Korean J. Chem. Eng., **26**(4), 969-979 (2009)
DOI: 10.2478/s11814-009-0161-3

Mapping multi-class cancers and clinical outcomes prediction for multiple classifications of microarray gene expression data

Yong Su Kim, Sun Jin Hwang, Jong Min Oh, Gye Dae Whang, and Chang Kyoo Yoo†

Center for Environmental Studies, Department of Environmental Science and Engineering, College of Engineering,
Kyung Hee University, Suwon 446-701, Korea

(*Received 5 January 2009 • accepted 1 February 2009*)

Abstract−DNA microarray analysis of gene expression is useful for discriminating between the various subtypes of cancer, which is necessary for the accurate diagnosis and treatment of patients. Particularly, assigning biological samples into subclasses or obtaining detailed phenotypes is an important practical application for microarray gene expression profiles. In the present study, a hierarchical framework of a nonlinear mapping classification was developed for elucidating data and classifying multiclass cancers based on microarray data sets. This classification maps the gene expression profiles of multi-class cancers to the visualized latent space and predicts the clinical output through high-dimensional computational biology. The proposed method was used to interpret and analyze four leukemia subtypes from microarray data. The results demonstrate that, using a high-dimensional nonlinear mapping to extract biological insights from microarray data, the proposed method can identify leukemia subtypes on the basis of molecular-level monitoring and improve the interpretability of leukemia clinical outputs. Furthermore, this nonlinear mapping of cancer subtypes is used to establish a relationship between expression-based subclasses of leukemia tumors and leukemia patient treatment outcomes. The proposed method may be used to guide efficient and effective approaches for the treatment of leukemia subclasses.

Key words: Bioinformatics, Cancer Classification, Clinical Outcome, Hierarchical Framework, Generative Topographic Mapping, Microarray Gene Expression

INTRODUCTION

DNA microarray technologies, which simultaneously monitor the expression pattern of thousands of genes, have resulted in a tremendous increase in the amount of available gene expression data. There is a great need to interpret, visualize, and analyze the information contained in gene expression profiles. However, gene expression data are characterized by very high dimensionality (number of genes), relatively small number of samples (observations), irrelevant features, and collinear and multivariate characteristics. Thus, comprehensible interpretation and analysis of gene expression profiles is difficult and entails a high computational cost. Typically, the first step in solving the difficulties in interpreting gene expression profiles consists of extracting the fundamental genes (or features) of the gene expression data, resulting in a dimensionality reduction of the original data set. In a second step, the selected genes are analyzed with computational biology (bioinformatics) tools. On the other hand, classifying biological samples into known classes or phenotypes is an important practical application for microarray gene expression profiles. For example, gene expression profiles obtained from patient tissue samples allow for the classification of cancers [1-6].

Many clustering and classification methods for gene expression profiles have recently been applied to cancer classifications for colon and breast cancer, for leukemia and for other tumors [7-18]. These investigations have clearly shown the capability of gene expression profiling for classifying tumors. Gene expression profiles may give more objective information than traditional morphological tumor characterization methods. For example, Lu and Han [13] provide a detailed review on methodologies that are commonly used for classification. A number of relevant papers and methodologies are mentioned to illustrate that considerable research has been conducted in the area, focusing on the use of a range of different algorithms for the classification of gene expression data. Golub et al. [14] used a weighted voting scheme for molecular classification of acute leukemia subtypes. This scheme predicts leukemia subtypes by means of a supervised learning algorithm and discriminant decision rules that were derived on the basis of the magnitude and threshold of the prediction strength. It provided strong evidence that gene expression profiles can be used for cancer classification. Among the researchers of microarray mapping, Tamayo et al. [15] and Toronen et al. [16] applied self-organizing maps (SOM) to analyze and visualize gene expression data. Alizadeh et al. [17] studied gene expression in the three most prevalent adult lymphoid malignancies and identified two previously unrecognized types of diffuse large B-cell lymphoma that exhibited a distinct clinical behavior. Average linkage hierarchical clustering was applied to identify the two tumor subclasses as well as to group genes with similar expression patterns across the different samples.

Recently, self-organized mapping (SOM) and generative topographic mapping (GTM) have been used to visualize the high-dimensional data [18-20]. The GTM model assumes a latent variable model and selects each node with some probability for each input (gene expression profile) to estimate the model. The latent variables are set to fixed grid points and the corresponding winning probabili-

†To whom correspondence should be addressed.
E-mail: ckyoo@khu.ac.kr

ties are learned. That is, the probabilities are learned by using the latent variable model, which generates the co-regulation expression features from a given latent space. The expression profiles of many genes can be expressed in a more compact form by latent features and can provide useful insights into understanding the data [3]; while PCA can introduce in the visualization only "global" stretching along the principal axes, the nonlinear projection manifold of GTMs can locally stretch and fold in the data space. GTM represents the whole data set in low dimensions without loss of information. This enables the complex systems of biosystems or gene expression microarray data to make full use of the latent space when describing the local distributions of data [21].

However, when dealing with large and complex data sets such as microarray data, a single global classification model is often not sufficient to get a good understanding of the relationships in the data. Moreover, the standard mapping algorithms tend to perform poorly when there is a great deal of noise and the data is not split into well-separated clusters, such as in microarray data or biosystems [22]. To tackle this problem, a hierarchical framework for mapping microarray data from multi-class cancer samples to nonlinear latent space is proposed in this research. The hierarchical approach allows users to make an intensive investigation into several interest regions and find out more about the data as well as the clusters.

The outline of this paper is as follows: first, current generative topographic mapping (GTM) is briefly presented. Second, a hybrid framework of topographical local mapping of the microarray gene expression data set is proposed. Third, the proposed method is applied to locally cluster the data and predict the unknown samples in a leukemia cancer microarray. Finally, the conclusion of this work is given.

CANCER CLASSIFICATION BY HIERARCHICAL GENERATIVE TOPOGRAPHICAL MAPPING

The expression level of a gene is usually quantified based on the approximate number of copies of that gene's RNA present in a cell, and is assumed to be correlated with the production rate of the corresponding protein. Thus, the expression level of a gene provides a measure of the activity (or the transcription rate) of that gene under certain biochemical conditions. The occurrence of certain diseases such as cancer will be reflected as a change of the expression level of the genes that are affected by the disease. Specifically for cancer, normal cells can evolve into malignant cancer cells through a series of mutations in genes that control the cell cycle, apoptosis, and ge-

