

An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data

Arthur L. Hsu^{1,*}, Sen-Lin Tang² and Saman K. Halgamuge¹

¹*Mechatronics Research Group, Department of Mechanical and Manufacturing Engineering, University of Melbourne, Victoria, Australia 3010* and ²*School of Veterinary Science, University of Melbourne, Victoria, Australia 3010*

Received on October 29, 2003; revised on April 2, 2003; accepted on April 17, 2003

ABSTRACT

Motivation: Current Self-Organizing Maps (SOMs) approaches to gene expression pattern clustering require the user to predefine the number of clusters likely to be expected. Hierarchical clustering methods used in this area do not provide unique partitioning of data. We describe an unsupervised dynamic hierarchical self-organizing approach, which suggests an appropriate number of clusters, to perform class discovery and marker gene identification in microarray data. In the process of class discovery, the proposed algorithm identifies corresponding sets of predictor genes that best distinguish one class from other classes. The approach integrates merits of hierarchical clustering with robustness against noise known from self-organizing approaches.

Results: The proposed algorithm applied to DNA microarray data sets of two types of cancers has demonstrated its ability to produce the most suitable number of clusters. Further, the corresponding marker genes identified through the unsupervised algorithm also have a strong biological relationship to the specific cancer class. The algorithm tested on leukemia microarray data, which contains three leukemia types, was able to determine three major and one minor cluster. Prediction models built for the four clusters indicate that the prediction strength for the smaller cluster is generally low, therefore labelled as uncertain cluster. Further analysis shows that the uncertain cluster can be subdivided further, and the subdivisions are related to two of the original clusters. Another test performed using colon cancer microarray data has automatically derived two clusters, which is consistent with the number of classes in data (cancerous and normal).

Availability: JAVA software of dynamic SOM tree algorithm is available upon request for academic use.

Contact: alhs@mame.mu.oz.au

Supplementary information: A comparison of rectangular and hexagonal topologies for GSOM is available from http://www.mame.mu.oz.au/mechatronics/journalinfo/Hsu_2003supp.pdf

INTRODUCTION

Microarray technologies (Schena *et al.*, 1995; Shalon *et al.*, 1996) have enabled genomic studies to advance faster than ever, as it allows expression levels of thousands of genes to be recorded, monitored and analysed simultaneously under different conditions. However, the vast amount of information obtained from the DNA microarray requires much computational aid to interpret and to extract useful information. This task imposes the need for multivariate analysis where hundreds, if not thousands, of genes are compared at the same time. One frequently performed task for the study of microarray data is clustering of samples based on expression patterns of a set of common genes, which are of strong interest to many authors in terms of cancer class discovery and significant gene identification (Getz *et al.*, 2000; Golub *et al.*, 1999). Meaningful discovery of a group of samples that have similar genes expression levels can provide insight to therapeutic and pathogenetic studies. For instance, the conventional way to identify tumour type is to divide visible abnormal cells into morphological groups (Triche *et al.*, 1988), but the clustering of samples based on gene expression levels can at least identify the genes that play an important role in the classification of tumour types, or even opens the door to understanding of the genetic events and mechanisms of tumour progression.

Two frequently employed methods for clustering gene expression levels (Tibshirani *et al.*, 1999) are hierarchical clustering (Sneath and Sokal, 1973) and Self-Organizing Maps (SOMs) (Kohonen, 1997). Classical hierarchical clustering methods have been reported to have several drawbacks such as lack of robustness when there is strong presence of

*To whom correspondence should be addressed.

noise in data. In addition, it may not provide unique clustering solution, since the clustering result is mostly in the form of a binary tree that can be segmented in many ways to yield a given number of clusters. Further, it may depend strongly on the order of data and consumes exponential time in complete clustering (Tamayo *et al.*, 1999). For the above reasons, many authors have used self-organizing networks (Herrero *et al.*, 2001; Kaski, 2001; Tamayo *et al.*, 1999; Toronen *et al.*, 1999), due to their robustness against noisy data. However, the SOM algorithm requires the number of clusters to be previously defined by the user. Even though hierarchical clustering methods with SOM already exist (Vesanto and Alhoniemi, 2000), the partitioning of clusters is still not uniquely defined in those methods, as for all the classical hierarchical clustering methods. The main functionality of producing SOM for hierarchical clustering is SOM's data compression property, where nodes of SOM serve as prototypes (or mean values) for a number of similar data entries, so that complete clustering can be achieved in a reasonable time.

In our context of analysis of patients' gene expression patterns, the focus involves two main aspects—cancer class discovery and significant gene identification. Cancer class discovery based on clustering of gene expression patterns have been used in different cancer types by many authors (Alizadeh *et al.*, 2000; Alon *et al.*, 1999; Ben-Dor *et al.*, 2000; Dudoit *et al.*, 2000; Getz *et al.*, 2000; Golub *et al.*, 1999; Sharan and Shamir, 2000). The challenge is not solely in the clustering quality alone, but also to obtain meaningful and adequate number of clusters. With meaningful clusters, grouped in appropriate numbers, identification of the genes that contribute significantly to the differentiation of clusters would become a simpler task. Golub has adopted the SOM approach for the class discovery task, but the number of clusters is pre-defined based on the a priori knowledge of data (Golub *et al.*, 1999), which may compromise the class discovery capability. Getz, on the other hand, used a coupled two-way clustering algorithm to search for clusters of genes that optimally partition clusters of patients (or samples) (Getz *et al.*, 2000). The clustering method allows number of clusters to be automatically defined, but the performance remains to be evaluated by the user community. We propose here a fully unsupervised methodology that uses a combination of dynamic SOM tree and Growing Self-Organizing Maps (GSOMs) in microarray analysis of class discovery and marker gene identification tasks. The known class labels of the sample data are only used to evaluate the performance of the method.

