Machine Learning, 52, 45–66, 2003

Analysis and Visualization of Gene Expression Microarray Data in Human Cancer Using Self-Organizing Maps

SAMPSA HAUTANIEMI sampsa.hautaniemi@tut.fi
OLLI YLI-HARJA yliharja@cs.tut.fi
JAAKKO ASTOLA jaakko.astola@tut.fi
Institute of Signal Processing, Tampere University of Technology, PO Box 553, 33101 Tampere, Finland

PÄIVIKKI KAURANIEMI paivikki.kauraniemi@uta.fi
ANNE KALLIONIEMI anne.kallioniemi@uta.fi
Laboratory of Cancer Genetics, Institute of Medical Technology, University of Tampere and Tampere University Hospital, FIN-33520 Tampere, Finland

MAIJA WOLF Maija.Wolf@Helsinki.fi
Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, USA; Medical Biotechnology Group, VTT Technical Research Centre of Finland and University of Turku, PO Box 106, 20521 Turku, Finland

JIMMY RUIZ jruiz@siumed.edu
SPYRO MOUSSES smousses@tgen.org
Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, USA

OLLI-P. KALLIONIEMI Olli.Kallioniemi@vtt.fi
Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, USA; Medical Biotechnology Group, VTT Technical Research Centre of Finland and University of Turku, PO Box 106, 20521 Turku, Finland

Editors: Paola Sebastiani, Isaac S. Kohane and Marco F. Ramoni

Abstract. cDNA microarrays permit massively parallel gene expression analysis and have spawned a new paradigm in the study of molecular biology. One of the significant challenges in this genomic revolution is to develop sophisticated approaches to facilitate the visualization, analysis, and interpretation of the vast amounts of multi-dimensional gene expression data. We have applied self-organizing map (SOM) in order to meet these challenges. In essence, we utilize U-matrix and component planes in microarray data visualization and introduce general procedure for assessing significance for a cluster detected from U-matrix. Our case studies consist of two data sets. First, we have analyzed a data set containing 13,824 genes in 14 breast cancer cell lines. In the second case we show an example of the SOM in drug treatment of prostate cancer cells. Our results indicate that (1) SOM is capable of helping finding certain biologically meaningful clusters, (2) clustering algorithms could be used for finding a set of potential predictor genes for classification purposes, and (3) comparison and visualization of the effects of different drugs is straightforward with the SOM. In summary, the SOM provides an excellent format for visualization and analysis of gene microarray data, and is likely to facilitate extraction of biologically and medically useful information.

Keywords: bioinformatics, gene expression in human cancer, self-organizing map

1. Introduction

Microarray experiments typically produce data for tens of thousands of genes simultaneously. Comparing gene expression profiles of this scale, especially when comparing data from multiple samples or experimental conditions that result in high dimensional data, presents a formidable challenge in both visualization and pattern recognition. Although several statistical approaches have been developed to address this issue (e.g. Parmigiani et al., 2002; Raychaudhuri, Stuart, & Altman, 2000; Wall, Dyck, & Brettin, 2001), they often do not permit a representative visualization of the data.

Clustering methods have become standard tools for microarray data analysis because they enable identification and representative visualization of biologically relevant groups of genes or samples. One of the most frequently used clustering algorithm in the microarray data context is hierarchical clustering (Eisen et al., 1998). Hierarchical clustering results in a dendogram, where genes and samples are arranged according to their pair-wise similarity. Hierarchical clustering is known to suffer from many severe problems such as low noise tolerance and sensitiveness to the choice of the linkage method (Mangiameli, Chen, & West, 1996). Furthermore, (Gibbons & Roth, 2002) report that clusters of genes derived from single- and average-linkage hierarchical clustering tend to be worse than random results. Therefore, need for an alternative for hierarchical clustering methods is evident. In this study we have employed the self-organizing map (SOM) in gene expression data analysis. The SOM is an unsupervised neural network algorithm, which has been used with great level of success in various clustering and visualization tasks (see Kaski, Kangas, & Kohonen, 1998). Moreover, several studies report that for a noisy data set the SOM outperforms hierarchical clustering and many other clustering methods in various critical areas such as noise tolerance, speed and robustness (Mangiameli, Chen, & West, 1996; Chen et al., 2002; Gibbons & Roth, 2002).

The purpose of this study is analysis and visualization of gene expression data obtained from cDNA microarray experiments on human cancer using the SOM. The SOM has been used earlier in clustering gene expression patterns from yeast or *C.elegans* (e.g. Tamayo et al., 1999; Törönen et al., 1999; Hill et al., 2000; Kaski et al., 2001) and recently also in cancer data sets (e.g. Chen, Chang, & Huang, 2000; Ramaswamy et al., 2001). However, while these studies show that the SOM is capable of finding biologically meaningful clusters, many of them do not fully utilize the visualization capabilities of the SOM. In addition, the analysis of gene expression data obtained from human cancers differs from the analysis of yeast or *C.elegans* partly due to the fact that our current knowledge on the biology of yeast (or *C.elegans*) is much more advanced than that of human cancer.

The order of this study is as follows. In Section 2 we describe the SOM algorithm, the U-matrix and the component plane representation effectively used in visualization. We also introduce a general procedure for computing p-value for a cluster. Section 3 involves discussion on general aspects of the application of the SOM to the microarray data analysis. In Section 4 we apply the SOM to two cancer data sets. First data set contains initially 13,824 genes from 14 breast cancer cell lines. In the second data case we demonstrate how the SOM could be used in identifying putative targets for therapeutic intervention. In Section 5 we compare the U-matrix and the component plane representation to another way to visualize the results presented in Tamayo et al. (1999). In Section 6 we discuss applications and limitations of the SOM, as well as future directions.

2. Brief description of the principles of the SOM

The SOM belongs to vector quantization algorithms. Basically the SOM transforms high dimensional input data into a lower dimensional display, which is normally two-dimensional. The elements of the display are called *neurons* (also referred as map units or cells). Each neuron contains a *reference vector* (also codebook vector), which usually has same dimension as the gene expression patterns. Optimization of the SOM is based on the quantization error in the reference vector space. Let $\mathbf{x} \in \mathbb{R}^{n\times 1}$, where n is dimension of the gene expression pattern, be an gene expression pattern drawn from the input data set. Further, let $\mathbf{m}_i \in \mathbb{R}^{n\times 1}$ be i:th (initialized) reference vector. The input gene expression pattern is connected to all neurons and the distances ($d(\mathbf{x}, \mathbf{m}_i)$) between the gene expression patterns and reference vectors are computed. The neuron having the closest reference vector (referred as $\mathbf{m}_c$) to the current gene input pattern is declared as a winner:

$$d(\mathbf{x}, \mathbf{m}_c) = \min_i(d(\mathbf{x}, \mathbf{m}_i)). \tag{1}$$

Normally the winner is surrounded by a (topological) neighborhood region, N_c, and all neurons belonging to the neighborhood are permitted to update. The intensity of the updating is controlled by neighborhood function, h_{ci}, which is centered to the neuron having closest reference vector $\mathbf{m}_c$.

