Self-organizing maps in mining gene expression data

Kari Torkkola *, Robert Mike Gardner, Tamma Kaysser-Kranich, Calvin Ma

Motorola Labs, MD EL508, 2100 East Elliot Road, Tempe AZ 85284, USA

Received 10 September 1999; received in revised form 24 February 2001; accepted 1 April 2001

Abstract

Modern DNA microarray technology provides means of measuring gene expression patterns of the whole genome of simple organisms at once. Exploratory analysis of these large-scale expression datasets is becoming vital to extracting functional information from the measurements. We demonstrate how self-organizing maps (SOM) can be applied to exploratory analysis of gene expression data from a yeast DNA microarray database in order to very rapidly find gene families with similar expression patterns. SOM not only enabled quickly selecting the gene families identified in previous work, but it facilitated identifying additional genes with similar expression patterns. Identifying new families of genes also appears to be possible as demonstrated by additional clusters of genes discovered from the data. Moreover, further insight into the primary pattern variations that discriminate between the families became explicit. © 2001 Elsevier Science Inc. All rights reserved.

1. Introduction

One of the ultimate goals in biological research today is to determine the proteins involved in specific physiological pathways. This is important for several reasons. First, basic research stems from a desire to understand the

* Corresponding author. Present address: Motorola Labs, MD ML28, 7700 South River Parkway, Tempe, AZ 85284, USA.
E-mail address: kari.torkkola@motorola.com (K. Torkkola).

0020-0255/01/$ - see front matter © 2001 Elsevier Science Inc. All rights reserved.
PII: S0020-0255(01)00158-X
reactions of proteins to environmental stimuli, the interactions of biological entities (proteins or DNA) with each other, and the influence of activators and repressors on other members of the pathways; that is, the control mechanisms that exist to keep all biological players in check. Beyond a basic understanding, the knowledge of the proteins involved in a particular pathway within the body can give us a clue as to the altered protein in a disease state affecting that pathway. Lastly, the hope is that this ultimately will lead to the identification of a particular protein that might be targeted by a new drug to help alleviate the disease state.

Monitoring of the expression levels of messenger RNA (mRNA), which encodes for the corresponding proteins, is one way in which the discovery of pathways and families of similarly acting proteins is facilitated. The state of a cell and its fundamental processes is reflected in the levels of mRNA. Subjecting a cell to environmental stimuli and measuring the mRNA levels of genes of interest over time provides expression patterns for the genes. Genes with similar expression are likely to be involved in similar mechanisms.

DNA microarrays have been used to measure the expression levels of many genes simultaneously. A microarray is an ordered array of DNAs printed on a small glass slide by a robotic device [1]. The amount of mRNA can be measured by corresponding cDNA, which is fluorescently labeled. The cDNA then hybridizes (binds to) the DNA on the microarray, and the resulting amount of fluorescence at each position on the array indicates the level of mRNA in the cell for each DNA sequence (gene) printed on the array.

Although in this work we use data generated by a 6400 element microarray, which should be considered ‘tiny’ by today’s data mining standards, it is expected that 1000 × 1000 arrays will be possible. The volume of data generated by these microarrays over many time points make any kind of manual analysis impossible. Here we show that the expression data is amenable for exploratory data analysis techniques, and that it is especially suitable for self-organizing maps (SOM) since these combine the two techniques required by this data: clustering and visualization. The purpose of this paper is to point out the applicability of SOM to gene expression data. Thus we have not attempted to draw any biologically valid conclusions, but we evaluate our work by comparing it to previous work on the same database.

The structure of this paper is as follows. We describe the baseline work of DeRisi et al. [1] and their analysis methods. Clustering and visualization methods are reviewed, especially as used in previous work about gene expression analysis. We describe the SOM-based analysis method and present results with gene expression data. We conclude with a discussion of relative merits of the SOM in this application.
2. Previous work

In previous work, experiments were carried out on yeast, a single-celled eukaryotic organism [1]. The entire sequence of the yeast genome is known, so a complete genome study can be done. To focus on particular pathways, the yeast is subjected to a shift in its environment. A culture has been grown in a high glucose medium, and the yeast has been allowed to deplete the glucose through utilization for energy. As the glucose is depleted, certain genes required for glucose utilization will be turned off (repressed), since they are no longer needed. As the depletion of the glucose progresses, the yeast will begin to utilize ethanol as a carbon source instead, and will grow aerobically. This shift (termed the diauxic shift) leads to the induction (turning on or increase in expression) of other genes.