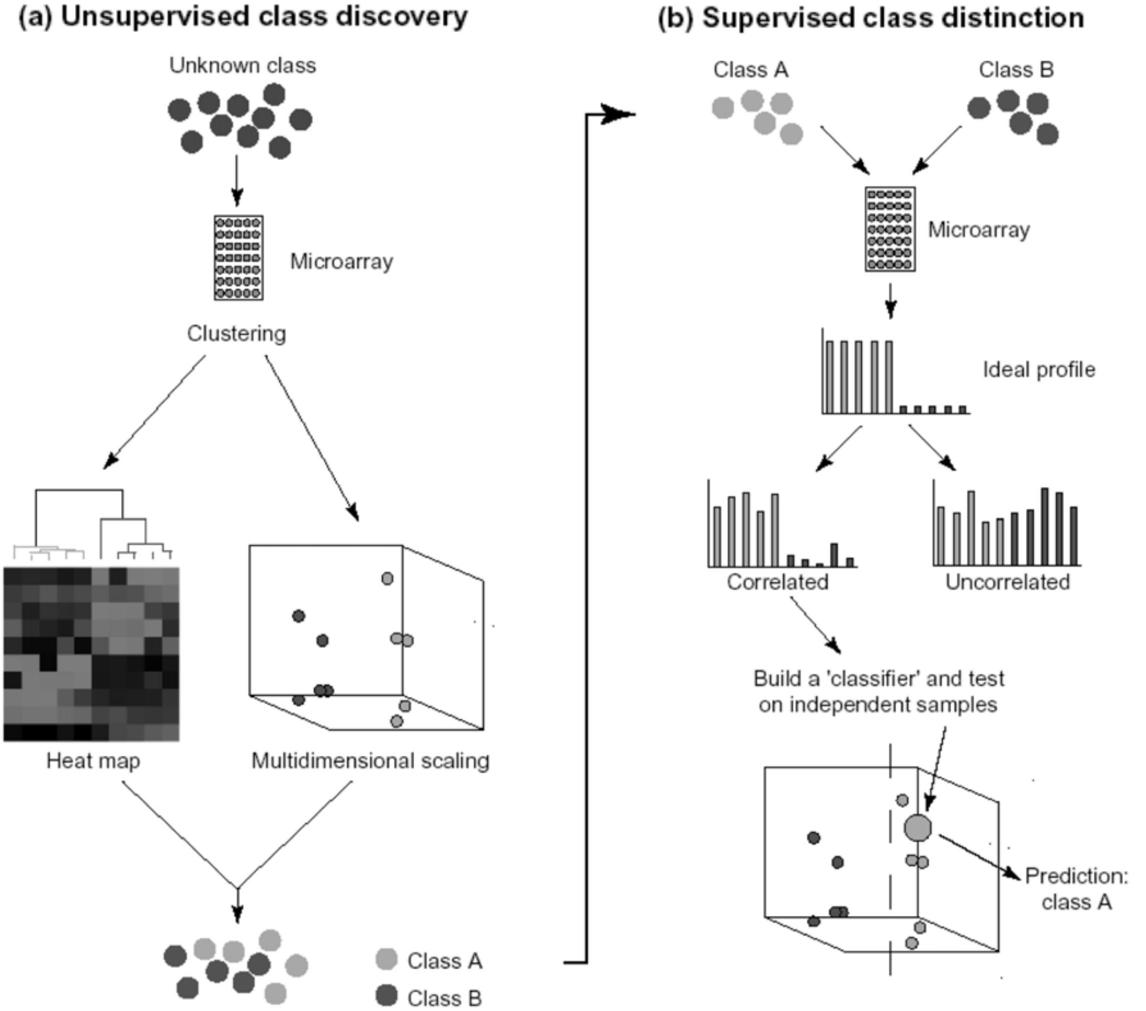

Fig. 1. Overview of cancer discovery and classification in microarray data analysis [1].

nome integrity [13].

The general principles of unsupervised class discovery and supervised class distinction are illustrated in Fig. 1 as applied to microarray data analysis. In the unsupervised class discovery approach, RNA from tumor samples is profiled for expressed genes on a microarray (either cDNA- or oligonucleotide-based), and the individual samples are clustered on the basis of similarities in expression patterns. Clustering results can be visualized in many ways. One method is to use 2-dimensional or hierarchical clustering 'heat maps', in which the gene expression levels are represented by colors: red signifies a high relative expression of a gene, and green signifies a low relative expression. The similarity among tumor samples can be evaluated using a dendrogram, in which the distance between the branches and sub-branches illustrates the relative similarity. The relative similarity among samples can also be represented in three-dimensional space, either by multidimensional scaling, where the expression levels of each gene dictate the position of the samples in space, or via principal component analysis (PCA) where, for example, the complexity of gene expression is collapsed into so-called supergenes. This might, for example, lead to the conclusion that there are several molecular classes among the tumor samples analyzed (class A and class B in the example in Fig. 1). In a supervised approach, the classification of tumor samples and the ideal profile that would distinguish between tumor sample classes are known *a priori* (e.g., high expression of a specific gene for one tumor sample class, low expression for another one). One or more methods are then used to rank the genes of interest according to the ideal profile (i.e., to identify those genes whose expression is correlated with the class distinction), and a subset of these genes is used for building a classifier to predict the class of a test sample. The processes used to infer the existence of a new classification of tumors in Fig. 1(a) can be verified by the supervised methods in Fig. 1(b) [1].

1. Generative Topographic Mapping (GTM)

Generative topographic mapping (GTM) is a probability density model which describes the distribution of data in a space of several dimensions in terms of a smaller number of latent (or hidden) variables. Mathematically, principal component analysis (PCA) is based on a linear transformation from latent space to data space, but GTM is extended to allow non-linear transformation and is based on a constrained mixture of Gaussians whose parameters can be optimized by using the EM (expectation-maximization) algorithm. The latent space is used to visualize the data by introducing a regular array of latent space centers in the two-dimensional interval [-1,1] [-1,1]. Fig. 2(a) shows the conceptual diagram of GTM by Bishop et al. [20]. It uses a nonlinear relationship between the latent space {**u**} and the data space {**x**}. It can be fitted to a data set {**xn**}, where n=1, 2, …, N, by a non-linear latent variable with a set of M fixed non-linear basis functions ϕ(u)={ϕ_j(u)}. The centers of non-linear basis functions are positioned in the latent space on a regular grid [19,20].

Data space variables are denoted in **x**=[x1, x2, …, xD]T, with gene expression profile dataset, and latent space variables in **u**=[u1, u2, …, uL]T, where a regular array of nodes, i=1, …, K, within the latent space. The GTM defines a nonlinear, parametric mapping **y**(**u**, **W**) from a latent space (**u**∈RL) to a data space (**x**∈Rp) where normally L<D. Points in the low dimensional latent space are mapped to corresponding centers of a Gaussian in the observable high-dimensional data space. The probability of data over the latent space p(**x**|**u**) will induce a probability over the data space, p(**x**|**u**, **W**, β) and is defined as a Gaussian distribution:

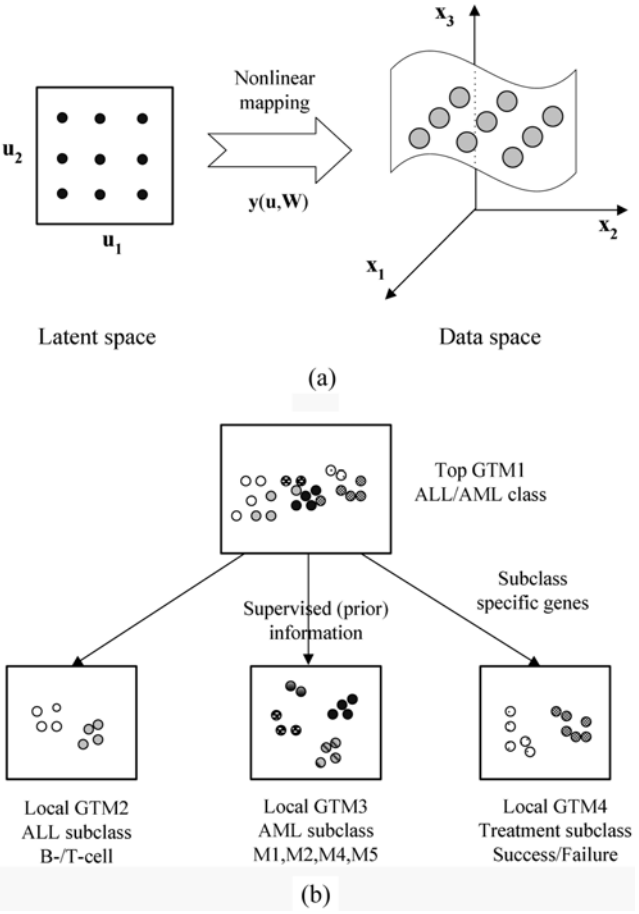

Fig. 2. (a) The nonlinear mapping of the GTM [19], and (b) the proposed hierarchical scheme for mapping multi-class cancer classification.

$$p(\mathbf{x}|\mathbf{u},\mathbf{W},\beta)=\left(\frac{\beta}{2\pi}\right)^{m/2}\exp\left[-\frac{\beta}{2}\|\mathbf{x}-\mathbf{y}(\mathbf{u},\mathbf{W},\beta)\|^2\right] \qquad (1)$$

where **x** is a point in the data space, **W** is a weight matrix and β–1 is the noise variance, which allows for some variance in the observed variables not explained by the latent variables. Both **W** and β–1 are estimated by an expectation-maximization (EM) algorithm [20,23]. To summarize, the GTM model can be regarded as a constrained mixture of Gaussians, as illustrated schematically in Fig. 2(a), in which the Gaussian components are isotropic with an inverse variance b and have centers given by an EM algorithm. A detailed derivation of the GTM can be found in the appendix.