ALGORITHM AND IMPLEMENTATION

Current classifiers and gene selection methods

Recent publications involve two main types of classifiers for cancer classification. The first type—cluster analysis, is used for the identification of new or unknown cancer classes using gene expression profiles (Sneath and Sokal, 1973; Dudoit

et al., 2000; Ben-Dor *et al.*, 2000; Hartuv *et al.*, 2000; Getz *et al.*, 2000; Golub *et al.*, 1999). The second type, discriminant analysis (or supervised learning), is used for the classification of malignancies into known classes (Ramaswamy *et al.*, 2001; Dudoit *et al.*, 2000; Zhang *et al.*, 2001; Khan *et al.*, 2001).

Gene selection, or otherwise known as feature selection, has been applied to genomic data to reduce the dimensions of the data and improve classification accuracy (Campbell *et al.*, 2001; Guyon *et al.*, 2000; Xing *et al.*, 2001). Other benefits of gene selection include: reduction of noise in the data and avoiding over-fitting. Relevant selection methods include: (a) a correlation metric that measures the relative class separation produced by the expression values of a gene (Golub *et al.*, 1999); (b) a log likelihood function for evaluating the suitability of a gene in class discrimination (Keller *et al.*, 2000); (c) recursive gene elimination (Guyon *et al.*, 2000).

Modified GSOM with hexagonal topology

A number of structure adaptive algorithms have been proposed to compensate for the static nature of SOM. Algorithms such as Growing Cell Structure (GCS) (Fritzke, 1994), Incremental Growing Grid (IGG) (Blackmore and Miiikkulainen, 1995), Growing Neural Gas (GNG) (Fritzke, 1995) and GSOM (Alahakoon *et al.*, 2000) all have the capability to attach more nodes to the network during training. Since the dynamic SOM tree requires the use of GSOM, we will introduce GSOM in further detail.

As the original GSOM only supports rectangular topology (Alahakoon *et al.*, 2000), we have, in this work, modified GSOM to support hexagonal topology, which is known to have better topology preservation for SOM (Kohonen, 1997). This implementation involves several changes to the original GSOM algorithm, which includes changes in initial grid shape and new node initialization methods, as detailed later. As a variant of SOM and inherited from SOM, GSOM has the same topology structure and weight vector adaptation rules as SOM, but grows according to its own growing criterion, whereas SOM is not capable to grow. A parameter of growth, Growth Threshold (GT), is defined as:

$$GT = -D \times \ln(SF)$$

where D is the dimensionality of data and SF is the user defined Spread Factor that takes values [0, 1], with 0 representing minimum and 1 representing maximum growth. GSOM is initialized to have one lattice structure as illustrated in Figure 1.

When the winning node is identified, then an accumulated error counter E of the winning node is updated by the following rule:

$$E(t + 1) = E(t) + \|I - w_{\text{winner}}\|$$

where I is the input vector and w_{winner} is the weight vector of the winning node. If the winning node is on the boundary of

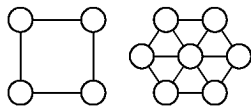


Fig. 1. Initial GSOM grid for rectangular topology (left) and hexagonal topology (right).

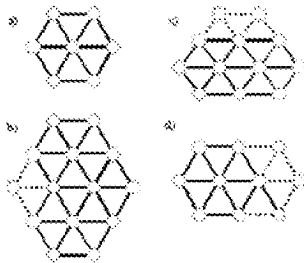


Fig. 2. (a) Initial GSOM. (b)–(d) All possible growing modes of hexagonal GSOM. In all cases, each new node will have an existing node topologically opposite to it.

GSOM (boundary node) and E exceeds GT, growing is initiated on that node to fill the surrounding unoccupied spaces of the lattice (again, rectangular or hexagonal). Using this growing phase more nodes can be added to the network to provide adequate resolution as specified by SF. In the case when E of the winning node exceeds GT and the winning node is not a boundary node, E is propagated outwards to the node’s neighbouring nodes.

Weights of the new nodes will be initialized according to the equations:

$$w_{new} = 2w_{winner} - w_{opposite}$$

where $w_{opposite}$ is the weights of the node topologically opposite to the new node if it exists, otherwise

$$w_{new} = w_{winner} + w_{other1} - w_{other2}$$

where w_{other1} and w_{other2} are weights of the nodes nearest to the new node, but are not the winning nodes. However, for hexagonal topology, there will always be a neighbour of the winning node that is topologically on the opposite side of the new node (Fig. 2b–d), therefore only the first equation is needed to determine weights of the new node.

Dynamic SOM tree

Conventional ways of clustering SOMs have some limitations. Visually identifying clusters on SOM maps by aid of colour code has limited accuracy, particularly when cluster boundaries are not well defined. Specifying the number of nodes to coincide with the number of expected clusters will require a priori knowledge of the data. The dynamic SOM tree algorithm is a robust hierarchical clustering application based on the GSOM that is able to improve the clustering process and

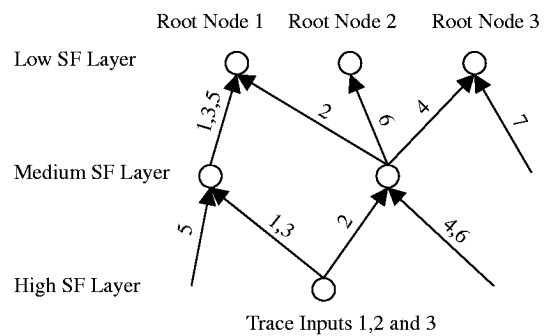


Fig. 3. Tracing inputs to identify root nodes and construct dynamic SOM tree.

to assist the analyst in understanding cluster properties (Hsu *et al.*, 2000). In this paper, we used the modified GSOM with hexagonal topology, which provides better map and clustering quality (in supplementary information).

