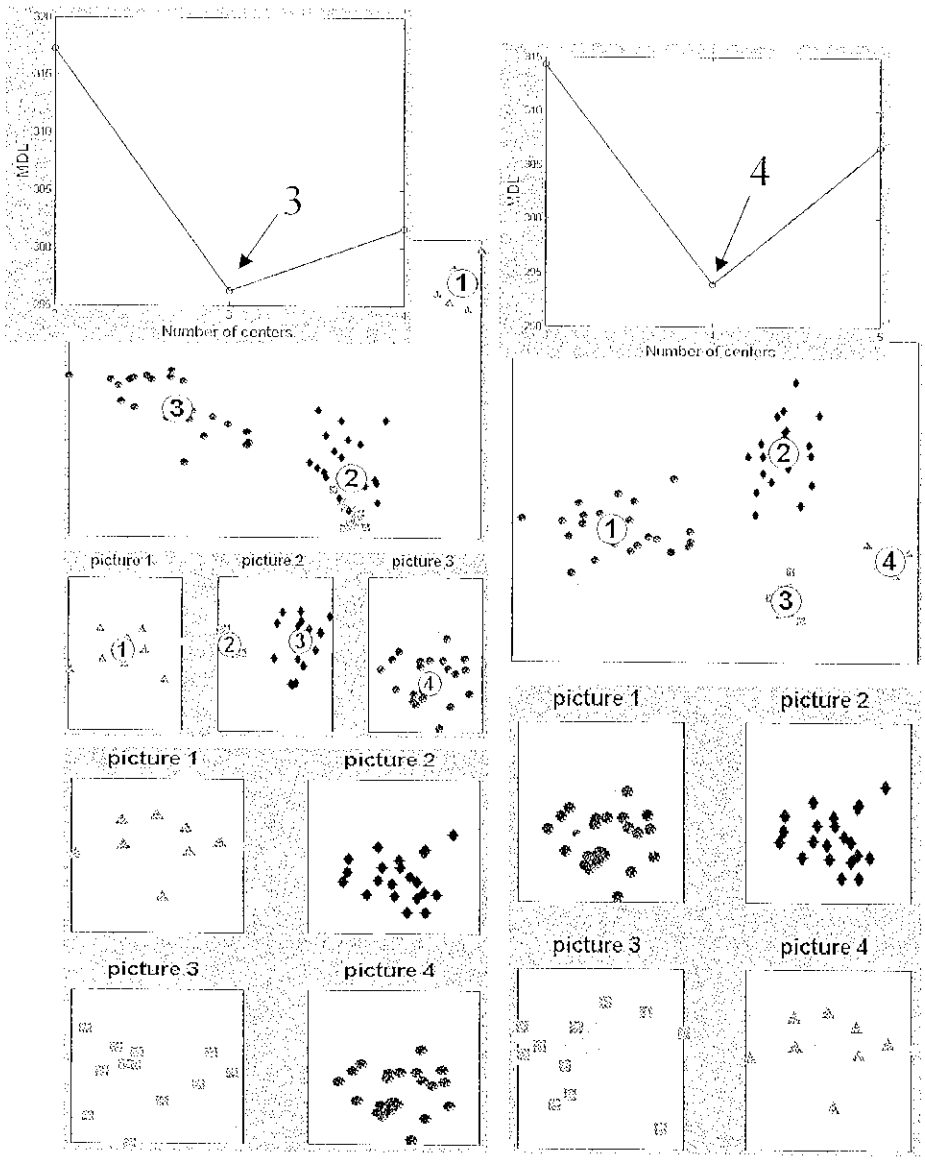


Figure 5 The hierarchical data exploration on NCI data. The left figure (a) is generated by using PCA, and right figure (b) is generated by using PCA-PPM. MDL curves for the top level projections are also included.

recommended by the MDL measure also indicates the consistency of the biological phenotype information and the final clustering in the bottom level of Fig. 5. In Fig. 5(a), although MDL cannot find the four-cluster structure on the top level, the projections on the second level provide a good opportunity for discovering two clusters within sub-cluster #2 (mixed square and diamond) on the top level. This also demonstrates the advantages of hierarchical visual exploration scheme on the cluster discovery. MDL recommends a four cluster structure on the top level of Fig. 5(b), the well partitioned four clusters on the bottom level show the ability of MDL to the finding the true cluster structure.