From the topological view, note that the projected points ϕ(**u**; **W**) in the data space will necessarily have a topographic ordering in the sense that any two points **uA** and **uB** that are close in latent space will map to points ϕ(**uA**; **W**) and ϕ(**uB**; **W**), which are close in data space [18]. If a sample in latent space is close to a class sample which is a representative of a specific cancer type, it is probabilistically the specific class sample. That is, close samples in the latent space have the same type and the same subclass of leukemia in the input space. This means that close samples in the input space have close representations in the latent space since GTM can preserve a topology between input and latent space [21].

2. The Hierarchical Framework for Mapping Multi-class Cancer to Nonlinear Latent Space

To represent complex intrinsic information when visualizing large data sets, hierarchical visualization systems have been proposed and developed in the literature [19,22]. Several approaches are proposed to develop a model involving multiple two-dimensional visualization spaces. The reason for this approach is that the lack of flexibility of individual models can be compensated by the overall flexibility of the complete hierarchy [21].

When dealing with large and complex data sets such as microarray data, a single global model is often not sufficient to get a good understanding of the relationships in the data. For a microarray data set with multiple subtypes of cancer, it would be appropriate to use multiple models to capture the local variations of each specific cancer type. If microarray data corresponding to different classes exhibit dissimilarities due to different cancer subclasses, each local model can capture its operating region better than a global model, at the cost of poor characterization of the other modes. Also, gene selection is a fundamental issue in gene expression-based tumor classification. One of the challenging problems in DNA microarray data analysis is how to find a limited number of underlying genes which account for the multiple cancers. Here, the issue is to select the specific key gene sets appropriate for multiple cancers and multiple models.

In this research, to enhance the classification performance and reduce missed treatments in the microarray, a hierarchical framework for mapping multi-class cancer to nonlinear latent space is proposed. The hierarchical GTM is composed of a set of GTMs and their corresponding plots in a sequential structure. Synergistic effects of the supervised knowledge that is responsible for selecting the key genes are accounted for by sequentially using GTMs in a hierarchical mapping method.

Fig. 2(b) shows the proposed framework of local mapping using multiple hierarchical models. Proper gene selection is crucial for successful identification and interpretation of microarray gene expression data. First, the relevant genes for the interested subclasses of cancer are selected by using the DPLS model, which considers synergistic effects of the supervised knowledge that is responsible for selecting the key genes [9]. Second, the global models with *a priori* knowledge are constructed by a high-dimensional nonlinear mapping, here, the top GTM model. Third, local mapping models with *a priori* genes for the cancer subclasses are used to build the corresponding subprojection for classifying data into several clusters using *a posterior* probability. In this research, microarray data from four leukemia subtypes are classified: (1) acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), (2) ALL subtype (T-cell or B-cell), (3) AML subtype (M1, M2, M4, or M5), and (4) AML subtype by clinical outcome (success or failure). From these nonlinear mappings, each subtype of cancer is classified by mapping those data into multi-cancer regions and highlighting the regions of interest. Based on the *a posterior* probability of each local model, the local model that best represents a current operating condition is selected and used for the cancer classification. This can be accomplished by establishing a relationship between expression-based subclasses of leukemia tumors and leukemia patient treatment outcome by a hierarchical GTM approach.

RESULTS AND DISCUSSION

The proposed method is applied to the microarray data set of acute leukemia cancers published by Golub et al. [14]. The data set consists of a set of high density oligonucleotide microarrays (Affymetrix) with probes from 7129 human genes obtained from 72 patients. 47 patients were affected with ALL (38 B-ALL and 9 T-ALL), and 25 patients were affected with AML. The training data set consists of 38 bone marrow samples: 27 samples were taken from ALL patients (19 B-ALL and 8 T-ALL) and 11 were taken from AML patients. The independent (test) data set consisted of 34 samples: 20 ALL patients and 14 AML patients. Furthermore, a description of cancer subtypes, treatment response, patient gender, and the laboratory that performed the analysis was provided with the data. In addition, the results of the subsequent treatment (success or failure) are provided for a limited number of samples.

In this research, a nonlinear mapping for four subclasses is studied: (1) acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), (2) ALL subtype (T-cell or B-cell), (3) AML subtype (M1, M2, M4, or M5), and (4) AML subtype by clinical outcome (success or failure). The gene expression profiles of the original data set are represented as log10 normalized expression values, such that overall intensities for each chip are equivalent. To remove systematic sources of variation in the microarray experiments (i.e., different labeling efficiencies and scanning properties, print-tip or spatial effects, and different noise levels in each array), the expression level of each gene was normalized to have a zero mean and a standard deviation of one.

1. Mapping of ALL and AML Classes of Leukemia to Latent Space: The Top GTM Model

Proper gene selection is crucial for successful identification and interpretation of microarray gene expression data. The DPLS method suggested by Yoo et al. [9] is used to select the genes that are most suited to discriminate between AML and ALL. Out of the 7129 available genes in the expression data, 23 genes were selected for ALL or AML classification. Most genes in the 23-gene set have been identified previously as being abnormal in AML or ALL [14,24].

Fig. 3 shows a hierarchical visualization plot of heat map and class prediction by using four GTM models. It shows gene expression maps of leukemia subclasses, where the CLUSTER and TREEVIEW software was used, which are both publicly available at http://rana.lbl.gov. This analysis confirms that the heat map of the selected genes is significantly different for each subclass, and the local GTM for the selected genes can distinguish the subclass and predict unknown samples well. When building the GTM, a modified algorithm of GTM with a regularization term was used and the latent space of the GTM was chosen to have two dimensionality (q=2). The latent space centers, *ui*, were positioned on a regular 25×25 square grid, and 81 Gaussian basis functions were used with the width =1.0, and the regularization coefficient 1. The weight matrix of the GTM, **W**, was initialized by the principal components, and the initial value for β was calculated based on the noise of the data (the L+1th eigenvalue) and the interdistances between Gaussian