Let us define map resolution by average number of nodes or neurons per input, so that with the same number of inputs a larger map will give a higher map resolution. In the case of GSOM, this can be achieved by specifying a higher SF as the SF implicitly determines the map resolution. Even though selecting a larger size for conventional SOM will have a similar effect, it also involves the difficulty of first determining an appropriate size and aspect ratio for the SOM.

To understand and visualize cluster relationships using GSOM, at least two GSOMs of different map resolutions are required. From two layers of GSOMs, i.e. one high and one low SF value GSOM, we are able to visualize cluster separation and merging by tracing their input mapping. For example, when 10 inputs that are mapped by two nodes (representing two clusters) in the high SF GSOM is mapped by only one node (representing a single cluster) in the low SF GSOM. This can be interpreted as the single cluster of 10 inputs and is separated into two clusters as SF increases. In the scenario of a dynamic SOM tree model, starting with a high SF layer and then gradually decreasing SF values for subsequent layers supports the merging of clusters based on map resolution.

A dynamic SOM tree is constructed using input tracing (Fig. 3). After each layer has been trained and inputs calibrated to nodes, the inputs are traced from the highest SF layer nodes (leaf nodes). A set of root nodes can be identified for each leaf node and leaf nodes having overlapping root nodes belong to the same cluster. With such tracing, distinct clusters having distinct sets of root nodes will form.

The use of a self-organizing network as the medium of the tree model provides with two additional strengths. First, it reduces the dimensionality of the input space to lower dimensions. Secondly, the amount of data is compressed when nodes act as averages of a number of similar data and the two-dimensional projection approximates probability density distribution of the input space.

As a rule of thumb of applying dynamic SOM tree to data, one can start constructing dynamic SOM tree from very high SF values (e.g. SF = 0.99 or SF = 0.9) and reduce by 0.1 per layer until terminating criterion is satisfied and then stop adding more layers. The terminating criterion is satisfied when the minimum possible resolution (SF = 0) is reached or stable clusters are formed. The latter is the case when the partitions of clusters remain unchanged after several additions of GSOM layers (indicating that the partitions are stable).

Dynamic SOM tree has demonstrated its robustness in identifying appropriate number of clusters in an unsupervised way for various data and shown strength in good clustering quality even compared to well-known clustering algorithms such as k-means and fuzzy c-means that predefined correct number of clusters (Hsu and Halgamuge, 2003). The robustness of dynamic SOM tree is attributed to several advantages inherited from both hierarchical clustering method and the SOM algorithm.

The number of clusters generated is uniquely defined for a specified clustering resolution with small variation that may depend on the initialization method of the GSOM. The resolution is interactively set by user to increase or decrease the number of clusters depending on the suitable hierarchical clustering strategy at the current resolution. This property is of great value in our context of investigation, since most current algorithms such as SOM and hierarchical clustering algorithms do not automatically find the optimal number of clusters or lack uniquely defined partitioning. With the use of an algorithm that is able to identify the number of likely clusters, autonomous class discovery from unlabelled microarray data (in our case, illness types are considered as unknown) becomes possible. For example, a group of cancerous patients having particular genes abnormally expressed, either over-expressed or under-expressed in comparison to normal patients, can be identified.

Dynamic SOM tree and GSOM combination for class discovery and marker gene identification

In this approach, we propose to use a combination of the dynamic SOM tree and a modified Golub's method for automatic class discovery and marker gene identification, respectively. The approach comprises three phases, Class discovery, marker gene identification and partition refinement (Fig. 4), as detailed later.

In the class discovery, preprocessed gene expression data is being clustered by dynamic SOM tree to automatically identify a number of natural partitions in the data set (automatic, as suggested by the algorithm). Under the restriction that samples are collected from cells related to the cancer type. For example, since leukemia is a blood related disease, samples collected by Golub *et al.* (1999). are bone marrow cells, which are responsible for producing blood. Such as to ensure the resulting partitions will have biological meanings. These clusters are then assumed to be actual classes in

Phase 1 – Class Discovery

1. Automatically identify N predicted classes (PC_1, \dots, PC_N) by clustering samples at level t (initially $t=1$) with Dynamic SOM Tree and assign each sample s_j to a predicted class i , such that $PC(s_j)=PC_i$

Phase 2 – Marker Gene Identification

2. Compute C_2^N class distinction scores for all genes
3. For each predicted class $PC_i, i \in N$
 - (a) Train a GSOM from $N-1$ class discrimination scores that are relevant with respect to i
 - (b) Identify the node with highest average class discrimination score and use all genes it represents as predictor genes

Phase 3 – Partition Refinement

4. Use weighted voting of predictor genes, calculate the $PC(s_j)^{new}$ and prediction strength $PS(s_j)$ for each sample s_j
5. Iterate to refine original clustering
 - (a) If $PC(s_j)^{new} \neq PC(s_j)$, repeat steps 2 to 5
 - (b) If $PC(s_j)^{new} = PC(s_j)$ and at least 5 samples with $PS(s_j) < \alpha$, a defined threshold value, then go to step 6
 - (c) If $PC(s_j)^{new} = PC(s_j)$ and not more than 5 samples with $PS(s_j) < \alpha$, then algorithm stops
6. Isolate all such s_j with $PS(s_j) < \alpha$ and then apply steps 1 to 5 on these sub-samples, $t=t+1$

Fig. 4. Pseudo-code of Dynamic SOM Tree and GSOM combination approach.

the data, therefore will be referred to as predicted class (PC) hereafter.