In studying leukemia, MIT published a 7129 dimensional microarray data sets that contain two classes, acute lymphoblastic leukemia (ALL; $N = 47$) and acute myeloid leukemia (AML; $N = 25$). The 2-D projections of the MIT data are presented in Fig. 6. All four projections on the leukemia data sets explore the data structure similarly to the other cases (simulated and NCI data sets).

DCA is meaningful with the model selection and the cluster partitioning even under unsupervised condition, i.e. no class information is known. We demonstrate this idea by combining PCA and DCA dynamically at the top level projection. In Fig. 7, the top level projection in the top figure is generated by PCA (unsupervised analysis) without knowing any class information. After model selection and partitioning (provide class information), DCA can now produce the re-projection of the data (middle figure). The bottom figure is a partition of the re-projected data. Even though the PCA projection (top figure) can be directly partitioned into sub-clusters (bottom figure), the additional step using DCA (middle figure) provides another chance to visualize the complete data set from different angles in which cluster structure is emphasized. As in this example, this data projection scenario is especially useful when cluster structure is ambiguous to the user in one projection, but becomes clear after DCA is applied based on the user model selection and cluster partitioning. From the clustering results shown on the bottom level of Fig. 7, we can conclude that the clustering is consistent with the biological phenotype information, since only a few mixed groupings occur.

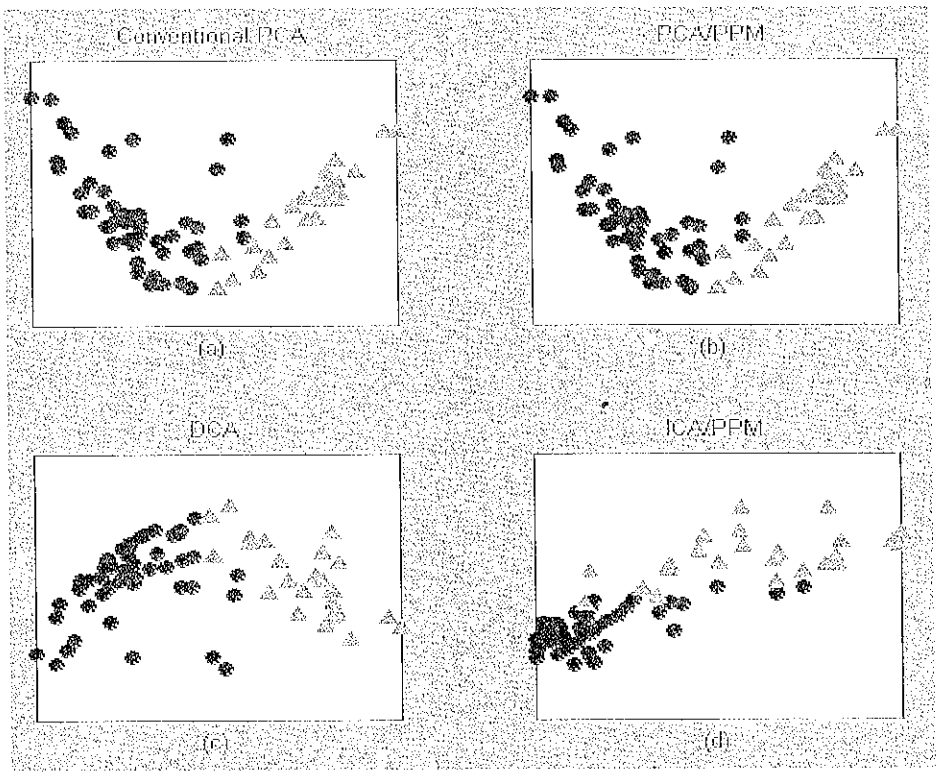


Figure 6 The 2-D projections of MIT data resulting from PCA, PCA-PPM, DCA, and ICA-PPM