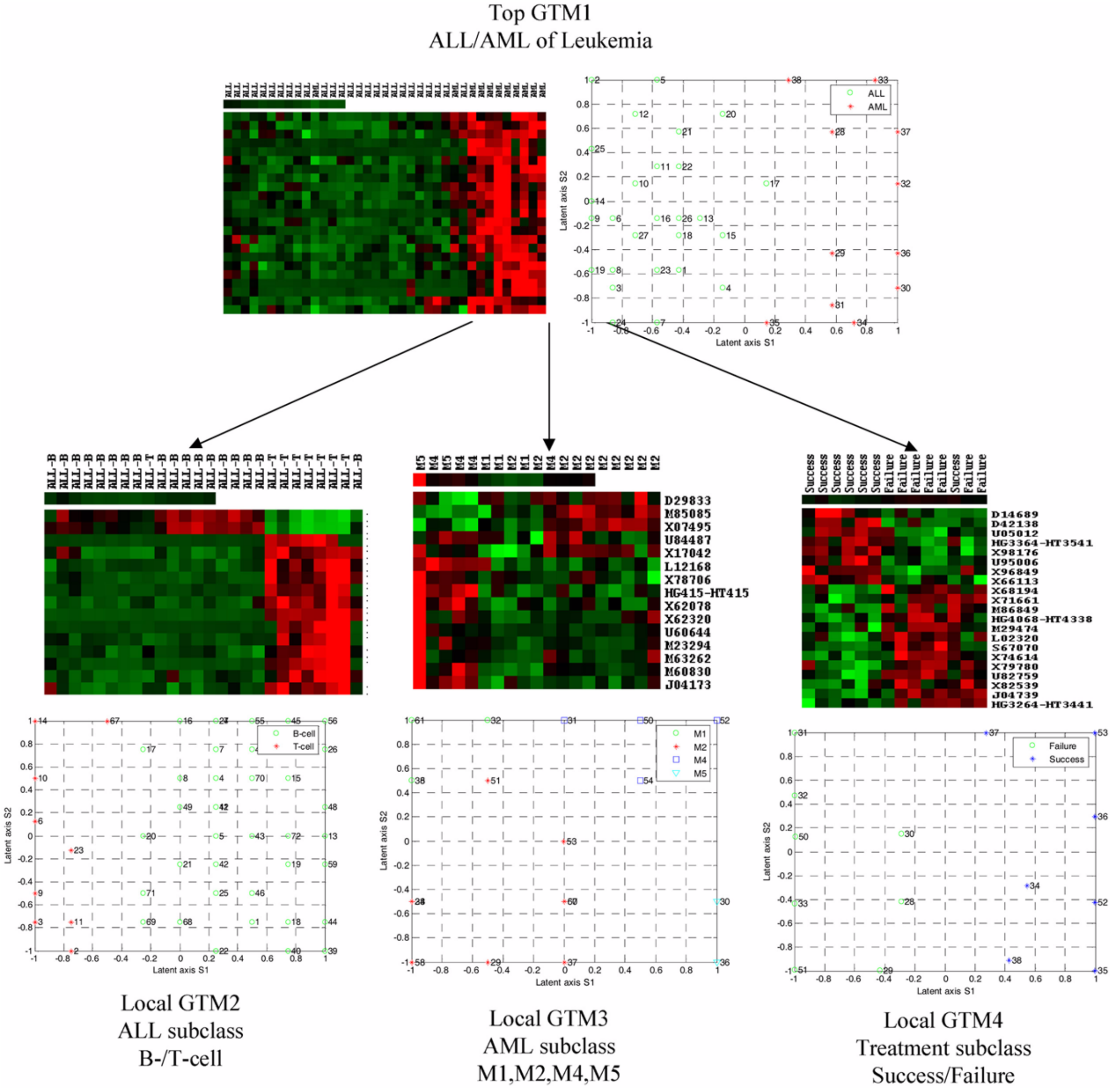

Fig. 3. A hierarchical visualization plot of the heat map and class prediction using four GTM models.

mixture centers in the data space.

Fig. 4 shows the representation of autoscaled gene expression profiles and posterior probability distribution over the latent space posterior of the GTM model for the given training samples. Fig. 4(a) shows the mutually co-regulated expression profiles whose expression levels of one gene activate or deactivate those of another gene through gene regulation mechanisms. From the expression profiles of Fig. 4(a), we find that using this simple technique is not easy when classifying the gene expression for the clinical outcome of leukemia patients. Fig. 4(b) shows the posterior probability distribution over the latent space of GTM. In this figure, the probability of data points for 38 training samples is multi-modal. Note that in practice, the distribution induced on the latent space by a data point represents the probability of that data point being generated by a particular set of latent points. In the case of multi-modal distribution, the importance of the latent point, obtained by the mode of the distribution, is much higher than for any other point on the latent space [25]. By comparing means and models, we confirm that the mean of the data is almost the same as the mode. Fig. 5 shows the learning trajectory of Gaussian centers of the GTM model 1. This indicates that the learning of the GTM is accomplished within 10 iterations.

Fig. 6(a) and (b) show the score plot of the PCA and GTM 1 model for 38 training sets of ALL and AML. Fig. 6(b) shows the mapping of the 38 training data sets with ALL and AML into the latent space of the GTM model 1. For the visualization of each data point, the posterior mean vectors are plotted in the latent variable space of GTM, where each point is labeled according to its sample number. For comparison, Fig. 6(a) also shows the corresponding results obtained with PCA. Note that the nonlinearity of GTM improves the separation of each class. Compared to PCA, GTM maps AML and ALL into two clusters in latent space, without any overlap, such

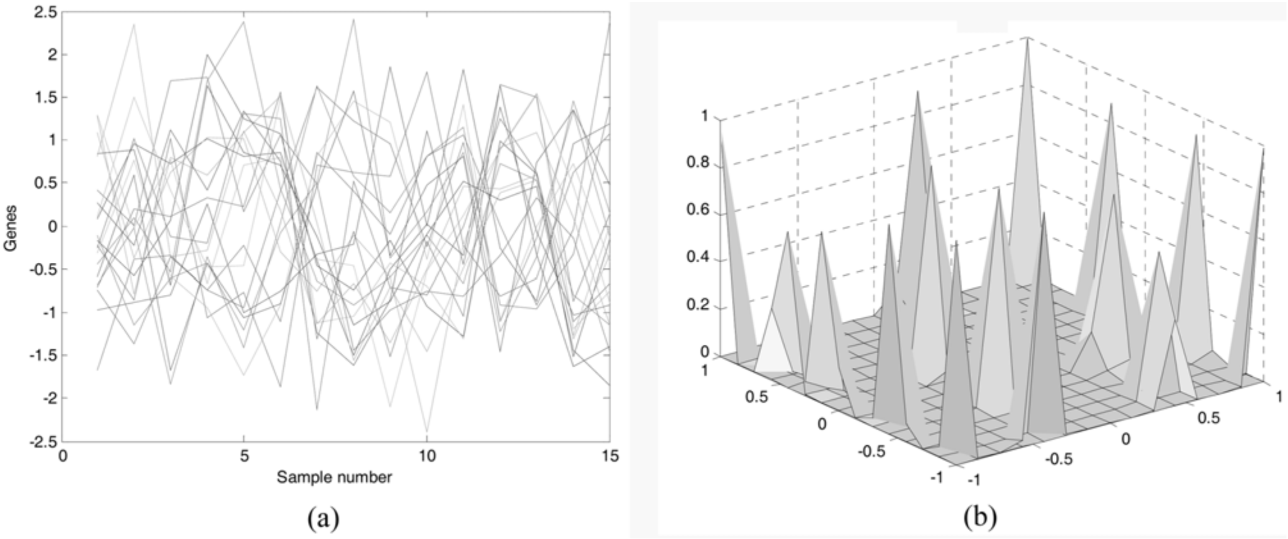

Fig. 4. Representation for the given samples of leukemia gene expression data (a) autoscaled gene expression profiles, (b) Posterior probability distribution over the latent space pos- terior to the GTM model.