The second phase searches for the most significant genes, often referred to as the predictor genes, which contributed predominantly to the formation of the PCs by means of modified neighbourhood analysis (Golub *et al.*, 1999). Neighbourhood analysis involves computing class discrimination scores of all genes from the PCs, which evaluates a given gene's suitability to distinguish two specified PCs and is defined as:

$$P(g, i, j) = \frac{\mu_i(g) - \mu_j(g)}{\sigma_i(g) + \sigma_j(g)}$$

where i and j are the specified PCs and $\mu_i(g)$ is the mean value, $\sigma_i(g)$ is the standard deviation, of the expression level of gene g for all samples belonging to PC_i . P has high absolute value if the gene is strongly suitable for discriminating class i from j , strongly negative if in favour of PC_j and strongly positive if in favour of PC_i .

There will be C_2^N pair-wise class discrimination scores for each gene, where N is the number of PCs, but only $N - 1$ scores that are related to differentiating PC_i from the other PCs will be used. Since the predictor genes will be selected

Table 1. Sample class discrimination scores as input for GSOM training for PC_2

Score	$P(g, 2, 1)$	$P(g, 2, 3)$...	$P(g, 2, N)$
Gene, g	0.83	0.71	...	0.87

based on highest average of class discrimination scores, if all C_2^N scores values are considered together the true class distinction ability of a gene may be masked when it performs well to distinguish the specified PC (PC_k) but can be averaged to low value by other irrelevant scores [i.e. $P(g, i, j)$ where $i, j \neq k$]. Therefore, we propose to analyse class discrimination score based on a class specific manner that for each PC_k , only $P(g, k, j)$, where $j = [1, N]$ and $j \neq k$, will be used to train a GSOM (step 2 of Fig. 4). An example of input vector to GSOM is given in Table 1.

Training GSOM from class discrimination scores for each PC (step 3a) gives the advantages of clustering genes have similar class discrimination ability and visually provides their extent of class discrimination strength on the two-dimensional GSOM map (Fig. 6). Weight vectors of nodes on GSOM, or any self-organizing neural networks, acts as mean values of a number of genes with small discrepancy from mean (Herrero *et al.*, 2001). Therefore, we propose that only the genes represented by the node with the strongest positive average class discrimination score (i.e. one versus all other classes) are used as predictor genes for the specified PC (step 3b). This allows a variable number of genes to be selected by the algorithm itself to act as predictor genes and participate in weighted voting, as opposed to fixed number of predictors used by Golub *et al.* Although only one node is being selected on the GSOM, neighbouring nodes of the selected genes are also well suited for class discrimination, thus should also be of interest to the analysts and should not be discarded from presentation.

The third phase of the approach aims at refining the PCs. Since the predictor genes have been identified in the second phase, it is necessary to verify that if predictions yielded by these genes agrees only with the original partitioning. Class prediction of each sample is determined using weighted voting (step 4). In weighted voting, when one sample needs to be associated to a class, each predictor gene for the specific PC has a vote to decide the sample's likelihood of being in the PC. For each of PC_i 's predictor gene g , the vote is calculated as:

$$v(g) = P(g, i, j) \times \left[x(g) - \frac{\mu_i(g) + \mu_j(g)}{2} \right]$$

where i is the PC under question and for all other PCs j , and $x(g)$ is the expression value of gene g for the sample being tested. All votes are then summed to give the overall prediction score, but since the number of predictor genes is variable for each PC, the prediction score is evaluated by calculating the average vote per gene. The PC with the highest prediction

score is the PC for this sample obtained from weighted voting (PC^{new}). Modified from Golub's approach, this measure is found to be reliable and more suitable in this investigation.

Prediction strength (PS) for a sample, a measure that reflects the confidence of prediction result, is also evaluated in a one-versus-all fashion:

$$PS = \frac{v_{\text{for}}(i) + \sum v_{\text{against}}(j)}{v_{\text{against}}(i) + \sum v_{\text{for}}(j)}$$

where i is the nominated 'winning class' (with strongest vote), $v_{\text{for}}(i)$ is the number of votes for the class and $v_{\text{against}}(j)$ is the number of votes against all other classes ($j \neq i$). The sum of which indicates the 'winning strength' and $v_{\text{against}}(i)$ and $v_{\text{for}}(j)$ are summed to give the 'losing strength'. Low PS $PS < \alpha$ indicates an uncertain prediction. A value of $PS < 0.2$ is a good indication that the prediction is uncertain, and is independent of number of predictors used. If the predictions from weighted voting disagree with original clustering from dynamic SOM tree, the predictor gene identification process needs to be iterated to find new sets of genes that best distinguish classes, thus refining the sets of predictor genes as well as the partitioning of data. Iteration will continue until new predictions are the same as predictions in previous iteration.

Evaluating clustering results

The quality of clusters can be analysed in terms of *purity* and *efficiency* (Getz *et al.*, 2000) by supervised testing. However, this is only used to evaluate the clustering quality produced by the proposed method and is not used in any way to identify cancer classes or marker genes (described in Fig. 4).

Purity and efficiency of clusters, defined as

$$\text{purity}(s|c) = \frac{|s \cap c|}{|s|}$$

$$\text{efficiency}(s|c) = \frac{|s \cap c|}{|c|}$$

reflect the extend to which assignment of the samples in class s corresponds to the samples in cluster c . Purity indicates the proportion of samples in class s being clustered correctly in cluster c , so that if all samples of the same class are clustered into one cluster, purity will be 100%. Efficiency evaluates the proportion of cluster c is of sample class s , so that if a cluster contains only samples of the class, the cluster will be 100% efficient.