Equation (1) defines a *quantization error*. If the number of reference vectors is L, a *distortion measure* can be defined by:

$$e(\mathbf{x}) = \sum_{i=1}^{L} h_{ci} \cdot d(\mathbf{x}, \mathbf{m}_i). \tag{2}$$

Using Eqs. (1) and (2) an *average expected distortion measure* can be defined as

$$E = \int e(\mathbf{x})p(\mathbf{x})\,d\mathbf{x} = \int \left(\sum_{i=1}^{L} h_{ci}(t) \cdot d(\mathbf{x}, \mathbf{m}_i)p(\mathbf{x}) \right) d\mathbf{x}, \tag{3}$$

where $p(\mathbf{x})$ denotes the probability density function of $\mathbf{x}$. The SOM can be defined by finding a set of $\mathbf{m}_i$ that minimizes average distortion measure (Eq. (3)). However, the minimum of Eq. (3) cannot usually be found in closed form and in practice we do not have the density $p(\mathbf{x})$. The best approximate solution is based on Robbins-Munro stochastic approximation. The idea of the Robbins-Munro stochastic approximation is to find the optimum by taking a gradient of Eq. (3) recursively. Starting with an arbitrary initial value, the sequence $\mathbf{m}_i(t)$ converges to the neighborhood of the optimum. Now we have $\mathbf{x}(t)$, $t = 1, 2, \ldots$, a sequence of input samples and $\mathbf{m}_i(t)$ recursive defined sequence of reference vector $\mathbf{m}_i$, then

$$e(t) = \sum_{i=1}^{L} h_{ci}(t) \cdot d(\mathbf{x}(t), \mathbf{m}_i(t)) \tag{4}$$

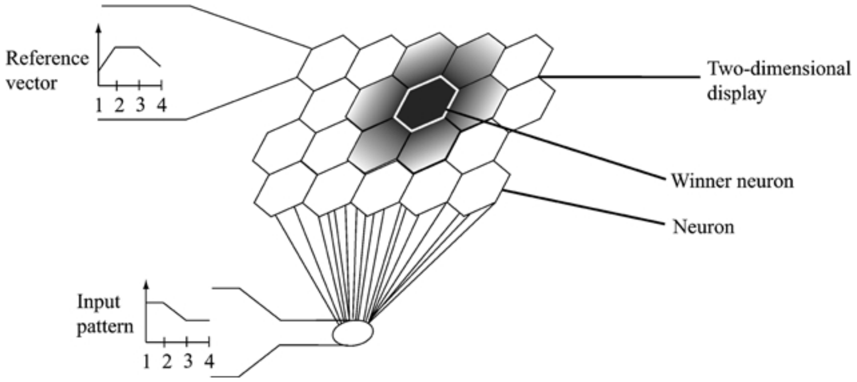

Figure 1. Idea of the SOM. All neurons contain a reference vector, whose dimension is the same as the dimension of the input data. Gene expression pattern is compared to all reference vectors and the neuron containing the closest reference (black with white boundaries) is permitted to update with neurons belonging to the neighborhood region (shaded).

is a random variable, and the sequence

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) - \mathbf{G}_t \nabla_{m_{i(t)}} e(t), \tag{5}$$

where $\mathbf{G}_t$ is *gain matrix*, converges to approximately optimum value. In this study we choose $\mathbf{G}_t = \alpha(t)\mathbf{I}$, where $\alpha(t)$ is called a learning factor (Kohonen, 2001).

Equations (4) and (5) can be taken as definition of a class of SOM algorithms. The reason why we chose vector quantization approach to define the SOM is that Eq. (5) allows a simple way to implement new distance measures to the SOM. For example, if $d(\mathbf{x}, \mathbf{m}_i)$ is defined to be Euclidian distance, the "traditional" SOM algorithm follows (Kohonen, 2001)

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t) \cdot h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)]. \tag{6}$$

The idea of the SOM algorithm is illustrated in figure 1, where the size of the output layer is 4 × 5 (i.e. there are 20 neurons in the display). Let us assume that we have measured the expression levels of the genes from four samples. As a consequence, the gene expression patterns are four-dimensional as are the reference vectors in the output layer. A gene expression pattern is drawn from the data set and distances between the gene expression pattern and all reference vectors are computed. The closest neuron, winner neuron (black neuron with white edge in figure 1), is permitted to update as well as the neurons belonging to the neighborhood (shaded region in figure 1).

2.1. Batch learning algorithm

The SOM defined by Eq. (5) is called the *sequential* (also on-line, stochastic or incremental) learning algorithm since reference vectors are updated after a single input vector is presented. Another way to update the reference vectors is the *batch* learning algorithm. In the batch algorithm update is done using all the input vectors. The pseudo-code for the batch SOM is (Kohonen, 2001):

```
Initialize reference vectors $\mathbf{m}_i$
Repeat until converged
    For each neuron $i$ collect a list ($\boldsymbol{S}_i$) of input vectors $\mathbf{x}$ whose nearest reference vector
        belongs to neuron $i$.
    Update each $\mathbf{m}_i$ using $\boldsymbol{S}_i$.
end
```

The batch algorithm is much faster than the sequential algorithm. Moreover, the batch algorithm, unlike the sequential one, does not suffer from convergence problems due to $\alpha(t)$. General guidelines for deciding whether to use sequential or batch learning algorithm are as follows. Sequential learning algorithm should be used in situations where data set is both redundant (i.e. data contains several samples having identical values) and large. The sequential algorithm is stochastic of its nature meaning that it is less likely trapped to a local minimum than the batch learning algorithm. However, the stochastic nature also makes it sometimes difficult to determine the conditions for convergence while batch learning algorithm converges under simple conditions. In many cases fast computation and guarantee of convergence makes batch learning algorithm more appealing than the sequential learning algorithm (Haykin, 1999).

2.2. Parameters

Although the SOM is an unsupervised algorithm some parameters must be defined before the analysis.

- Topology of the SOM map. Popular alternatives are sheet, cylindrical and toroid topology. Sheet topology is easy to interpret and most common in practice.
- Distance measure for computing the difference between data and reference vectors. Choice of distance measure is depending on the data and purpose of the experiment. We will discuss the distance measure in greater details in Section 3.2.
- Number of neurons. There are no explicit rules for choosing the number of neurons, a heuristic guideline $5 \cdot \sqrt{p}$, where p is number of input vectors, is given in Vesanto et al. (2000).
- Initialization of reference vectors. Initialization can be done in arbitrary many ways. Linear initialization, where the reference vectors are initialized using principal component analysis, has been shown to be a better choice than random initialization (Kohonen, 2001).
- Learning algorithm. If sequential learning algorithm is chosen, one has to decide also the value of the learning factor.
- Neighborhood function and corresponding dynamical parameters.

The neighborhood function plays central role in the SOM algorithm regardless of the type of the learning algorithm. Three frequently used neighborhood functions are presented in figure 2 (Kohonen, 2001; Vesanto et al., 2000).

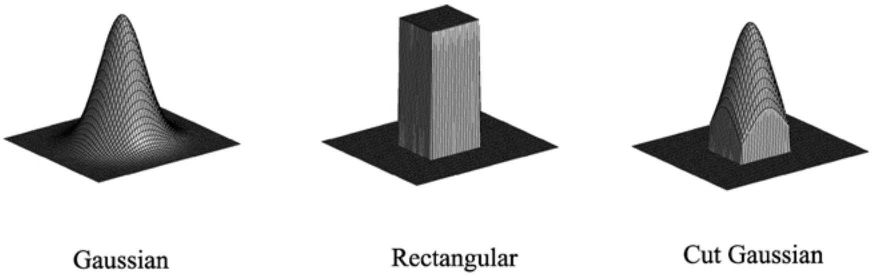

Figure 2. Three neighborhood functions frequently used in updating the reference vectors.