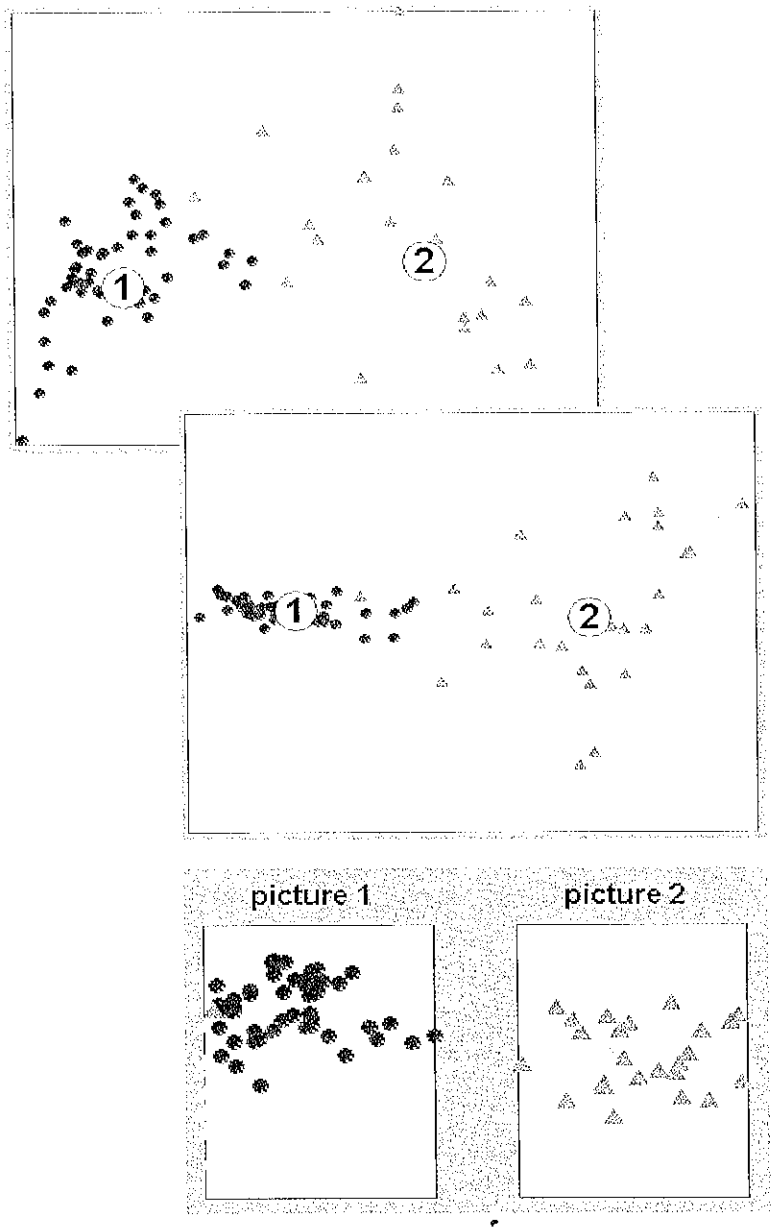


Figure 7 The dynamic combination of PCA and DCA in the exploration of MIT leukemia microarray data

5. Discussion

The growing volume of high dimensional and multimodal data sets demands a data mining tool, different from conventional data visualization methods, which is capable of dealing with high dimensional data. The hierarchical visualization paradigm involving hierarchical statistical models and visualization space is proven to be able to effectively discover data structure and capture all interesting aspects of the data. Using several complementary visualization subspaces makes this complicated task feasible. The strategy of the hierarchical data exploration and mining tool used for cluster discovery is that the top level model and projection should explain as much structure information of the entire data set as possible, while lower level models explain the local and internal structure between individual cluster, which may not be

obvious in the high level models. With several complementary mixture models and visualization projections, each level will be relatively simple while the complete hierarchy maintains overall flexibility yet conveys considerable cluster information. In this algorithm, dimensionality reduction and cluster decomposition are two major components. Dimensionality reduction allows visualization of high dimensional data and less computational demand. Cluster decomposition provides relatively simple models by partitioning large and complicated mixture models into small local structure, which offers great ease of interpretation and many benefits of analytical and computational simplification.