SIMULATION RESULTS

Class and marker gene discovery from leukemia microarray data

The leukemia data, available on and obtained from MIT's website (Golub *et al.*, 1999), contains two data sets one for training the model and the other for testing. The training data set contains microarray data from 38 samples (patients) and each

Table 2. Summary of quality and contents of PC from level 1 analysis of leukemia data

Level 1	PC_1	PC_2	PC_3	PC_4	Max. purity
ALL-B	16	2	1	0	0.84
ALL-T	1	7	0	0	0.88
AML	1	0	2	8	0.73
Max. eff.	0.89	0.78	0.67	1.0	

sample is composed of 7129 genes' expression levels. The test data set also contains gene expression levels for the same 7129 genes, but from a difference of 34 samples. All of the patients were diagnosed to have one of the three Acute Leukemia types, namely Acute Lymphoblastic Leukemia B-Cell (ALL-B), Acute Lymphoblastic Leukemia T-Cell (ALL-T) and Acute Myelogenous Leukemia (AML).

The original data sets are raw data therefore requires some preprocessing. Data are first preprocessed by replacing values greater than 20 000 with 20 000 and values less than 20 with 20. Second, we apply variation filtering so that only genes with clear highs and lows in expression pattern are selected. Such filtering involves removal of genes with difference between maximum and minimum values (i.e. among all samples) less than 100 and ratio of maximum over minimum less than three from the data set. The data is then normalized before using dynamic SOM tree. After preprocessing, 5855 genes satisfy the filtering criterions and remain in the data set.

We applied the dynamic SOM tree clustering (i.e. unsupervised training) to the preprocessed train data set containing the expression patterns of these 5855 genes (step 1 of pseudo-code) and obtained four clusters/PCs. The PCs are verified by examining their contents and summarized in Table 2. Note that the verification is only performed to validate the clustering ability of the dynamic SOM tree and has no effect on subsequent procedures. PC_1 , PC_2 and PC_4 can be identified as PCs representing 84% of ALL-B, 88% of ALL-T and 73% of AML samples (purity) respectively. PC_3 , unfortunately consist of one ALL-B and two AML samples, do not represent the majority of any known class. The efficiencies of the clusters, PC_1 (89%), PC_2 (78%), PC_3 (67%) and PC_4 (100%), are estimated considering the presence of samples of other classes with respect to its representing class. Initial clustering results and actual leukemia class of samples can be interpreted as in (Fig. 5).

It seems that it is easier to separate between ALL-T and AML patterns using unsupervised learning, whereas it is difficult to separate ALL-B from AML and ALL-B from ALL-T. From a biological perspective, it is likely that since B-cell and myeloid cells all originated from blast cells of the bone marrow they have more genes expressed in common.

To demonstrate the granularity of the dynamic SOM tree algorithm, two tests were performed using only ALL-B and

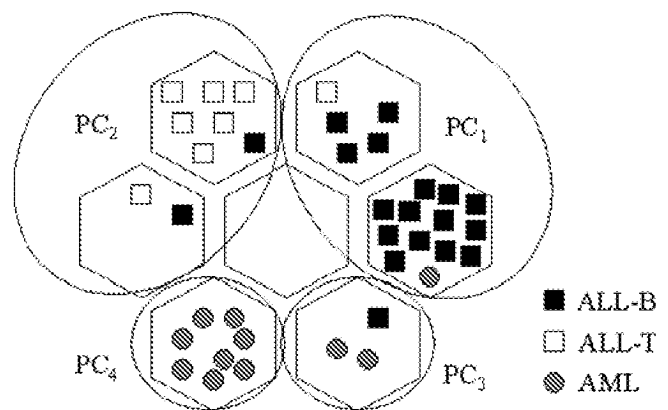


Fig. 5. Topological and clustering relationships between PCs and actual classes produced by using dynamic SOM tree.

Table 3. Marker genes for diagnosing leukemia at St. Jude Children's Research Hospital.

Leukemia type	Highly sensitive gene	Highly specific gene
ALL-B	CD19	Cytopl. CD79a
ALL-T	CD7	CD3
AML	CD13 or CD33	Cytopl. Myeloperoxidase

only AML data, respectively. Constructing dynamic SOM tree trained from ALL-B data yielded two clusters. The first cluster contains 16 samples, 15 of which are the ones in PC_1 . Similar result was achieved with AML samples, where two clusters are formed with eight and three samples. All eight samples are contained in PC_4 . The deviation in ALL-B clustering result with respect to PC_1 may be due to various reasons. One reason could be that the total number of genes remained after preprocessing the data belonging to a single class is different from that of the complete data set.

In the next stage of analysis (step 3 of pseudo-code) the pair-wise correlation values are computed for all genes and a GSOM is trained for all pair-wise correlations relating to PC_i ($i \in [1, 4]$) (step 3a of pseudo-code). At this stage, genes are grouped by their strength in distinguishing PC_i from the rest on the GSOM. The highest average distinction node is selected and all the genes it represents are used as class predictors for this class (step 3b of pseudo-code).

A list of clinically used marker genes for leukemia diagnosis at St Jude Children's Research Hospital is provided in Table 3 (Pui and Evans, 1998). It is interesting to see that many clinically used marker genes are either in the vicinity of the predictors on the trained GSOM (Fig. 6) or are included in our predictor genes. For example, CD7, which is highly sensitive for T lineage cells of ALL, is included along with many other T lineage cell specific genes as the predictors for PC_2 , which contains mainly ALL-T samples. In the group

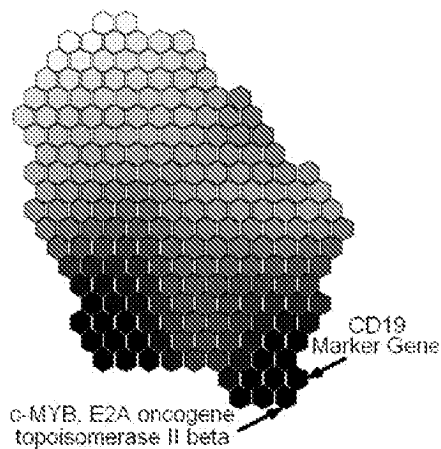


Fig. 6. GSOM trained from class discrimination scores for PC_1 showing locations of predictor genes and the marker gene. Darker shades indicates higher average class discrimination score value.

of predictor genes for PC_1 that contains mostly of ALL-B samples, a few interesting aspects are noted. Firstly, although CD19, the marker gene for ALL-B, is not included in the group of predictor genes discovered here, it is located in their vicinity on the GSOM (in fact, right next to the predictor node). Secondly, a number of genes that are of significance to ALL-B are included. Genes like known oncogenes c-MYB, E2A, principle antileukemic drug target—topoisomerase II beta, and a few other genes relevant to S-phase cell cycle and transcription are used as predictor genes for PC_1 , and the result is consistent with Golub's findings.