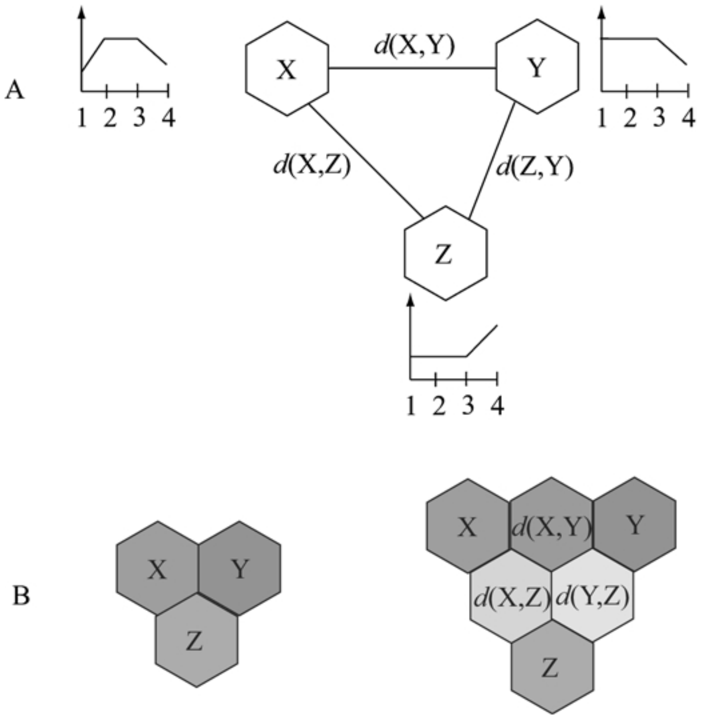

Figure 4. Idea of the U-matrix. (A) an example of three neurons and their reference vectors. (B) Neurons in the two-dimensional graph are colored according to their similarities to adjacent neurons. The result of the U-matrix procedure is an extra element between the neurons. *X* and *Y* are similar to each other while *Z* has a distinct reference vector.

In figure 2 X-Y-plane corresponds to topological order of the neurons and Z-axis is the updating intensity. Rectangular is the simplest neighborhood function and all the neurons belonging to the neighborhood set are updated with same intensity. Gaussian function is a very popular choice for neighborhood function because it is biologically more suitable than the rectangular one (Haykin, 1999).

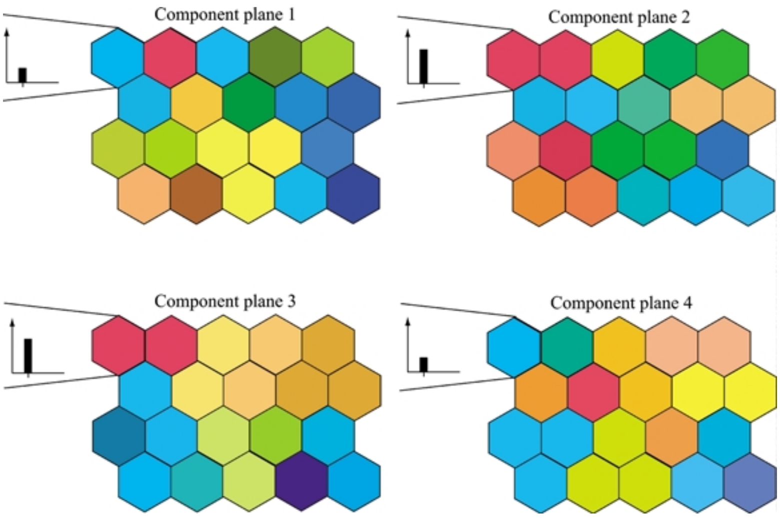

Figure 3. Component plane representation of the SOM map in figure 1. Bar diagrams correspond to the reference vector pattern of the first neuron in the left upper corner in figure 1. Colors code the values of the reference vector in corresponding neuron such that shades of blue mean low expression, yellow and green moderate expression and shades of red high expression.

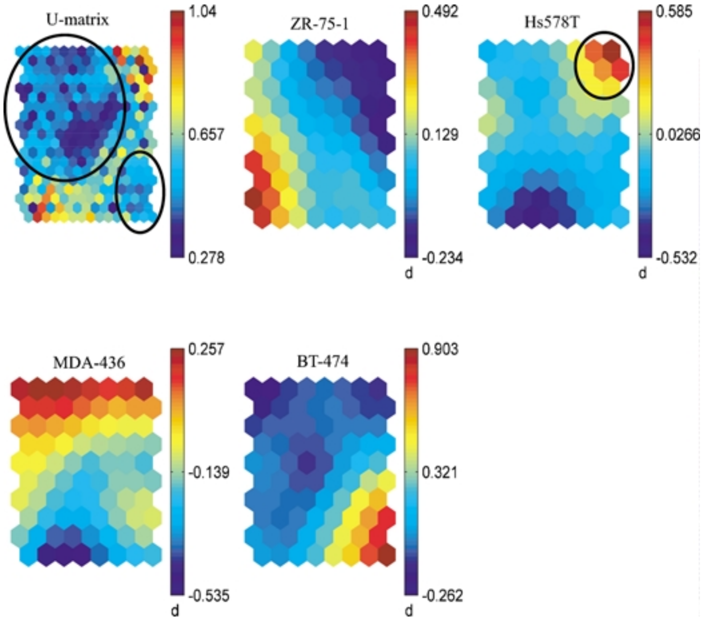

Figure 5. An example of representation of the clustering results. U-matrix tells the distances between the neurons and allows straightforward interpretation of clusters. Two circles in the U-matrix connotate clusters. Component planes allow direct comparison of the samples. Circled neurons in Hs578T contain genes that overexpressed in Hs578T and MDA-436 but underexpressed in ZR-75-1 and BT-474.

2.3. Finding correlations between samples

The two-dimensional display (figure 1) can be divided into component planes thus enhancing the visualization capabilities of the SOM. Component planes are formed from the reference vector by splitting it to n components, where n is dimension of the reference vector. In the context of the microarray data analysis a component plane corresponds to a sample.

Neurons in the component planes are colored with gray levels or color shades. Throughout this study we assume that ratios are constructed by dividing the test sample intensity by the control sample intensity. Accordingly, the shades of red correspond to high expression, the shades of yellow to moderate expression and the shades of blue to low expression. Colored component planes are a powerful tool for visualization of the correlation between the samples. An example of the component plane representation is illustrated in figure 3. The coloring of the left upper-most neuron in every component plane follows the profile given in figure 1. Correlations are readily seen from the component planes. For example, it is evident that low right-most corner has three neurons that consist of genes being underexpressed in all samples. Moreover, the color of the upper left-most neuron suggests that genes in that neuron are highly expressed in samples 2 and 3, but underexpressed in samples 1 and 4. The component plane representation allows straightforward way to compare effects of the drug treatments as shown in Section 4.2.

2.4. Interpreting clusters in the SOM

Elements of the data that are close to each other and comprehend a disjunctive subset are said to belong to the same cluster. As a result of the SOM algorithm, n-dimensional data are presented in two-dimensional display. Elements of the data that are close to each other in $\mathbb{R}^n$ are arranged to neurons close to each other also in the two-dimensional display.

All clustering algorithms share a problem of deciding boundaries of the clusters. The SOM is not an exception. In fact (Ultsch & Siemon, 1989), showed with a simple experiment that clusters cannot be detected in a reliable manner using the two-dimensional display only. In order to solve this problem (Ultsch & Siemon, 1989) developed an unified-distance matrix (U-matrix), which allows easy and straightforward detection of the clusters. The idea of the U-matrix is to compute the difference between two adjacent reference vectors and illustrate the difference as an extra neuron between the original neurons in the U-matrix figure. An example of the construction of the U-matrix is given in figure 4, where distances between three top left-most neurons in figure 3 are computed. If the distance between the neurons is small, extra neuron depicting the distance is colored with shades of blue, and if the distance is big with shades of red. Clusters can be seen as ‘valleys’ (neurons colored with shades of blue) separated by ‘hills’ (neurons colored with shades of red). Relative location of the clusters in the U-matrix reflect their similarities in $\mathbb{R}^n$. The higher the hill (shades of red) the more dissimilar the cluster are in $\mathbb{R}^n$. Distance $d(X, Y)$ can be chosen arbitrarily, but usually average is used. The process for creating the U-matrix is as follows (Ultsch & Siemon, 1989).