The techniques involved are statistical modeling of the high dimensional data with SFNM distribution, 2-D data projection jointly presented by an unsupervised and supervised data mining scheme, and evaluation of cluster structure produced by such scheme using microarray experiments with known phenotypes. Unlike conventional PCA, PCA-PPM, DCA and ICA-PPM project the data set into 2-D visual subspaces, which allows the data set to be discriminatively explored so that cluster structure is effectively revealed. Furthermore, IEM procedure are implemented to probabilistically estimate SFNM distribution. With the model-based approach, a model selection procedure is used to determine the number of sub-clusters within each cluster using the minimum description length criterion. This approach allows the algorithm to determine automatically whether a further split of a subspace should continue in completing the whole hierarchy [10]. User interaction with the algorithm is also an important issue. The user-friendly graphical interface facilitates the data visualization purpose, which allows the user to select initial centers of the data clusters. The initialization for the clustering procedure made by a user and further validated by MDL is usually better than a random initialization based on our experience, and the reduction of both computational complexity and local optimum likelihood is expected and the outcomes of our experiments do agree with the expectation, i.e., less iterations in IEM are needed for its convergence to the correct clustering results that are shown in the experiments in Results Section. While the final SFNM model can be estimated, the pathways of achieving cluster decomposition may be multiple. This user-driven nature of the current algorithm is also highly appropriate for the visualization context. With the advantages discussed above, our data mining algorithm can explore data structure in great extents with no standard data analyses may compare.

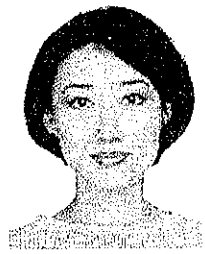
Acknowledgments

The authors wish to thank Dr. T. R. Golub and his group of Whitehead Institute, MIT, for the leukemia microarray data published in their web site.

References

1. D. J. Duggan, M. L. Bittner, Y. Chen, P. Meltzer, and J. M. Trent, "Expression Profiling Using cDNA Microarrays," *Nature Genetics*, vol. 21, 1999, pp. 10–14.
2. U. Scherf, D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, W. C. Reinhold, T. G. Myers, D. T. Andrews, D. A. Scudiero, M. B. Eisen, E. A. Sausville, Y. Pommier, D. Botstein, P. O. Brown, and J. N. Weinstein, "A Gene Expression Database for the Molecular Pharmacology of Cancer," *Nature Genetics*, vol. 24, 2000, pp. 236–244.
3. M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent, "Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling," *Nature*, vol. 406, no. 3, 2000, pp. 536–540.
4. H. Zhang, C.-Y. Yu, B. Singer, and M. Xiong, "Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data," *Proc. Natl. Acad. Sci.*, vol. 98, no. 12, 2001, pp. 6730–6735.
5. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, 1999, pp. 531–537.
6. J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine*, vol. 7, no. 6, 2001, pp. 673–679.
7. P. Tamayo, D. Slonim, J. Mesirov et al., "Interpreting Pattern of Gene Expression with Self Organizing Maps: Methods and Application to Hematopoietic Differentiation," *Proc. Natl. Acad. Sci.*, vol. 96, 1999, pp. 2907–2912.
8. E. Hartuv, A. O. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach, and R. Shamir, "An Algorithm for Clustering cDNA Fingerprints," *Genomics*, vol. 66, 2000, pp. 249–256.
9. A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support Vector Clustering," *J. Machine Learning Research*, vol. 2, 2001, pp. 125–137.