AML samples that are marked as uncertain [by French-American-British (FAB) classification] are either of AML subtype M2 or M5, which have increased frequency of CD13 and CD14 expression (Amirghofran *et al.*, 1999). However, the number of M2 and M5 samples in AML is relatively small in the data set, with only two M5, five M2, three M1 and one M4 samples. Therefore a strong correlation cannot be made to their clinically known marker genes. The reason why CD14 was not identified as a predictor gene even though CD14 expresses in 75% of the M5 cases has been investigated. It was removed in preprocessing due to the fact that it only expressed in one of the two M5 samples.

Predictions are made and PSs are evaluated for all original samples (step 4 of pseudo-code). In this application, most of the samples remain in the same PC after the iteration described in step 5 of Figure 4 [$PC(s_i)^{new} = PC(s_i)$ as shown in step 5c].

We revised the predictors (step 5a of the pseudo-code) to obtain our final predictor genes for level 1. Thus, repeating steps 2–5 of Figure 4. In the above-described process of finalising the predictors, it is found that not only the clinically known marker genes are still present, but also some of the seemingly less relevant genes are excluded. Also, in the

Table 4. Summary of contents of PCs from level 2 analysis of leukemia data (step 6 of pseudo-code)

Level 2	PC_1	PC_2	PC_3
ALL-B	3	1	0
ALL-T	0	1	0
AML	0	1	3

case of PC_2 predictors, CD3 (a marker gene that expresses highly specific to ALL-T patients) and the second instance of CD7 related gene (the first instance is already discovered as a predictor in previous steps) are included. This shows that the iteration not only refines the predictor genes, but also improves them.

However, the analysis of the PS (in step 5b) indicates that all incorrectly clustered samples, which include all samples predicted to be in PC_3 and a few other incorrectly clustered samples in other PC, have low PS [PS < α , and $\alpha = 0.2$ according to Golub *et al.* (1999)]. Therefore further analysis is performed on these uncertain samples by repeating the process, thus refining predictions.

Nine samples at level 2 (four ALL-B, one ALL-T and four AML samples) that have been marked uncertain at level 1 step 6 (PS < 0.2) are analysed further. Variation filtering is again applied to remove genes with negligible variation (max/min < 3) from analysis and only 4132 out of 5855 genes satisfy the criterion. Clustering of uncertain samples using dynamic SOM tree produces three PCs (Table 4). PC_1 contains three ALL-B samples, PC_2 contains a mixture of samples with one sample from each acute leukemia type, and PC_3 contains three AML samples. Predictors for PC_1 consist of some interesting genes like CD9 antigen, Interleukin 1 beta and CD63 antigen. CD9 antigen is expressed during early stage B-cell differentiation or activation. The main function of Interleukin 1 beta is on B-cell maturation and proliferation. CD63 antigen and another lysosome-associated membrane glycoprotein are associated with early stages of tumour progression and may play a role in growth regulation. Furthermore, in PC_3 , predictor HOXB2 is also present. HOXB genes are known to express in acute myelogenous leukemia and turned off in chronic myelogenous leukemia (Magli *et al.*, 1997).

Predictions are made for the uncertain samples and PSs are low (two of three samples have PS < 0.2) for PC_2 (step 5c). The two levels of analysis can be related to each other from two possible ways. One method is to investigate whether the PCs analysed in level 2 has strong correlation with any of the PC in level 1, in terms of expression pattern. The other method is to examine whether the pathological or pharmacological functions of the predictor genes in level 2 are comparable to those of predictors in level 1. In this work, the latter method

Table 5. Summary of prediction results using test samples

	PC_1	PC_2	PC_3	PC_4
<i>Level 1</i>				
Samples	7	0	22	5
Correct	7	—	—	5
<i>Level 2</i>				
Samples	6	9	7	
Correct	6	—	6	

was used. If neither of those methods confirms/supports relationship between the PCs in level 2 and the PCs in level 1 then it should be considered as a separate class.

Following the construction of class and predictor models we can now use the test data set to evaluate prediction of new samples. Test data set contains 34 acute leukemia samples, 19 of which are ALL-B samples, 14 AML samples and only one ALL-T sample. Testing was performed using the previously identified predictor genes from training data and the results are presented in Table 5. Test results using predictors found in level 1 of training data predicted only 12 samples correctly. The remaining 22 samples predicted to be in PC_3 have $PS < 0.2$ and therefore considered for level 2. Uncertain test samples are then further tested using the predictor genes found in level 2 in training data. Thirteen more samples are predicted with $PS > 0.2$ at level 2, with one sample (the only ALL-T sample) predicted incorrectly. The remaining samples (three AML samples and six ALL-B samples) are predicted to be in PC_2 with $PS < 0.2$. Unfortunately, the training data did not allow us to go to level 3 and therefore those nine samples could not be processed further. Altogether, out of 34 test samples, one prediction was incorrect and nine were uncertain, thus the prediction accuracy is around 71%.