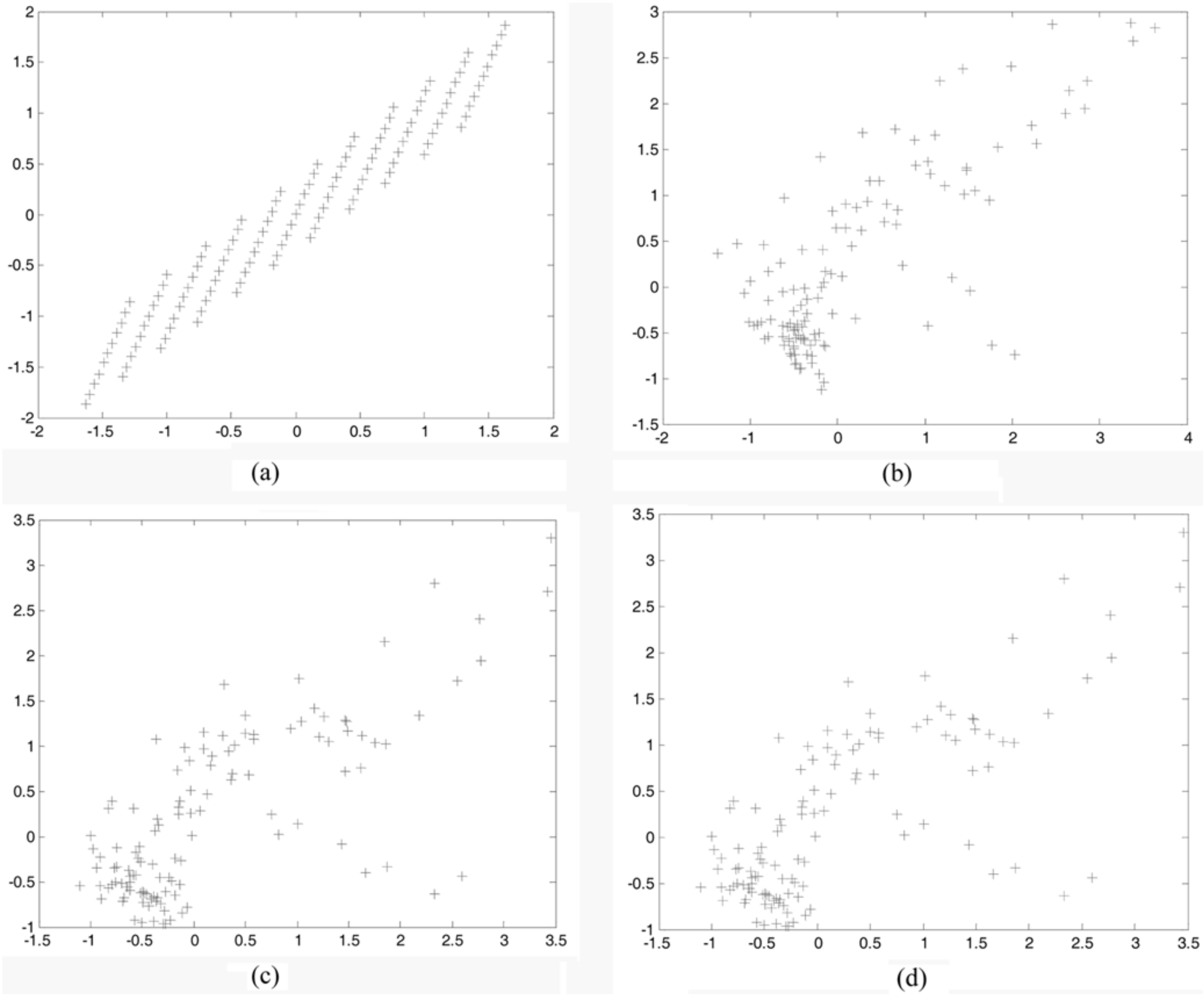

Fig. 5. The trajectory of Gaussian centers of GTM model 1 after, (a) initial, (b) 10 iterations, (c) 20 iterations, (d) 50 iterations.

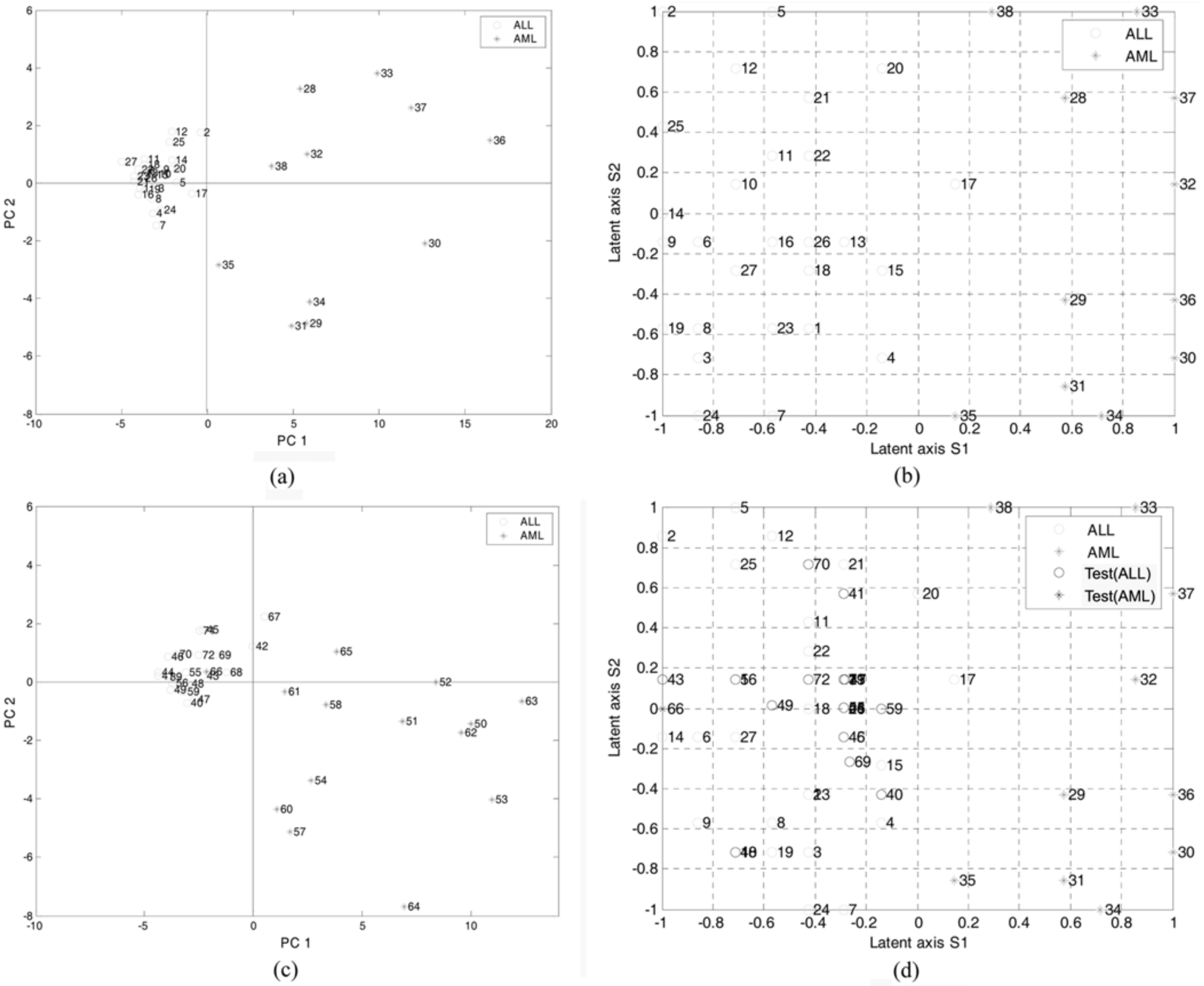

Fig. 6. Classification results of the ALL and AML training samples in the latent space of the (a) PCA, and (b) GTM model 1, and test samples in the (c) PCA, and (d) GTM model 1.

that AML samples are mapped into right plane and ALL samples are mapped into left plane.

For validation, the projected samples of an independent matrix of expression data from 34 unknown patients using posterior mean projection of GTM are shown in Fig. 6(c) and (d). The 34 test samples of ALL and AML are distinguished with 97% accuracy (32/34 patients) by the PCA and with 94% accuracy (33/34 patients) by the GTM, the exception being the 66th patient, which indicates that one of the test samples is misclassified. The only misclassified sample corresponds to patient 66, which was classified as ALL but actually labeled AML. Several investigations indicated that the leukemia data set of Golub et al. [14] contains at least one sample including patient 66 that is mislabeled and patient 66 has unusually low gene expression levels compared to other AML patients. Yoo et al. [9] showed that the contributions of most genes in the 66th patient are negative, contrary to the behavior of the other AML patients. In particular, the top-ranked gene, Zyxin, shows an abnormally low expression level for patient 66. This means that there are actually no classification errors for the 34 test data sets since sample number 66 is known as the mislabeled sample in the leukemia data set. Thus, it can be concluded that the nonlinear mapping of GTM enables the extraction of meaningful features that permit distinguishing between ALL and AML.

The prediction result of the proposed method is superior to those of Yoo et al. [9], which have two misclassifications, and Golub et al. [14], which have six misclassifications. On the other hand, PCA shows a dense distribution of ALL patients but quite a sparse distribution among AML patients in latent space. This indicates that the linear mapping of PCA cannot be enough to consider the precise relationship between ALL and AML. It also confirms that the nonlinear mapping of GTM can be used to discriminate AML and ALL subclasses.