10 Y Wang, J Luo, M J Freedman, and S -Y Kung, "Probabilistic Principal Component Subspaces: A Hierarchical Finite Mixture Model for Data Visualization," *IEEE Trans Neural Nets*, vol 11, no 3, 2000, pp 625–636
11 Y Wang, J Lu, and Z Wang et al , "Discriminative Mining of Gene Microarray Data," in *Proc of IEEE Neural Network for Signal Processing Workshop*, Sept 2001, pp 23–32
12 S T Roweis and L K Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol 290, 2000, pp 2323–2326
13 R Lotlikar and R Kothari, "Fractional-Step Dimensionality Reduction," *IEEE Trans Pattern Anal Machine Intell*, vol 22, no 6, 2000, pp 623–627
14 G E Hinton, P Dayan, and M Revow, "Modeling the Manifolds of Images of Handwritten Digits," *IEEE Trans Neural Net*, vol 8, no 1, 1997, pp 65–74
15 N Kambhatla and T K Leen, "Dimension Reduction by Local Principal Component Analysis," *Neural Computation*, vol 9, no 7, 1997, pp 1493–1516
16 M E Tipping and C M Bishop, "Mixtures of Probabilistic Principal Component Analyzers," *Neural Computation*, vol 11, 1999, pp 443–482
17 C M Bishop and M E Tipping, "A Hierarchical Latent Variable Model for Data Visualization," *IEEE Trans Pattern Anal Machine Intell*, vol 20, no 3, 1998, pp 282–293
18 S Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed , Upper Saddle River, New Jersey: Prentice-Hall, Inc , 1999
19 D M Titterington, A F M Smith, and U E Markov, *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley, 1985
20 E Mjolsness and D DeCoste, "Machine Learning for Science: State of the Art and Future Prospects," *Science*, vol 293, 2001, pp 2051–2055
21 J Rissanen, "Modeling by Shortest Data Description," *Automatica*, vol 14, 1978, pp 465–471
22 A K Jain, R P W Duin, and J Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans Pattern Anal Machine Intell*, vol 22, no 1, 2000, pp 4–37
23 J H Friedman, "Exploratory Projection Pursuit," *J Amc Stat Asso*, vol 82, no 397, 1987, pp 249–266
24 A Hyvarinen and E Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, vol 13, 2000, pp 411–430
25 B Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996
26 Y Wang, S -H Lin, H Li, and S -Y Kung, "Data Mapping by Probabilistic Modular Networks and Information-Theoretic Criteria," *IEEE Trans Signal Processing*, vol 46, no 12, 1998, pp 3378–3397
27 K Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed , New York: Academic Press, 1990
28 S -Y Kung, *Principal Component Neural Network*, New York: Wiley, 1996
29 R N Bracewell, *Two-Dimensional Imaging*, Prentice-Hall, Inc , 1995



Znyi Wang received the Ph D degree in biomedical engineering at the Catholic University of America (CUA), Washington, DC in 2003 She received her M S degree from the Department of Electrical Engineering, CUA She is currently working at Department of Electrical Engineering, CUA, and Center for Genetic Medicine Research, Children National Medical Center Her research interests are in computational bioinformatics and medical imaging
zwang@cumcresearch org; zwang@pluto ee cua edu

Yue Wang received the Ph D degree in electrical engineering from the University of Maryland in 1995 He is currently an Associate Professor of Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Alexandria, VA He is also affiliated with the Johns Hopkins Medical Institutions as an Adjunct Professor of Radiology His recent research focuses on computational bioinformatics and molecular imaging
yuewang@vt edu

Jianping Lu is a visiting researcher at The Catholic University of America, Washington DC, and an Associate Professor of Computer Science at SuZhou University, Su Zhou, China His recent research focuses on bioinformatics, image processing, network, and database lu@pluto ee.cua edu

Sun-Yuan Kung (F'88) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

In 1974, he was an Associate Engineer of Amdahl Corporation, Sunnyvale, CA. From 1977 to 1987, he was a Professor of Electrical Engineering-Systems, University of Southern California, Los Angeles. Since 1987, he has been a Professor of Electrical Engineering at Princeton University, Princeton, NJ. He has authored more than 300 technical publications, including three books: *VLSI Array Processors* (Englewood Cliffs, NJ: Prentice-Hall, 1988) (with Russian and Chinese translations), *Digital Neural Networks* (Englewood Cliffs, NJ: Prentice-Hall, 1993), and *Principal Component Neural Networks* (New York: Wiley, 1996).