Class discovery of colon cancer microarray data

The colon cancer data is obtained from Weizmann Institute of Science's website (Getz *et al.*, 2000). There are 62 samples in the data set, 40 of which are tumorous samples and 22 normal samples, diagnosed using two different protocols. The data is available in a processed form, but follows a different filtering/selection process to the leukemia data. The data is filtered such that only the most-expressed 2000 genes are used and then normalized each sample by dividing the expression value by the sum of expression values. This filtering process involves, in our case, a few undesirable characteristics. First, the 2000 genes with maximum expression levels do not necessarily include marker genes for validating our result. Second, as the preprocessing method is different to our previously used leukemia data, it is hard to compare the outcome with the previous application. However, the data set is still useful to illustrate class discovery part of the algorithm. There may have other biological limitations to the colon cancer data. Genes

Table 6. Summary of contents of PCs from level 1 analysis of colon cancer data

Level 1	PC_1	PC_2	Max. purity
Tumorous	37	3	0.93
Normal	6	19	0.76
Max. eff.	0.86	0.86	

that express abnormally depends on the stage of tumour progression. Therefore, clustering with gene expression patterns becomes more difficult. And it is less likely to predict tumorous samples with a fixed set of predictor genes. Different stages of tumour progression actually have different marker genes. Mutation of the APC gene occurs in the early stage of tumour development. As small adenomas progress to larger forms, mutations in the transforming oncogene *ki-ras* occur frequently. In later stages necrosis or inactivation of other tumour suppressing genes like Deleted in Colorectal Cancer and p53 occur (Rumsby and Davies, 1995). Unfortunately, such marker genes are not present in the data set, thus we are unable to test whether these genes do influence the clustering or demonstrate their correlation to the predictor genes we used.

The dynamic SOM tree, trained using the colon cancer data (step 1 of Fig. 4), automatically suggests that two clusters or PCs are formed and the contents of each cluster are presented in Table 6. This is a good indication of robustness of the dynamic SOM tree, since the aim of clustering cancerous samples here is to distinguish between tumorous patients and normal patients. PC_1 contains 43 samples of which 37 tumorous and six normal. PC_2 contains 19 samples of which 16 normal and three tumorous. Of the six normal samples clustered in PC_1 , three are diagnosed using protocol A, and all the tumorous samples clustered in PC_2 are diagnosed using protocol B. Perhaps it is possible to infer that protocol A is a stricter test, such that clustering gene expression data can identify all tumorous samples diagnosed using it.

Predictor genes are found from GSOMs trained from class distinction values (steps 2–3). Fourteen samples are marked as uncertain (again, $PS < 0.2$) and require next level of analysis. Out of the 48 samples with strong prediction strength only four cases (three tumorous and one normal) are predicted incorrectly. Predictions by weighted voting produces identical result as clustering results using dynamic SOM tree (step 4), thus predictors need not be refined (step 5b).

PC_1 predictors contain a number of genes that produce tumour-associated proteins like, for example, thioredoxin and putative NDP kinase. Compared to normal tissue, over half of the human primary lung, colon and gastric cancers over-express thioredoxin. Therefore thioredoxin is often conceived as an oncogene (Grogan *et al.*, 2000). Putative NDP kinase is

Table 7. Summary of contents of PCs from level 2 analysis of colon cancer data

Level 2	PC_1	PC_2	PC_3
Tumorous	4	2	1
Normal	2	4	1

likely to be found in reduced amount in tumour cells of high metastasis potential (Wang *et al.*, 1993).

Clustering the 14 uncertain samples (step 6 or step 1 of level 2), again using dynamic SOM tree, came to three clusters or PCs (Table 7). There are six samples in PC_1 , 4 tumorous and two normal; six samples in PC_2 , four normal and two tumorous; two samples in PC_3 , one normal and one tumorous. As we can see that strong distinction between tumorous and normal samples cannot be made clearly here. This is also reflected upon the prediction strength, where 6 out of the 14 samples still have low prediction strengths having two uncertain samples in each cluster.

In the training set, 52 samples can be predicted to be belonging to a PC that represents the majority of a class. Also, the maximum purities and efficiencies of the PCs are quite high, which indicates that a large majority of one class is clustered in one single cluster.

DISCUSSION

The proposed algorithm automatically (and unsupervised) identified suitable number of clusters and likely marker genes for each data set used. In class discovery, the dynamic SOM tree identified three major and one minor cluster for the leukemia data that has three actual classes, and two clusters for the colon cancer data that has two actual classes; thus indicating the automatically identified number of clusters is appropriate. Likely marker genes identified by the algorithm for each of the cancer classes have strong relevancy and many are backed by existing literature. Prediction accuracies for the training sets are 84% for colon cancer data and 92% for leukemia data. Prediction accuracy for the leukemia test set is 71%.

We understand that a new identified leukemia type known as Mixed Lineage Leukemia (MLL) was identified after the leukemia data used in this work has been collected. MLL has quite distinct expression pattern from known leukemia types (Armstrong *et al.*, 2002). It is worth analysing further to determine whether the small cluster, PC_3 , contains MLL samples. Since at the time of diagnosis of the patients the MLL was not identified, therefore they may have been assigned to one of the known leukemia types.

The simulations were carried out on a P3 800 MHz PC running Sun's JAVA VM 1.3. Training GSOMs consumes most of the analysis time. On an average, training GSOMs using

the two data sets take approximately 21 s for each layer of the dynamic SOM tree and approximately 220 s for class discrimination scores.

ACKNOWLEDGEMENTS

We thank Gaddy Getz of Weizmann Institute of Science, Israel, for valuable help in providing the colon cancer data set and details of preprocessing used. We also thank the referees for their valuable comments.