2. Local Mapping of the ALL Subclass (B-cell and T-cell): GTM Model 2

ALL can be further classified into T-cell and B-cell lineages. In clinical practice, the B-cell lineage responds better to treatment than

Table 1. Comparison of classification results by applying Voting machine, PCA, and GTM to discriminate between the leukemia gene expression profiles. * indicates that the test in the method is not available

No.	Leukemia subclasses		Voting machine[a]	WEKA-machine learning[b]	Multi-PCA model	SVM[c]	Hierarchical GTM
1	ALL/AML	Validation error	5	3	2(#61, 66)	2-4	1(#66)
2	ALL subclass (B-cell and T-cell)	Validation error	*	*	2(#14, 67)	*	0
3	AML subclasses	Prediction of Unknown samples	*	*	vague samples 2(#14, 67)	*	0
4	AML patients (failure and success)	Prediction of Unknown samples	*	*	*	*	0

[a]Golub et al. [14]; [b]Wang et al. [26]; [c]Fuery et al. [27]

the T-cell lineage. Therefore, it is important to distinguish between these lineages. The 47 ALL patients were analyzed by using PCA and GTM: 38 B-cell and 9 T-cell lineages. 15 genes were selected by the DPLS method in order to allow discrimination between T-cell ALL (T-ALL) and B-cell ALL (B-ALL).

Table 1 describes the mapping of the 47 ALL data sets with B-cell and T-cell lineages into the latent space of the PCA and GTM model 2. In the GTM, the AML patients are clearly separated into two clusters in nonlinear latent space, one for each subtype, without any overlap, such that the T-cell subclass samples (#2, 3, 6, 9, 10, 11, 14, 23, 67) are mapped into the left plane and the B-cell subclass (the other samples) are mapped into the right plane. This confirms that the nonlinear mapping of GTM can be used to discriminate B-ALL and T-ALL subclasses. In PCA, samples 14 and 67 in linear latent space are not separated into the T-cell region, which demonstrates that the linear mapping of PCA cannot efficiently predict the ALL subclasses of T-cell and B-cell. This can be a reason why the classification results of PCA are not in good agreement with the results of the GTM when only the first two PCs are taken into account. The classification location and centers cannot be discriminated in the linear analysis of PCA. Unlike PCA, local GTMs approximate smooth two-dimensional manifolds with quantities useful for capturing the amount of nonlinear mapping.

3. Local Mapping of the AML Subclass (M1, M2, M4, M5): GTM Model 3

AML can be classified into six subtypes: M1, M2, M3, M4, M5, and M6. Although patients tend to be classified into either the M2 or M4 subtype under the FAB system, it is difficult for most doctors to discriminate sharply between these subtypes. Identifying the M3 subtype is important because this subtype usually responds well to treatment with retinoids. The M5 subtype is not easy to detect with the FAB system, and usually shows poor response to treatment. Most doctors recommend intensive chemotherapy for patients with this subtype. Correct determination of the AML subtype is important, since different subtypes will respond differently to treatment [14].

Among the 25 AML patients, data from 20 patients were used as a training data set (four M1 cases: samples 32, 35, 38, 61; ten M2 cases: samples 28, 29, 33, 34, 37, 51, 53, 57, 58, 60; four M4 cases: samples 31, 50, 52, 54; two M5 cases: samples 30, 36). The remaining five patients (samples 62-66), which could not be classified by Golub et al. [14], were used as a test data set. 15 genes were selected by using the DPLS method to allow discrimination between the AML subclasses. Many of the top 15 genes are relevant for AML subclass discrimination.

Table 1 demonstrates the prediction result of the AML subclass (M1, M2, M4, or M5) comprising the training samples and the five test samples (62-66) using PCA and GTM model 3. In the GTM of training samples, the AML patients are clearly separated into four clusters in nonlinear latent space, a separate one for each subtype, without any overlap, such that the M1 subclass data (#32, 38, 61) are mapped into the top left, the M2 subclass data (#28, 29, 33, 34, 37, 51, 53, 57, 58, 60) are mapped into the bottom left, the M4 subclass data (#31, 50, 52, 54) are mapped into the top right, and the M5 subclass data (#30, 36) are mapped into the bottom right. Based on the results of the training samples, GTM can be used to predict the subtype of the five AML samples that could not be predicted by the method of Golub et al. [14]. The prediction results for the five unknown test samples (#62-66) indicate that four AML patients (samples #63, 64, 65, 66) are of subtype M1, and one patient (sample #62) is of subtype M2. In the PCA, two samples (#62, 63) in linear latent space were not easily assigned to a subclass.

4. Clinical Outcome Prediction of AML Patients (Failure and Success): GTM Model 4

Relating gene expression patterns to clinical outcomes is a key issue in cancer genetics. One of the most promising aspects of gene expression profiling is the hope that it will enable more accurate identification of patients who are at a high risk of failing conventional therapy. To search for additional sets of genes useful for predicting the clinical outcome of leukemia patients, additional gene selections were performed for the prediction of the clinical output of AML treatment using DPLS [9]. The 21 most relevant genes were selected to discriminate between the failure and success of the clinical treatment of AML patients. Almost all of the 21 gene set (HoxA9, PIG-B, MACH-alpha-2 protein, BPI Bactericidal/permeability increasing protein, Autoantigen PM-SCL, ERGIC-53 Protein, and others) have been identified previously as being abnormal in AML or another form of leukemia [24]. Among the 25 AML patients of Golub et al. [14], 15 patients are used as a training data set, of whom 7 patients (#34-38 and 52-53) survived and 8 patients (#28-33, 50, and 51) died during treatment, and we used the remaining 10 patients (samples # 54, 57, 58, and 60-66), who did not respond to treatment, as a test data set.

Fig. 7(a)-(d) shows the prediction results of the 15 training samples of the AML subclass (success and failure) in the latent space of the (a) PCA, and (b) GTM model 4. In both plots, circles and stars represent the clinical outcome of failure and success, respec-

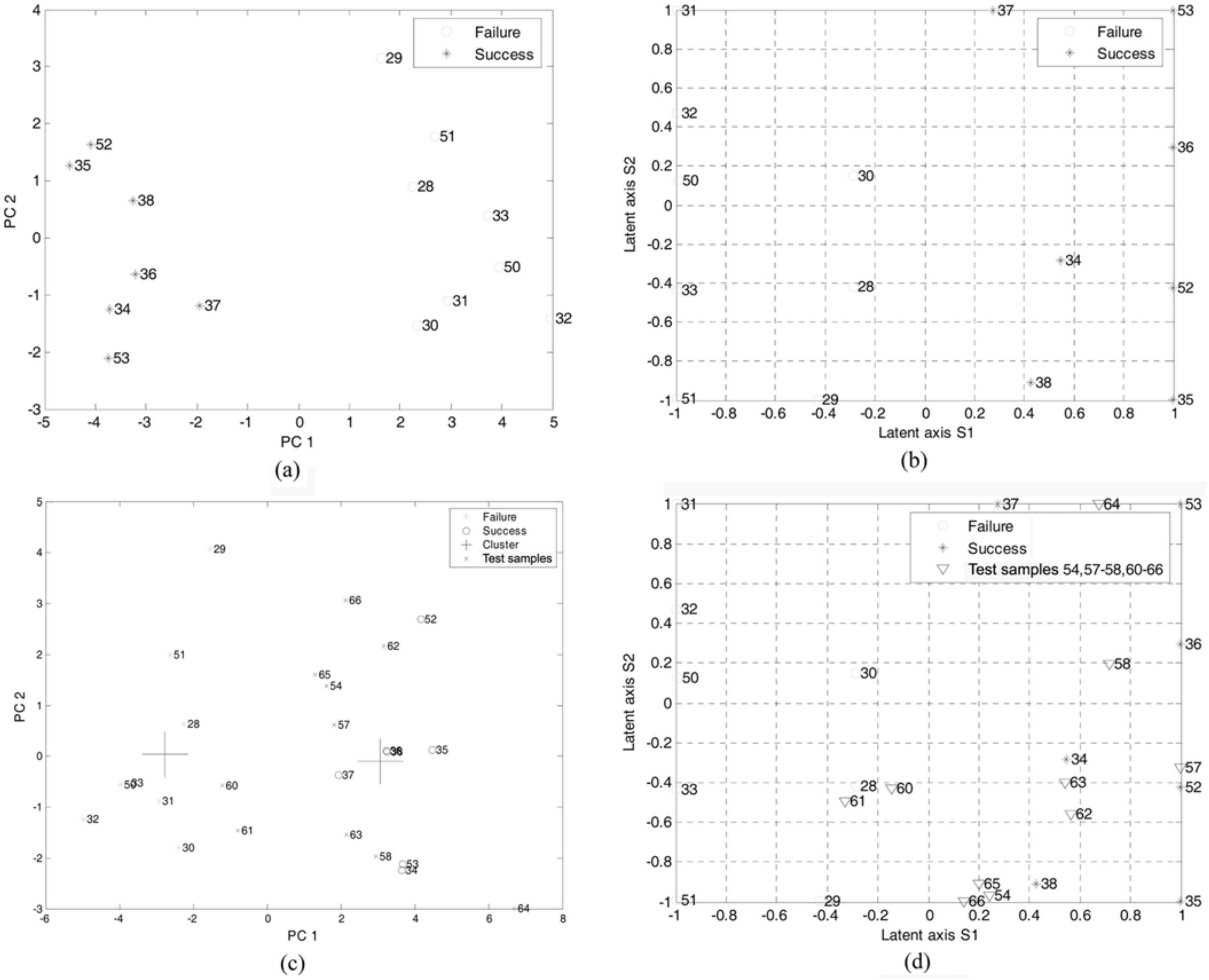

Fig. 7. Prediction results of the 15 training samples from the AML subclass in the latent space of the (a) PCA, and (b) GTM model 4, and the 10 test samples of the (c) PCA, and (d) GTM model 4.