1. Analyze all data with the SOM.
2. Compute the difference between the neurons.

3. Add an element between all adjacent neurons, color neurons and added elements according to distance between neurons.
4. If clusters are geometrically far away from each other or they are separated by high (red) wall, then there is a large dissimilarity between the clusters.

Combination of the U-matrix and component plane representation enables a researcher to obtain both clusters and correlations between the samples from the same figure. For example, the U-matrix and component planes in figure 5 are constructed from relative gene expression levels of 270 genes across four breast cancer cell lines (ZR-75-1, Hs578T, MDA-436, and BT-474).

Circled neurons in the U-matrix (figure 5) comprehend clusters since they are shaded with blue and there are no high walls between them. Circled neurons in component plane Hs578T contain the most highly expressed genes in breast cancer cell line Hs578T. It is evident that highly expressed genes in Hs578T are suppressed in ZR-75-1 and BT-474. Moreover, Hs578T and MDA-436 have similar groups of suppressed genes while these cell lines have very different sets of high expressed genes. Some values in color bars next to component planes are negative due to logarithmic transformation prior the SOM analysis. Mark "d" below the color bar indicates that the values in the color bar are denormalized for visualization purposes.

2.5. Computing significance of clusters

Very often it is necessary to compute significances of the resulting clusters. In this section we suggest a general procedure for assessing significance of a cluster. Our approach is to compare the similarity of the gene expressions in a cluster to similarity of the gene expressions chosen randomly from the data set.

Explicitly stated H_0 hypothesis is "It is possible to choose randomly a group of genes whose expression patterns are more similar to each other than expression patterns between clustered genes". Naturally H_1 hypothesis is the complement of H_0. Pseudo-code for the procedure is as follows.

```
1. Choose test statistic $S(x)$.
2. Compute $S$ using gene expression patterns in a cluster.
3. If $S$ was computed using $c$ genes, randomly choose $c$ gene expression patterns from the
   data set and compute $S_{new}$ using them.
4. If $S_{new} > S$ (or $<$, depending on the choice of $S(x)$) increase counter by one.
5. Perform steps 2–4 $r$ times and finally divide counter by $r$. This value corresponds to
   one-tailed $p$-value of the cluster.
```

In this study we have chosen $S(x)$ to be the sum of the correlation distances between all pairs of the gene expression profiles in a cluster or in a randomly chosen set of genes. Further, we chose r to be 5000.

The procedure for assessing the significance of a cluster allows semi-automatic gene expression clustering and visualization as follows. Firstly, the data are analyzed with the SOM. Secondly, the user identifies a group of neurons that comprehend a cluster using the U-matrix. Thirdly, for the cluster determined by the user, p-value is computed using the procedure described above. In summary, the U-matrix and the component plane representation accompanied with the procedure described in this section speeds up microarray data analysis and produces assurance to the clusters under interest.

3. Analyzing microarray data with SOM

Originally, the SOM was developed for matching static signals only and it may not attain steady state if experiments are statistically dependent. In many cases, the underlying biological process is dynamic, which may lead to statistically dependent samples. For example, if one takes samples of a cell during the cell cycle, the samples are very likely statistically dependent on each other.

In a cDNA microarray experiment, the gene expression levels of millions of cells are determined. Due to the large amount of cells, it is very unlikely that all the cells are in same cell cycle phase unless synchronization is used. Thus in a standard microarray experiment the expression levels in each sample are “mixed” in sense that every hybridized cDNA strand belongs to an arbitrary cell cycle. As a consequence, samples may be considered to be independent on each other.

If the experiment leads to data that are dependent, there are two basic approaches for applying the SOM. The first is to define a time window for a sequence of samples and concatenate them. Another method is to filter the data with low-pass filters. More detailed discussion of dynamical SOM is given in Kohonen (2001).

3.1. Preprocessing

A preprocessing stage is obligatory in order to ensure that gene expression patterns across samples are comparable. Microarray data consist of ratios of gene expression levels between two samples (test and reference, the latter one being the same in all experiments), so a popular start is to take logarithm of the data. Very often it is helpful to do mean/median centering in order to remove the effect of the reference sample. In many cases median is preferred over mean since median is not affected by outliers.

3.2. Choosing the distance measure

One of the most difficult decisions when using a clustering algorithm is to choose the distance measure. The standard choice is the Euclidian distance, mostly because it is simple. However, there are many situations in microarray data analysis where Euclidian distance may not be the best choice. The reason is that, from a biological point of view, the direction of the change in the gene expression profile is very often more important than difference between ratios and Euclidian distance is incapable to take direction into account.

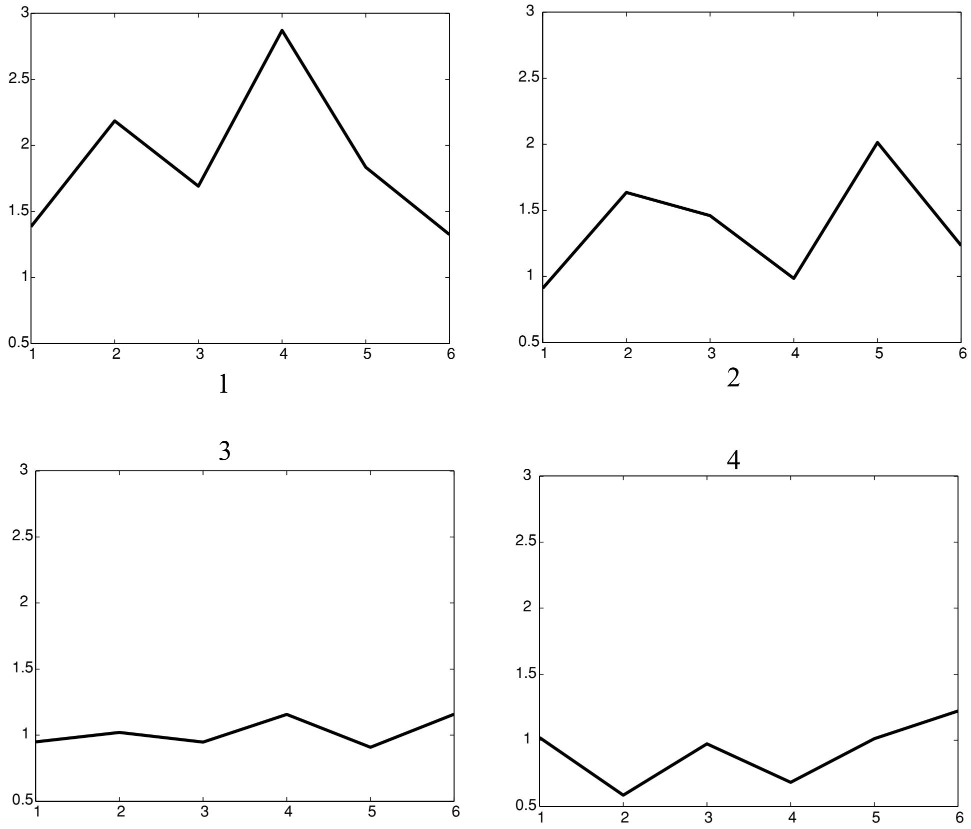

Figure 6. Four gene expression profiles as introduced to the expert committee.

Another distance measure, correlation distance, is capable for taking also directions of the changes in the profiles into account, so it may be better choice for a distance measure than the Euclidian distance. As the distance measure should meet biological requirements, proposed distance measures should be verified by comparing their performance to the opinions of biologists who conduct the microarray experiments.