REFERENCES

- Alahakoon, D., Halgamuge, S.K. and Srinivasan, B. (2000) Dynamic self-organising maps with controlled growth for knowledge discovery. *IEEE Transactions on Neural Networks, Special Issue on Knowledge Discovery and Data Mining*, **11**.
- Alizadeh, A.A., Eisen, M.B. *et al.* (2000) Different types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
- Alon, U., Barkai, N. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. **96**, 6745–6750.
- Amirghofran, Z., Shaseddin, A. *et al.* (1999) Immunophenotypic analysis in iranian patients with acute myelogenous leukemia: correlation with leukemia subgroups. *Iran. J. Med. Sci.*, **24**, 40–44.
- Armstrong, S.A., Staunton, J.E. *et al.* (2002) Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.*, **30**, 41–47.
- Ben-Dor, A., Bruhn, L. *et al.* (2000) Tissue classification with gene expression profiles. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, Tokyo, Japan.
- Blackmore, J. and Miikkulainen, R. (1995) Visualizing high-dimensional structure with the incremental grid growing neural network. In *Proceedings of the 12th International Conference on Machine Learning*.
- Campbell, C., Li, Y. and Tipping, M. (2001) An efficient feature selection algorithm for classification of gene expression data. In *NIPS 2001 Workshop on Machine Learning Techniques for Bioinformatics*, Vancouver, Canada.
- Dudoit, S., Fridlyand, J. and Speed, T. (2000) Comparison of discrimination methods for the classification of tumors using gene expression data. Tech. report, Berkeley.
- Fritzke, B. (1994) Growing cell structures—a self-organising network for unsupervised and supervised learning. *Neural Networks*, **7**, 1441–1460.
- Fritzke, B. (1995) A growing neural gas network learns topologies. In Tesauro, G., Touretzky, D.S. and Leen, T.K. (eds), *Advances in Neural Information Processing Systems 7*. Cambridge, MA: MIT Press.
- Getz, G., Levine, E. and Domany, E. (2000) Coupled two-way clustering analysis of gene microarray data. In *Proceedings of National Academy of Science*, Vol. 97. pp. 12079–12084, USA.
- Golub, T.R., Slonim, D.K. *et al.* (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–536.
- Grogan, T.M., Fenoglio-Prieser, C. *et al.* (2000) Thioredoxin, a putative oncogene product, is over-expressed in gastric carcinoma and

- associated with increased proliferation and increased cell survival. *Hum. Pathol.*, **31**, 475–481.
- Guyon, I., Weston, J. et al. (2000) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–420.
- Hartuv, E., Schmitt, A. et al. (2000) An algorithm for clustering cDNA fingerprints. *Genomics*, **66**, 249–256.
- Herrero, J., Valencia, A. and Dopazo, J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.
- Hsu, A., Alahakoon, D., Halgamuge, S.K. and Srinivasan, B. (2000) Automatic clustering and rule extraction using a dynamic som tree. In *Proceedings of the 6th International Conference on Automation, Robotics, Control and Vision*, Singapore.
- Hsu, A. and Halgamuge, S.K. (2003) Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualisation. *Int. J. Approximate Reasoning*, **23**, 259–279.
- Kaski, S. (2001) *Advances in Self-Organising Maps*. Springer, London.
- Keller, A., Schummer, M., Hood, L. and Ruzzo, W. (2000) Bayesian classification of DNA array expression data. Tech report, Tech Report, University of Washington.
- Khan, J., Wei, J.S. et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.
- Kohonen, T. (1997) *Self-Organising Maps*. Springer, Berlin.
- Magli, M.C., Largman, C. and Lawrence, H.J. (1997) Effects of hox homeobox genes in blood cell differentiation. *J. Cell Physiol.*, **173**, 168–177.
- Pui, C.H. and Evans, W.E. (1998) Drug therapy: Acute lymphoblastic leukemia. *New Engl J. Med.*, **339**, 605–615.
- Ramaswamy, S., Tamayo, P. et al. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci., USA*, **98**, 15149–15154.
- Rumsby, B. and Davies, M. (1995) Genetic events in the development of colon cancer. *Food Chem. Toxicol.*, **33**, 328–330.
- Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a DNA microarray. *Science*, **210**, 467–470.
- Shalon, D., Smith, S.J. and Brown, P.O. (1996) A DNA microarray system for analysing complex DNA samples using two-color fluorescent probe hybridisation. *Genome Res.*, **6**, 639–645.
- Sharan, R. and Shamir, R. (2000) Click: A clustering algorithm for gene expression analysis. In Miyano, S., Shamir, R. and Takagi, T. (eds), *Currents in Computational Molecular Biology*, Universal Academy Press, pp. 6–7.
- Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy*. Freeman, San Francisco.
- Tamayo, P., Slonim, D. et al. (1999) Interpreting patterns of gene expression with self-organising maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci., USA*, **96**, 2907–2912.
- Tibshirani, R., Hastie, T. et al. (1999) Clustering methods for the analysis of DNA microarray data. Tech. report, Tech. Report, Stanford University.
- Toronen, P., Kolehmainen, M., Wong, G. and Castren, E. (1999) Analysis of gene expression data using self-organizing maps. *FEBS Lett.*, **451**, 142–146.
- Triche, T.J., Cavazzana, A.O. et al. (1988) N-myc protein expression in small round cell tumors. *Prog. Clin. Biol. Res.*, **271**, 475–85.
- Vesanto, J. and Alhoniemi, E. (2000) Clustering of the self-organising map. *IEEE Trans. Neural Networks*, **11**, 586–600.
- Wang, L., Patel, U. et al. (1993) Mutation in the nm23 gene is associated with metastasis in colorectal cancer. *Cancer Res.*, **53**, 717–720.
- Xing, E., Jordan, M. and Karp, R. (2001) Feature selection for high-dimensional genomic microarray data. In *Proceedings of the 18th International Conference on Machine Learning*, Massachusetts, USA.
- Zhang, H., Yu, C., Singer, B. and Xiong, M. (2001) Recursive partitioning for tumor classification with gene expression microarray data. *Proc. Natl Acad. Sci., USA*, **98**, 6730–6735.