tively. These plots demonstrate that both PCA and GTM can exactly discriminate the clinical outcome of AML patients, and that GTM can extract key feature components.

A meaningful classifier can predict the clinical outcome of unknown leukemia patient treatment. Fig. 7(c) and (d) show the prediction results of the AML subclass (success and failure) of 10 unknown AML test patients (54, 57, 58, 60-66) whose clinical outcomes were not specified by Golub et al. [14] by the (c) PCA, and (d) GTM model 4. The results of GTM in Fig. 7(d) indicate that eight AML patients (#54, 57, 58, 62, 63, 64, 65, and 66) are close to treatment survival, and two AML patients (#60 and 61) are close to death. From the topological view, two patients (#60 and 61) in the latent space are close to patient (#28) who is AML type BM, M2 subclass under the FAB system and died after clinical treatment. On the other hand, two patients (#60 and 61) were found to be AML type BM, M2 subclass under the FAB system. This is an interesting finding in which close samples in the latent space have the same type and the same subclass of leukemia in the input space. This finding shows that close samples in the input space have close representations in the latent space since GTM can preserve a topology between input and latent space. From close representations between #60, 61 and #28 in the latent space, patients #60 and 61 are predicted to die after treatment. On the other hand, eight AML patients (#54, 57, 58, 62, 63, 64, 65, and 66) are predicted to survive after treatment. In more detail, the results indicate that three patients (#54, 65, 66) are close to a sample (#38) of the BM and M1 type, two patients (#62, 63) are close to a sample (#34) of the BM and M2 type, one patient (#57) is close to a sample (#52) of the PB and M4 type, and one patient (#58) may be in overlap between two samples (#36 and 34) in the progression of disease and may be a new undiscovered class. This sample #58 would need further study for clinical treatment and prognosis. All unknown samples were predicted for the clinical outcome of leukemia patients using the proposed method.

To evaluate the performance of the proposed method, we compared our method with previously developed methods. In general, it is somewhat difficult to directly compare these methods because they each use a difficult criterion. We compared their classification performances using the number of the misclassification samples. Table 1 compares the classification results of 4 subclasses by applying the voting machine [14], WEKA-machine learning [26], multi-PCA, support vector machine (SVM [27]) and hierarchical GTM models for discriminating between the leukemia gene expression profiles, where the validation error is the number of the misclassified unknown samples, and * means that the classification result in the method is not available. Note that the sample number 66 is known as the mislabeled sample in the leukemia dataset and may influence the error rate [28]. This data provides classification performances on leukemia data sets that are similar or better than other methods and other published results.

Our conclusion is that a nonlinear mapping by GTM enables extraction of meaningful features to discriminate between success and failure of AML treatment. Our findings show that the local mapping approach by the hierarchical GTM models can give more reliable and higher resolution classification results than the global and linear methods. Thus, the proposed method makes it possible to predict the clinical outcome of AML patients in more detail. Although the clinical outcome is also affected by many other factors, such as patient age, treatment regime, and time of diagnosis [9,29], the results presented here highlight the potential of the proposed method for uncovering the prognostic indicators of leukemia.

CONCLUSIONS

A hierarchical framework for mapping multi-class cancer microarray data to nonlinear latent space is proposed in this research. The hierarchical approach allows the users to make an intensive investigation into several interest regions and find out more about the data as well as the clusters. The application results indicate that the non-linearity of GTM and the local information of hierarchical models give an improved separation of the clusters over the linear PCA and the nonlinear SVM. Moreover, the proposed method allows the elucidation of leukemia subclasses with local visualization and prediction of the output of patients' treatments. Our findings show that the hierarchical nonlinear mapping approach can give more reliable and higher resolution classification results than the global and linear methods.

ACKNOWLEDGMENTS

This work was supported by Brain Korea21 project, the Korea Research Foundation by Grant funded by the Korean Government (MOEHRD) (KRF-2007-331-D00089) and funded by Seoul R&BD Program (CS070160).

NOMENCLATURE

EM	: expectation-maximization
G	: a K×K diagonal matrix
K	: number of regular array of nodes within the latent space
L	: log likelihood of posterior probability density model
R_{ni}	: responsibility of Gaussian $p(\mathbf{x}_n\|u_i, \mathbf{W}, \beta)$ generated the point x_n in the data space
R	: a K×N matrix with elements R_{ni}
u	: variables in latent space
W	: weight of the nonlinear mapping from latent space to data space
X	: an N×D matrix with microarray data
x	: variables in data space

Greek Letters

ϕ	: Gaussian function
Φ	: K×M matrix with elements $\phi_{ij}=\phi_j(u_i)$

REFERENCES

1. G. M. Hampton and H. F. Frierson, *Trends. Mol. Med.*, **9**, 5 (2003).
2. M. F. Ochs and A. K. Godwin, *Bio Techniques*, **34**, S4 (2003).
3. B. T. Zhang, J. S. Yang and S. W. Chi, *Machine Learning*, **52**, 67 (2003).
4. S. Bicciato, M. Pandin, G. Didone and C. Di Bello, *Biotechnol. Bioeng.*, **81**, 594 (2002).
5. S. Dudoit, J. Fridlyand and T. P. Speed, *J. Am. Stat. Assoc.*, **97**, 77 (2002).
6. D. V. Nguyen and D. M. Rocke, *Bioinformatics*, **18**(1), 39 (2002).
7. G. Stephanopoulos, D. H. Hwang, W. A. Schmit, J. Misra and G. Stephanopoulos, *Bioinformatics*, **18**(8), 1054 (2002).
8. N. L. W. van Hal, *J. Biotechnol.*, **3**, 271 (2002).
9. C. K. Yoo, I. Lee and P. A. van Walleghem, *Comp. & Chem. Eng.*, **29**, 1345 (2005).
10. Y. Gao and G. Church, *Bioinformatics*, **21**(21), 3970 (2005).
11. G. Sanguinetti, M. Milo, M. Rattray and N. D. Lawrence, *Bioinformatics*, **21**(19), 3748 (2005).
12. L. Li, *Bioinformatics*, **22**(4), 466 (2005).
13. Y. Lu and J. Han, *Information Systems*, **28**, 243 (2003).
14. T. R. Golub, D. K. Slonim, P. Tamayo and E. S. Lander, *Science*, **286**, 531 (1999).
15. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander and T. R. Golub, *Proc. Natl. Acad. Sci.*, **96**, 2907 (1999).
16. P. Toronen, M. Kolehmainen, G. Wong and E. Castren, *FEBS Lett.*, **451**, 142 (1999).
17. A. A. Alizadeh, *Nature*, **403**, 503 (2000).
18. C. M. Bishop and M. Svensen, *Neurocomputing*, **21**, 203 (1998).
19. C. M. Bishop and M. E. Tipping, *Pattern Analysis and Machine Intelligence*, **20**(3), 281 (1998).
20. C. M. Bishop, M. Svensen and C. K. I. Williams, *Neural Comput.*, **10**(1), 215 (1998).
21. P. Tino and I. Nabney, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **24**(5), 639 (2002).
22. I. T. Nabney, Y. Sun, P. Tino and A. Kaban, *IEEE Trans. Knowledge and Data Engineering*, **17**(3), 384 (2005).
23. J. F. M. Svensen, Ph. D Thesis, Aston University (1998).
24. J. Lyons-Weiler, S. Patel and S. A. Bhattacharya, *Genome Res.*, **13**, 503 (2003).
25. A. O. Andrade, S. Nasuto, P. Kyberd and C. M. Sweeney-Teed, *Biosystems*, **82**, 273 (2005).
26. Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. X. Mayer and H. W. Mewes, *Comput. Biology Chemistry*, **29**, 37 (2005).
27. T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler, *Bioinformatics*, **16**, 906 (2005).
28. M. L. Chow, J. Moler and I. S. Mian, *Physiol. Genomics*, **5**, 99

(2005).