In this section we compare results computed using four distance measures to opinion of a committee of nine biologists. The experiment was conducted as follows. Firstly, each one of the nine judges was given four gene expression profiles as illustrated in figure 6 and was asked to determine the similarity between the profiles on a scale of zero to ten. Then the evaluations were averaged and sorted. Secondly, we computed the distances between the profiles using the Euclidian distance, the correlation, the symmetric Kullback-Leibler, and the Tanimoto distance measure (definitions given in Appendix). Finally, we ranked the distances between gene expression profiles for all four distance measures. The results are given in Table 1. According to the committee, the most similar profiles are 1 and 3 (Column 1–3 in Table 1).

Table 1. Rank orders of distances between the gene expression profiles given in figure 6.

Distance measure	1–2	1–3	1–4	2–3	2–4	3–4
Euclidian distance	4	5	6	2	3	1
Kullback-Leibler (symm.)	5	1	6	3	4	2
Correlation	5	1	6	3	4	2
Tanimoto	3	5	6	2	4	1
Expert committee	2	1	6	3	4	5

Correlation distance and Kullback-Leibler agreed well with the committee (Kendall's tau resulted in correlation coefficient 0.333 for both). The assumption behind the Kullback-Leibler is that two distributions to be compared are probability distributions. As this assumption does not hold in the context of the microarray experiment, profiles must be converted to probability distributions. The conversion can be done in many ways that produce slightly dissimilar results. Also, conversion slows down the algorithm and sometimes creates false similarities when the number of samples is low. However, if the number of the samples is big and samples to be clustered do not share dimension (e.g. some samples measure gene expression ratios, some protein concentrations or gene ontology classes) Kullback-Leibler may turn out to be a very good choice since distances between the probability distributions are straightforward to compute.

From Table 1 one is able to see that Euclidian distance does not agree very well with the committee (Kendall's tau resulted in correlation coefficient −0.200). Also results obtained by Oja et al. (2002) suggest that Euclidian distance may not be optimal distance when analyzing gene expression data. However, in Gibbons & Roth (2002) Euclidian distance was found to be the best measure for computing dissimilarities of ratio-based measurements, while correlation distance was found to be the best for non-ratio-based measurements. Detailed comparison of the distance measures is beyond the scope of this study, but it would make an interesting topic for further study. As a consequence, we use both Euclidian distance and correlation distance in our case studies.

4. Case studies

In the case studies we have applied the SOM to breast cancer and prostate cancer data sets. For both data sets we used batch learning algorithm and Gaussian neighborhood function. Distance measure for breast cancer data set was correlation, while for prostate cancer data set we used Euclidian distance. All other a priori parameters are set for default values as presented in Vesanto et al. (2000).

In a typical cDNA microarray experiment, the majority of the genes are not over- or underexpressed but more or less constant. Naturally, the SOM assigns all genes having similar expression profile to the same cluster. When there are thousands of similar gene expression profiles, the result is one huge cluster and drawing conclusions is very difficult. For example, initially the breast cancer data set consisted of 13,824 genes. When all the genes were included in the SOM analysis, the result was a cluster consisting of nearly 10,000 genes (data not shown). Trying to find biological meaning in such a big cluster is ineffective. Our approach to overcome this problem is to identify a defined set of genes that show clear expression changes and then apply these genes to the SOM.

4.1. Breast cancer data set

DNA amplification is a known mechanism for increasing expression levels of genes that provide a growth or survival advantage in cancer and the aim of this study was to identify genes whose expression levels were elevated due to amplification (i.e. increased gene copy number). The original data set contained 14 breast cancer cell lines. Each cell line was subjected to two microarray experiments (for details on the microarray protocols and fluorescent image analysis (Hyman et al., 2002)). First experiment was a traditional cDNA microarray experiment revealing the expression levels of the 13,824 genes included in the array. In the second experiment, comparative genomic hybridization (CGH) on cDNA microarray (Kallioniemi et al., 1992; Monni et al., 2001; Pollack et al., 1999) was used to measure the copy numbers of the same set of genes. The data were normalized by taking logarithm and performing median centering for the samples. Using a random permutation test we identified a set of 270 genes that were most likely to have an increased expression level due to increased copy number (Hautaniemi et al., 2002; Hyman et al., 2002).

The expressions of 270 genes across 14 samples were applied to the SOM and the resulting SOM map is illustrated in figure 7. The U-matrix revealed several small clusters (figure 7) indicating that the SOM was able to further categorize these genes based on their expression patterns. Some of these clusters contained genes from a particular region of the genome, i.e. genes that were co-amplified. For example, 11 genes from the 17q23 chromosomal region were included in the SOM analysis and ten of these clustered together to neurons in the lower right-most corner (Cluster 1 in figure 7). We obtained $p < 0.028$ for Cluster 1 using procedure given in Section 2.5

Visual inspection of the SOM map in figure 7 indicates that the genes in Cluster 1 are highly expressed in MCF7, MDA361, and HCC1428 cell lines that are known to harbor 17q23 amplification (Monni et al., 2001). Similarly, most of the genes originating from the so-called *ERBB2* amplicon (four genes) are located in a cluster of three neurons (Cluster 2 in figure 7) with $p < 0.037$. The SOM map illustrates the high expression of Cluster 2 genes in cell lines, such as BT474, ZR7530, and UACC812, containing this amplicon as reported in Kauraniemi et al. (2001). Clustering of 11 of the 25 genes from the 20q13 region, another common amplification site in breast cancer, was also evident in the SOM analysis (Cluster 3 in figure 7). In this cluster there is only one gene outside the 20q13 region and we obtained $p < 0.0226$ for Cluster 3.

The above examples demonstrate that of the 270 genes whose expression were statistically attributable to increased copy number, the SOM clustered together genes that originated from a particular region of the genome and were co-amplified in a set of cell lines using only expression data in clustering analysis. This result indicates that the SOM is able to help identification of biologically meaningful clusters.

We also explored the applicability of Euclidian distance in the SOM analysis of the breast cancer data set. This approach resulted in a U-matrix with a single large cluster containing almost all of the 270 genes (data not shown) thus making the U-matrix useless. However,

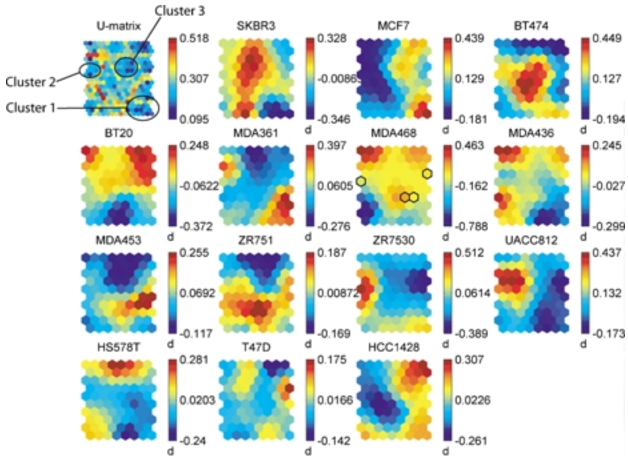

Figure 7. SOM map for breast cancer data set. Three examples of clusters are circled in the U-matrix. Cluster 1 and Cluster 2 contain neurons that contain genes from chromosome region 17q23 and *ERBB2* amplicon, respectively. Cluster 3 contains genes from 20q13. Gridded neurons in component plane MDA468 indicate positions of *BRCA2* predictor genes.