Dr. Kung received the 1992 IEEE Signal Processing Society's Technical Achievement Award for his contributions on parallel processing and neural network algorithms for signal processing. Since 1990, he has served as Editor-in-Chief of the *Journal of Signal Processing*. Recently, he served as a General Chair of the 1997 IEEE Workshop on Multimedia Signal Processing at Princeton University. He was appointed as IEEE-SP Distinguished Lecturer in 1994. He received the 1996 IEEE Signal Processing Society's Best Paper Award.

kung@ee.princeton.edu

Junying Zhang received the Ph.D. degree in the national key lab on radar signal processing from the Xidian University, Xi'an, P.R. China, in 1998. She is currently a professor of Computer Science at Xidian University, Xi'an, P.R. China. She is also affiliated with the national key lab on radar signal processing as a professor and researcher, and is currently a visiting scholar of Electrical Engineering at The Catholic University of America, Washington, DC. Her recent research focuses on computational bioinformatics and molecular imaging, feature selection and pattern recognition.

jyzhang@pluto.ee.cua.edu

Richard Lee received the Ph.D. degree from the Department of Physiology and Biophysics, Georgetown University, Washington, DC, in 2001. He is currently a postdoctoral fellow at the Lombardi Cancer Center to study the mechanisms of drug resistance in breast cancer using computational bioinformatics.

leer@georgetown.edu

Jianhua Xuan received his B.S. and M.S. degrees in Electrical Engineering from Zhejiang University in 1985 and 1988, and his Ph.D. degree in Electrical Engineering and Computer Science from the University of Maryland Baltimore County in 1997. He is currently an Assistant Professor of Computer Science at the Catholic University of America. His research interests include bioinformatics, medical imaging, computer vision, computer graphics and visualization.

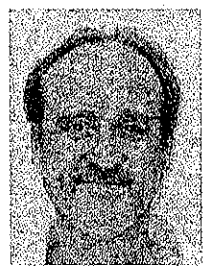
xuan@cua.edu

Javed Khan gained his Bachelor's degree at the University of Cambridge in England in 1984 in Immunolology and Parasitolgy. He subsequently obtained his Masters degree in 1998 in the same subjects. He went onto the school of clinical medicine, Cambridge University where he completed Doctor of Medicine (MD) degree. He has obtained the postgraduate degree of MRCP (Membership of the Royal College of Physicians), which is equivalent

to the Boards examination in the USA. From 1995–1998 Dr. Khan competed a Hematology & Oncology Fellowship at the Pediatric Oncology Branch, National Cancer Institute, NIH. In 1996 as part of the fellowship he moved to the Cancer Genetics Branch, National Human Genome Research Institute (NHGRI) and became involved in the pioneering work of developing cDNA microarrays for cancer research. During his five years at the NHGRI Dr Khan was recognized for his exceptional abilities by receiving both a Merit Award as well as an Outstanding Oral Presentation Award.

The initial paper using this cDNA microarray technology by Dr Khan (Cancer research, vol 58, 5009–5013, 1998) demonstrated that it can be used to identify genetic fingerprints of a specific type of muscle cancer, rhabdomyosarcoma (RMS) which is able to distinguish one type of cancer from another and was the first to apply hierarchical clustering and visualization tools including multidimensional scaling to demonstrate the relationships between cancers based on gene expression profiling. In April 2001 Dr. Khan was recognized by the American Association for Cancer Research for his pioneering work is tumor profiling by receiving a "Scholar in Training Award". In May 2001 Dr. Khan joined the Pediatric Branch, NCI as a tenure-track investigator having been selected among a distinguished applicant pool. During this transition to the NCI Dr. Khan and colleagues published a new model for diagnosis of cancer using artificial neural networks (ANN), a form of artificial intelligence, and microarray technology (Nature Medicine, vol. 7,6: 673–679) ANN were used to decipher gene-expression signatures collected with DNA microarrays and to classify cancers into specific categories.

khanjav@mail.nih.gov



Robert Clarke received the Ph.D. and D.Sc. degrees in biochemistry from the Queen's University of Belfast, U.K., in 1986 and 1999, respectively.

He is currently a Professor in oncology, physiology, and biophysics at Georgetown University, Washington, DC. His research interests focus on studies into the molecular biology and endocrinology of breast cancer.

Dr. Clarke is a Fellow of the Royal Society of Chemistry (U.K.), Royal Society of Medicine (U.K.), and Royal Institute of Biology (U.K.).

clarker@georgetown.edu