29. J. G. Thomas, J. M. Olson, S. J. Tapscott and L. P. Zhao, *Genome Res.*, **11**, 1227 (2001).

APPENDIX

1. Generative Topographic Mapping (GTM)

Data space variables are denoted in $\mathbf{x}=[x_1, x_2, \ldots, x_D]^T$, while the gene expression profiles dataset and latent space variables are denoted in $\mathbf{u}=[u_1, u_2, \ldots, u_L]^T$, a regular array of nodes, i=1, …, K, within the latent space. The GTM defines a nonlinear, parametric mapping $\mathbf{y}(\mathbf{u}, \mathbf{W})$ from a latent space ($\mathbf{u}\in R^L$) to a data space ($\mathbf{x}\in R^P$) where normally L<D. It can be fitted to a data set $\{\mathbf{x}_n\}$, where n=1, 2, …, N, by a non-linear latent variable with a set of M fixed nonlinear basis functions $\phi(u)=\{\phi_j(u)\}$ [18,19]. Points in the low dimensional latent space are mapped to corresponding centers of a Gaussian in the observable high-dimensional data space. The probability of data over the latent space p(**u**) will induce a probability over the data space, $p(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta)$ and is defined as a Gaussian distribution:

$$p(\mathbf{x}|\mathbf{u},\mathbf{W},\beta)=\left(\frac{\beta}{2\pi}\right)^{m/2}\exp\left[-\frac{\beta}{2}\|\mathbf{x}-\mathbf{y}(\mathbf{u},\mathbf{W},\beta)\|^2\right] \quad \text{(A1)}$$

where **x** is a point in the data space, **W** is a weight matrix, and β^{-1} is the noise variance, which allows for some variance in the observed variables not explained by the latent variables. Both **W** and β^{-1} are estimated by an expectation-maximization (EM) algorithm [20,23].

An optimal parameter estimation of a GTM model for a data set is obtained by maximizing the log likelihood function, L, of $p(\mathbf{x}|\mathbf{W}, \beta)$ with respect to **W** and β.

$$L=\sum_{n}^{N}\ln\left(\frac{1}{K}\sum_{k}^{K}p(\mathbf{x}_n|\mathbf{u}_k,\mathbf{W},\beta)\right) \quad \text{(A2)}$$

In the E-step, the current values of the parameters **W** and β evaluate the posterior probability, responsibility, which each component i takes for every data point. The responsibility of a data point, $\mathbf{x}_n$, is given by Svensen [23].

$$R_{ni}=p(\mathbf{x}_i|\mathbf{u}_n,\mathbf{W}_{old},\beta_{old})=\frac{p(\mathbf{u}_n|\mathbf{x}_i,\mathbf{W}_{old},\beta_{old})}{\sum_{j=1}^{K}p(\mathbf{u}_n|\mathbf{x}_j,\mathbf{W}_{old},\beta_{old})} \quad \text{(A3)}$$

Then, in the M-step, the responsibility is used to re-estimate the weight matrix **W** by solving the following linear equation.

$$(\Phi^T\mathbf{G}_{old}\Phi)\mathbf{W}^T_{new}=\Phi^T\mathbf{R}\mathbf{X} \quad \text{(A4)}$$

where Φ is a K×M matrix with elements $\phi_{ij}=\phi_j(u_i)$, **X** is an N×D matrix with elements x_{nk}, **R** is a K×N matrix with elements R_{ni}, and **G** is a K×K diagonal matrix with elements $G_{ii}=\sum_{n=1}^{N}R_{ni}(\mathbf{W},\beta)$. Here, the matrix, $\mathbf{W}_{new}$, can be easily solved using an efficient matrix computation of Cholesky decomposition or singular value decomposition (SVD) to solve the singularity problem in computation of the inverse of $(\Phi^T\mathbf{G}_{old}\Phi)$ [19]. In the paper, a modified algorithm of GTM, which adds a regularization term to the objective function, $1/2(\lambda\|\mathbf{W}\|^2)$, is used to control the nonlinear mapping. Therefore, the Eq. (A4) of M-step is modified as:

$$(\Phi^T\mathbf{G}_{old}\Phi+(\lambda+\beta)\mathbf{I})\mathbf{W}^T_{new}=\Phi^T\mathbf{R}\mathbf{X} \quad \text{(A5)}$$

where **I** is the M×M identity matrix andis a fixed hyperparameter. The following is a summary of GTM using an EM algorithm:

E-step: compute $R_{ni}(\mathbf{W}_{old}, \beta_{old})$

$$R_{ni}(\mathbf{W}_{old},\beta_{old})=\frac{p(\mathbf{x}_n|\mathbf{u}_i,\mathbf{W}_{old},\beta_{old})}{\sum_{i=1}^{K}p(\mathbf{x}_n|\mathbf{u}_i,\mathbf{W}_{old},\beta_{old})} \quad \text{(A6)}$$

where $p(\mathbf{x}|u,\mathbf{W}_{old},\beta_{old})=\left(\frac{\beta_{old}}{2\pi}\right)^{D/2}\exp\left[-\frac{\beta_{old}}{2}\|\mathbf{x}-\mathbf{W}\phi(u_i)\|^2\right]$

M-step: re-estimate **W**, β

$$\mathbf{W}^T_{new}=(\Phi^T\mathbf{G}_{old}\Phi+(\lambda/\beta)\mathbf{I})^{-1}\Phi^T\mathbf{R}_{old}\mathbf{X} \quad \text{(A7)}$$

$$\frac{1}{\beta_{new}}=\frac{1}{ND}\sum_{t=1}^{N}\sum_{i=1}^{K}R_{ni}\cdot\|\mathbf{x}_n-\mathbf{W}\phi(u_i)\|^2 \quad \text{(A8)}$$

A detailed derivation of the EM algorithm of GTM can be found in (Bishop et al. [20]).

2. Data Visualization of GTM in the Two-dimensional Latent Space

For the purpose of visualization, it is convenient to summarize each such distribution by its mean [19], where Bayes' theorem can be used to invert the nonlinear transformation from the latent space to the data space. GTM projects the point from the data space into the low-dimensional latent space. The posterior distribution of latent space, given a data point x_n, is a sum of delta functions centered data at u_i with coefficients given by the responsibilities

$$R_{i,n}=\frac{p(\mathbf{x}_n|u_i,\mathbf{W},\beta)}{\sum_{i=1}^{K}p(\mathbf{x}_n|u_i,\mathbf{W},\beta)} \quad \text{(A9)}$$

The responsibilities $R_{i,n}$ are the posterior probabilities that the Gaussian $p(\mathbf{x}_n|u_i, \mathbf{W}, \beta)$ is generated at the point x_n in the data space. For visualization, GTM projects points x_n from the data space into the low dimensional latent space. The latent space representation of the point x_n is taken to be the mean of the posterior distribution on the latent space [21]. It can be calculated using Eq. (A10) and averaging u_i weighted by $R_{i,n}$ over i

$$\sum_{i=1}^{K}R_{i,n}u_i \quad \text{(A10)}$$

Thus, a set of data points x_n, n=1, 2, …, N, is projected onto a corresponding set of the means of the posterior distribution in the two-dimensional latent space.