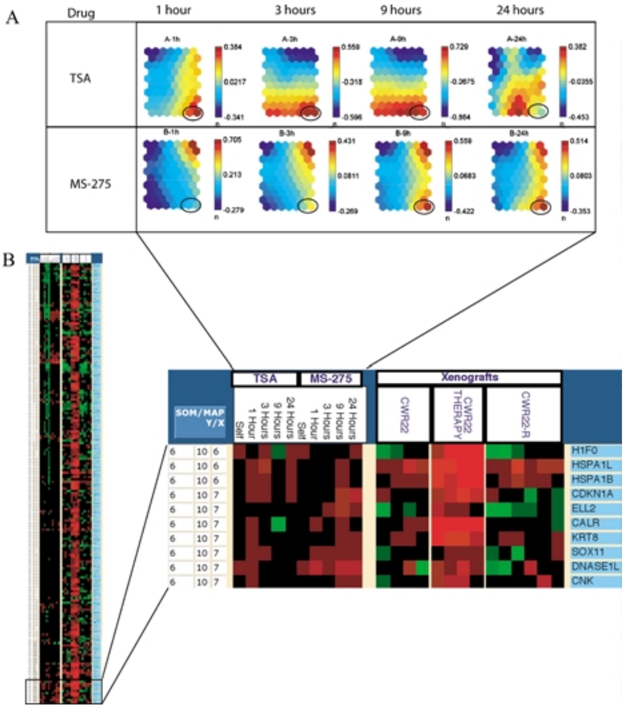

Figure 8. Using the SOM in comparison of prostate cancer drugs TSA and MS-275 that have anticancer effects. (A) Component planes of TSA and MS-275 across all time points. (B) Snapshot from relational database and enlargement of genes clustered to circled neurons.

genes belonging to Clusters 1, 2, and 3 were grouped close to one another also with the use of Euclidian distance.

One of the aims of cDNA microarray based gene expression studies has been improved disease classification. Previous studies have, for example, identified genes that predict the presence of *BRCA2* mutations that are known to be responsible for a substantial portion of inherited breast cancer (Hedenfalk et al., 2001). Gridded neurons in component plane MDA468 (figure 7) indicate the locations of the four top predictor genes identified in Hedenfalk et al. (2001) and Xiong, Fang, & Zhao (2001) that were included also to our breast cancer data set. These genes are (coordinates in brackets): *mitogen-activated protein kinase, kinase 1* (4.8), *protein phosphatase 1, catalytic subunit, alpha isoform* (7.6), *zinc finger protein 161* (7.7), and *DKFZP586L0724 protein* (1.5). The fact that the predictor genes are clustered far from each, and from U-matrix one can see that neuron containing *protein phosphatase 1* is very distinct compared to the neuron containing *zinc finger protein 161* ($p < 0.1602$, so H_0 cannot be rejected at significant level), may explain their predictor value and possibly implies that these genes highlight different pathways that are affected by the *BRCA2* mutation. Consequently, SOM clustering results may be used to aid in the selection of genes for disease classification purposes.

4.2. Prostate cancer data set

Androgen ablation therapy is one of the last lines of defense against prostate cancer. Unfortunately, resistance to this form of hormonal therapy is inevitable leading to death for patients with hormone refractory prostate cancer. It is therefore of great interest to identify the genes involved in the progression of this disease and to identify drugs and specific drug targets for the treatment and management of late stage prostate cancer. We have previously investigated the in vivo expression profile of hormone refractory cancer (Mousses et al., 2001). In addition, we are now investigating emerging anticancer compounds to identify genes whose expression is associated with a response and attempt to identify putative new targets for therapeutic intervention. To this end, we used cDNA microarrays to measure the gene expression profile in prostate cancer cells in vitro following a time course of treatment with several drugs. In each case, the cells were exposed to a chemical compound for either 1, 3, 9, and 24 hours, then the mRNA was isolated and subjected to hybridization on a cDNA microarray with 14,000 genes. Cy3-labeled cDNA from each treatment was hybridized against a Cy5-labeled reference cDNA sample resulting in a relative ratio for fluorescence intensity for each spot on the microarray (for details on the microarray protocols and fluorescent image analysis (Mousses et al., 2001)). The microarray data were then put into a relational database for further mining and analysis.

Since the drug response differed amongst the drugs tested, we analyzed the microarray data to discover the genes whose expression is associated with anticancer activity. We were particularly interested in a subset of progression associated genes, which we had previously shown to be associated with prostate cancer growth and survival during in vivo therapy resistance. Specifically we used the SOM to visualize the expression profile of a subset of 262 genes, which were not affected by hormonal therapy, but were either induced or repressed as tumors acquired resistance to hormonal therapy. As an example, figure 8 shows the component planes for two drugs that have anticancer effects on the CWR22R cell line. While these two histone deacetylase drugs target the same pathway, they also may target other pathways, and this may differ between them. It is obviously important from the drug development point of view, as well as the novel drug target discovery perspective, to identify the genes that are different, as well as those that are common and might mediate the anticancer effects. A group of neurons in the component plane show that there is a set of genes whose expression in induced by the two effective drugs. The specific genes in each neuron are mapped back to a database. For example, there are seven genes, which were clustered to the neuron at the lower right corner in figure 8(a). The expressions of these genes (figure 8(b)) is shown by the color plot taken from relational database (red is upregulated, green is downregulated, black is unaffected), match the pattern seen for the neuron across the samples and the response to the drugs. Furthermore, the pattern of expression for these genes in the drug treatments can be compared to their response in vivo.

Interestingly, the same genes that were induced by the effective drugs were also repressed during tumor progression, suggesting that they may play a role in suppressing the growth and survival of these cells since the anticancer compounds induce their expression. The clustering of these genes together indicates that the SOM is an effective means to analyze microarray data to discover and visualize patterns of expression that match phenotypic endpoints and to identify the groups of genes that fit these patterns.

5. Comparison of SOM visualization

In this section we compare a widely acknowledged way to visualize the SOM clustering results (Tamayo et al., 1999) to the U-matrix and the component plane representation using breast cancer data set.

We applied 270 genes that were preprocessed as explained in Section 4.1 to GeneCluster v.2.1 Beta (Tamayo et al., 1999) and performed the SOM analysis. Parameters were set to defaults, except the size of the SOM map, which was set to be 10 × 8 neurons as in Section 4.1. The resulting SOM map is given in figure 9.

In figure 9 each rectangular box denotes a neuron. Altogether there are 80 neurons and $c0$ is the name for the first neuron, $c1$ name for the second etc. Digit next to a neuron's name is the number of genes clustered to this neuron. Each neuron in figure 9 contains a line with dots that visualize reference vectors. According to Tamayo et al. (1999) two lines above and below the dotted line denote standard deviation of average expression of genes clustered in the neuron. There are two neurons that do not contain genes, for these neurons standard deviation lines are absent. GeneCluster treats each neuron as a cluster (c is abbreviation from cluster), which is a common fallacy. While this inaccuracy in notation does not affect actual clustering, it may mislead a researcher to concentrate only on single neurons instead of groups of neurons that constitute a cluster.

Cluster 1 in figure 7 consist of ten 17q23 genes. In figure 9 seven of these genes are clustered to neurons $c2$ and $c3$ and three to neurons $c22$ and $c23$. Four *ERBB2* amplicon genes were clustered to Cluster 2 (figure 7). In figure 9 these genes are clustered to neurons $c70$, and $c72$. Genes that belong to Cluster 3 in figure 7 are clustered to neurons $c7$, $c8$, $c9$, and $c18$ in figure 9.

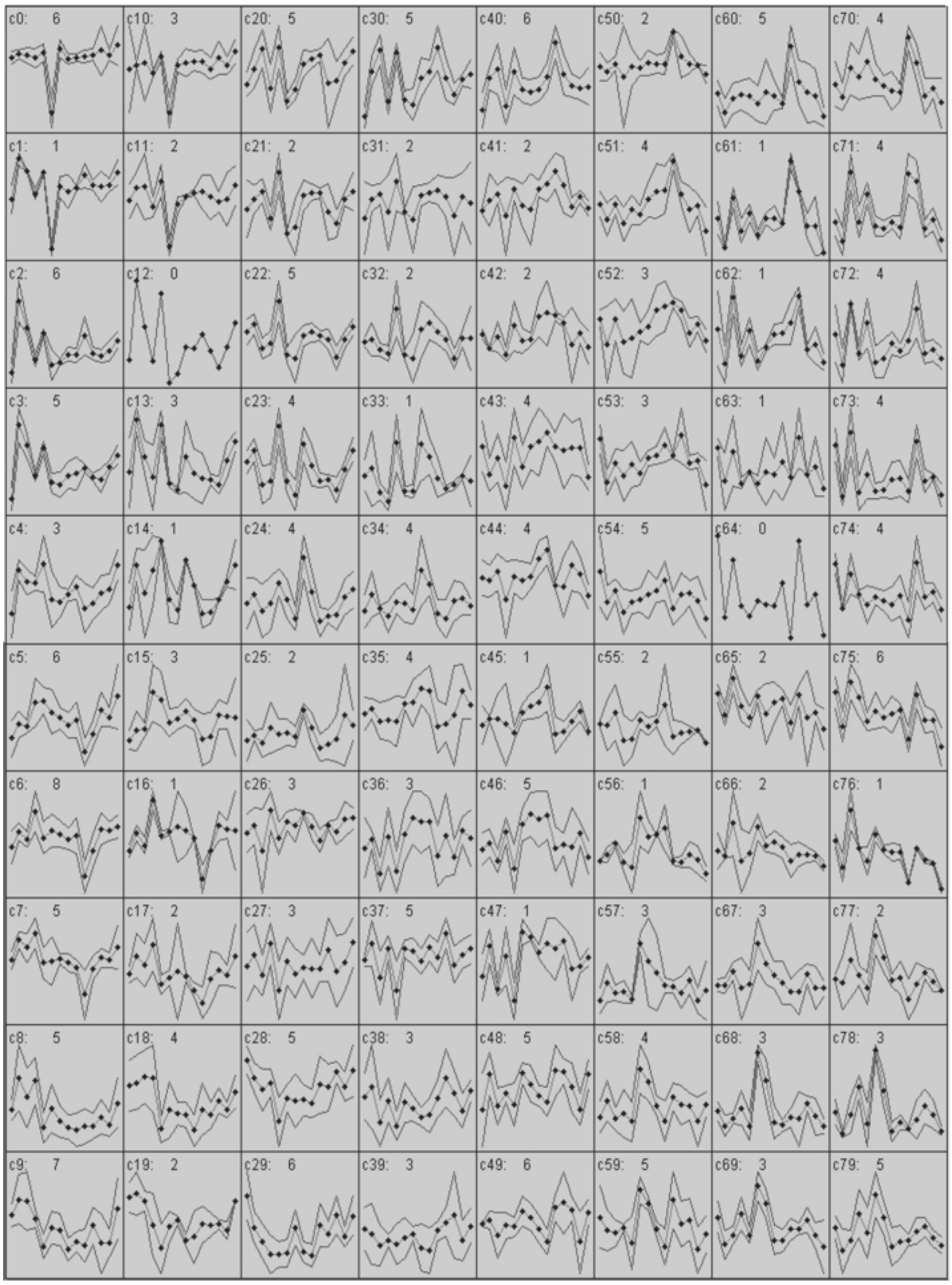

Figure 9. 10 × 8 SOM map from GeneCluster. Dotted lines in neurons denote reference vectors and lines above and below standard deviations of average expression.

Genes that belong to three clusters detected using U-matrix in figure 7 are mainly clustered close to each other also when using GeneCluster. Variation is mainly due to different settings of parameters, for example, in GeneCluster Euclidian distance must be used since there are no other options.

When SOM map in figure 9 is compared to SOM map in figure 7 two things are obvious. Firstly, correlations between the samples are practically impossible to identify from figure 9, while with component planes it is easy to find groups of neurons that consist of genes sharing similar expression profile across the samples. Secondly, identification of groups of neurons that comprehend a cluster using figure 9 is very difficult. One may try to compare the reference vectors between adjacent neurons, but clearly U-matrix offers much more convenient way to detect clusters of several neurons.

In this study we have used MATLAB software with SOM-toolbox (Vesanto et al., 2000), which is freely available at http://www.cis.hut.fi/projects/somtoolbox/. One of the most appealing features of the SOM-toolbox is that the source code is both permitted and easy to modify. This is not possible with Java-based microarray analysis softwares such as GeneCluster, GeneSpring v.5.0, and JExpress v.1.1, which visualize the SOM results in a similar fashion with figure 9. This is a remarkable disadvantage since one may want to cluster gene expression profiles with the correlation distance instead of Euclidian or make other alterations.

6. Discussion

Visualization of microarray experiments is very important but difficult due to high dimensionality of the data. The SOM algorithm offers a good choice for this purpose due to its unique and versatile visualization capabilities. In this study we have utilized the U-matrix and the component plane representation to illustrate the usefulness of the SOM for microarray visualization and analysis tasks.

Our case studies indicate that the SOM is applicable to observational data (breast cancer data set) as well as to intervened data (prostate cancer data set). Further, the case studies imply that, at least in these data sets, Euclidian distance and correlation distance cluster genes in a similar fashion; genes that were clustered next to each other with the correlation distance were clustered close to each other when the Euclidian distance was used. However, if correlation is used as a distance measure, the U-matrix is more useful and clusters can be found more easily. Thus, correlation distance measure may perform better in cases where the purpose is to find gene clusters.

The versatility and robustness of the SOM also brings forward some problems. The SOM contains far more parameters than, for example, hierarchical clustering. Moreover, the SOM is a mathematically complicated algorithm and some of its theoretical properties still remain without proof (Kohonen, 2001). Therefore the SOM is more difficult to apply to the microarray data than the hierarchical clustering. In addition, widely used visualization by showing reference vectors (figure 9) makes the identification of clusters and correlations between the samples difficult. These two issues might explain why the SOM has not been widely accepted as a standard clustering tool for microarray data analysis. It must be also noted that if the data set consists of hundreds of samples the component plane representation may be difficult to interpret.

Choice of the distance measure is problematic for all clustering algorithms. As clustering algorithms are used for analysis of the biological data, it is important that the chosen distance measure meets biological requirements. For example, from a biological point of view, the direction of a change might be more important than the actual difference between the ratios. Thus, new distance measures need to be carefully validated before they are applied to the microarray data analysis. The comparison experiment of the distance measures employed in this study, albeit being too small for making any global conclusion, is one of the first steps towards choosing the distance measure with less heuristic manner.

When the data set contains thousands of genes with similar expression patterns, they are clustered close to each other and the result may be difficult to interpret. Moreover, groups of neurons, having thousands of genes, constitute a formidable task to biologists or medical doctors who are analyzing the relevance of the results. Therefore, the number of genes to be analyzed must be limited. In this study we extracted sets of biologically or pharmacogenomically interesting genes for the SOM analysis. This requires a remarkable amount of biological knowledge and slows down the analysis. One of our future directions is to incorporate information from several sources, such as gene ontology databases, to gene expression data in order to identify groups of interesting genes.

In many recent studies clustering results are hoped to be helpful in construction of genetic regulatory networks. As clustering algorithms compare how similar gene expression profiles are, not how significant a gene is for another gene, one should use clustering results very cautiously when inferring interactions between genes. For example, if two gene expression profiles are mirror images, they are normally clustered far away from each other. However, biologically these kinds of genes may be very important since they are likely involved in the same regulatory network.

Clustering results may be very useful when searching for a set of genes that can be used for construction of a classifier for diagnostic purposes. Genes that are clustered far away from each other may turn out to be good predictor genes as indicated in the breast cancer case study. Therefore, clustering results can reduce the number of predictor genes considerably. As the SOM is an unsupervised algorithm, it cannot be used directly in classification. However, the SOM is closely related to learning vector quantization (LVQ) algorithm, which is a classification algorithm. The SOM is often used as a preprocessor and the LVQ is initialized using results from the SOM (Kohonen, 2001).

In summary, we have discussed theoretical and practical aspects of analysis of gene expression data in human cancers using the SOM. We conclude that the SOM provides an excellent format for visualization and analysis of microarray data, and is likely to facilitate extraction of biologically and medically useful information.

Appendix A: Definitions of the distance measures

Here we present definitions for distance measures we have used in this study. Vectors **x**, **y** $\in \mathbb{R}^{n\times 1}$ and their inner product is defined as:

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i. \tag{7}$$

Especially,

$$\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2. \tag{8}$$

A.1. Correlation distance

Correlation distance is defined as:

$$d_c(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} \tag{9}$$

A.2. Euclidian distance

Euclidian distance is defined as:

$$d_e(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \tag{10}$$

A.3. Tanimoto distance measure

Tanimoto distance has produced good results in finding relevancies between documents and is defined (Kohonen, 2001):

$$d_{ta}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\| + \|\mathbf{y}\| - \mathbf{x}^T\mathbf{y}} \tag{11}$$

A.4. Symmetric Kullback-Leibler distance measure

Kullback-Leibler distance is based on information theory. It is very widely in used in engineering science since it has many important properties (Haykin, 1999). Kullback-Leibler distance for discrete vectors is defined as:

$$d_{kl}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \log\left(\frac{x_i}{y_i}\right), \tag{12}$$

where n is the number of values in the vectors. The Kullback-Leibler is not symmetric and thus not true distance. However, it is possible to modify in order to satisfy symmetry property as follows

$$d_{kl2}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \left[x_i \log\left(\frac{x_i}{y_i}\right) + y_i \log\left(\frac{y_i}{x_i}\right)\right]. \tag{13}$$

d_{kl2} has been used in different applications earlier e.g. (Siegler et al., 1997). Non-symmetric Kullback-Leibler is used earlier with gene expression clustering with the SOM by Kaski and

Sinkkonen (2002). One assumption behind the Kullback-Leibler distance is that vectors **x** and **y** represent probability distributions. Haykin (1999) Therefore gene expression profiles must be converted to probability distributions, which may cause some loss of information.

Acknowledgments

This work was supported in part by the Foundation of Technology, which is gratefully acknowledged.

References

Chen, D.-R., Chang, R.-F., & Huang, Y.-L. (2000). Breast cancer diagnosis using self-organizing maps for sonography. *Ultrasound in Medicine and Biology, 26:3*, 405–411.

Chen, G., Jaradat, S., Banerjee, N., Tanaka, T., Ko, M., & Zhang, M. (2002). Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statistica Sinica, 12*, 241–262.

Eisen, M., Spellman, P., Brown, P., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, USA, 95*, 14863–14868.

Gibbons, F., & Roth, F. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research, 12*, 1574–1581.

Hautaniemi, S., Ringnér, M., Kauraniemi, P., Kallioniemi, A., Edgren, H., Yli-Harja, O., Astola, J., & Kallioniemi, O.-P. (2002). A strategy for identifying putative causes of gene expression variation in human cancer. In *Proceedings of Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Raleigh, NC, USA, Oct. 2002.

Haykin, S. (1999). *Neural Networks, a Comprehensive Foundation,* 2nd edition, Prentice Hall.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.-P., Wilfond, B., Borg, Å., & Trent, J. (2001). Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine, 344:8*, 539–548.

Hill, A., Hunter, C., Tsung, B., Tucker-Kellogg, G., & Brown, E. (2000). Genomic analysis of gene expression in C. elegans. *Science, 290*, 809–812.

Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringnér, M., Sauter, G., Monni, O., Elkahloun, A., Kallioniemi, A., & Kallioniemi, O.-P. (2002). Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Research*, in press.

Kallioniemi, A., Kallioniemi, O.-P., Sudar, D., Rutovitz, D., Gray, J., Waldman, F., & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science, 258*, 818–882.

Kaski, S., Kangas, J., & Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys 1,* 102–350.

Kaski, S., Nikkilä, J., Törönen, P., Castrén, E., & Wong, G. (2001). Analysis and visualization of gene expression data using self-organizing maps. In *Proceedings of NSIP-01, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing.*

Kaski, S., & Sinkkonen, J. (2002). Clustering based on conditional distributions in an auxiliary space. *Neural Computation, 14*, 217–239.

Kauraniemi, P., Bärlund, M., Monni, O., & Kallioniemi, A. (2001). New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. *Cancer Research, 61*, 8235–8240.

Kohonen, T. (2001). *Self-Organizing Maps,* 3rd edn., Springer.

Mangiameli, P., Chen, S., & West, D. (1996). A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research, 93*, 402–417.

Monni, O., Bärlund, M., Mousses, S., Kononen, J., Sauter, G., Heiskanen, M., Paavola, P., Avela, K., Chen, Y., Bittner, M., & Kallioniemi, A. (2001). Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. *Proceedings of the National Academy of Sciences, USA, 98:10*, 5711–5716.

Mousses, S., Wagner, U., Chen, Y., Kim, J., Bubendorf, L., Bittner, M., Pretlow, T., Elkahloun, A., Trepel, J., & Kallioniemi, O.-P. (2001). Failure of hormone therapy in prostate cancer involves systematic restoration of androgen responsive genes and activation of rapamycin sensitive signaling. *Oncogene, 20:46*, 6718–6723.

Oja, M., Nikkilä, J., Törönen, P., Wong, G., Castrén, E., & Kaski, S. (2002). Exploratory clustering of gene expression profiles of mutated yeast strains. In W. Zhang & I. Shmulevich (Eds.), *Computational and Statistical Approaches to Genomics*, Kluwer Academic Publishers.

Parmigiani, G., Garrett, E., Anbazhagan, R., & Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 1–20.

Pollack, J., Perou, C., Alizadeh, A., Eisen, M., Pergamenschikov, A., Williams, C., Jeffrey, S., Botstein, D., & Brown, P. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics, 23:1*, 41–46.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E., & Golub, T. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences, USA, 98:26*, 15149–15154.

Raychaudhuri, S., Stuart, J., & Altman, R. (2000). Principal components analysis to summarize microarray experiments: Application to sporulation time series. *Proceedings of the Pacific symposium on Bioinformatics, 5*, 452–463.

Siegler, M., Jain, U., Raj, B., & Stern, R. (1997). Automatic segmentation, classification and clustering of broadcast news audio. In *Proceedings of the DARPA Speech Recognition Workshop* (97–99).

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., & Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps; methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences, USA, 96*, 2907–2912.

Törönen, P., Kolehmainen, M., Wong, G., & Castrén, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Letters, 451:2*, 142–146.

Ultsch, A., & Siemon, H. (1989). Exploratory data analysis: Using Kohonen networks on transputers. *Technical Report 329,* University of Dortmund, Germany.

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). SOM toolbox for Matlab 5. Technical Report A57, Helsinki University of Technology, Finland.

Wall, M., Dyck, P., & Brettin, T. (2001). SVDMAN-singular value decomposition analysis of microarray data. *Bioinformatics, 17:6*, 566–568.

Xiong, M., Fang, X., & Zhao, J. (2001). Biomarker identification by feature wrappers. *Genome Research, 11:11*, 1878–1887.

Received June 18, 2002
Revised November 1, 2002
Accepted December 27, 2002
Final manuscript January 13, 2003