Although the entire genome of this yeast is known it is not completely characterized. The goal of the experiment was to identify groups of genes that react to the stimuli in a similar manner over the time points taken, thus identifying genes likely to be in a single pathway, or closely related pathways. Uncharacterized genes grouped together with known genes are now likely to be involved in similar functions. These gene families can then be compared to each other to obtain possible relational information and to predict interactions and time-sequence events in the process of induction/repression.

The challenge of these expression experiments is to be able to analyze all of the data generated, which over the seven time points analyzed for all of the approximately 6400 array elements, leads to about 43,000 expression-ratio measurements. DeRisi et al. analyzed interrelationships present in this large data set by hand. They utilized known genes and pathways to verify that the microarrays have generated correct and useful data. In addition, they searched for unknown genes with a matching profile on induction/repression graphs, thus delineating groups of genes which behave similarly. Five of these groups are shown in Figs. 5(b,c) in [1]. The results are presented as graphs of the ratio of induction or repression to the level of expression in the beginning of the experiment measured at seven time points. These five groups were chosen either by looking at induction/repression at a specific time point (such as, “Plot all expression patterns that exhibit more than 9-fold induction at last time point”), or by taking a group of known genes, and searching the database for other genes that exhibit similar expression patterns.

The results generated through this analysis strategy show that the array methodology leads to legitimate expression determination, and is useful in looking for unknown members of behavioral families. However, the analysis

---

1 The figures can be viewed at http://cmgm.stanford.edu/pbrown/explore/metaprofiles.html.
method is tedious and labor-intensive, and the large data set is difficult to analyze in its entirety for associations at all time points.

As emphasized at the end of the paper by DeRisi et al.: “The greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.”

3. Exploratory data analysis

A wealth of such methods does exist – they are the subject of so-called exploratory data analysis, which is a multidisciplinary field between statistics, machine learning, and computer science – to discover and to illustrate structures and similarities in a large sets of data. Two subsets of methods in this discipline are relevant to the task at hand, clustering, and data visualization.

3.1. Clustering methods

The goal of clustering is to categorize or group similar data items together. Two major clustering categories exist: partitional and hierarchical clustering [5].

Hierarchical clustering operates by merging small clusters into larger ones (starting from the actual data items as clusters), or by splitting large clusters (starting from all of the data viewed as a single cluster). A tree of clusters called a dendogram is produced, displaying the split/merge relationships between the clusters. By terminating the dendogram at a desired level, a clustering of the data items into separate groups is obtained. Examples of using this method in the analysis of gene expression data have been presented in [2,13,19,20]. Dendogram is essentially a one-dimensional ordering of the original data items where each item has two closest neighbors. This is also a limitation of the method since multiple kinds of relationships cannot be represented in a single dimension. Sometimes items that are not very dissimilar end up very far in the dendogram. The method is also sensitive to the data: addition or removal of a few data items may result in a drastically different hierarchy. Tibshirani et al. [16] investigated two-way clustering but concluded that it is limited in its ability to discover finer structure.

In partitional clustering, the data is directly divided into a set of clusters without any regard to the relationships between the clusters. These methods try to maximize some measure of similarity within the samples of each cluster, while minimizing the similarity between clusters. Perhaps the best known partitional clustering method is \( K \)-means [12]. These methods are not completely automatic: usually \( K \), the number of clusters needs to be prespecified, and quite different results may be produced with different values of \( K \). Each
method also has a bias of its own, preferring clusters of certain shapes, for example. Although clustering reduces the amount of data, the clusters/categories have limited value as such in exploratory data analysis. The clusters, which are still high-dimensional vectors, should be visualized somehow to help understanding how they are like and what kind of relationships exist between them.

3.2. Data visualization by projection

Methods also exist to reduce the dimensionality of the data items. These methods project the data items to a low-dimensional space at the same time preserving their mutual relations as well as possible. Projection to a plane is often used to visualize the data. Some methods in this category are principal components analysis [4], projection pursuit [3], Sammon mapping [14], and multidimensional scaling [11]. Each method has a different flavor of how the relations will be preserved and what kind of relations will be preserved in the projection. By visual inspection of the projected data the user can hopefully see ‘clusters’, points of data that are similar, and different from others, as indicated by their placement on the plane.

Combination of partitional clustering and projection methods would perhaps be desirable in gene expression analysis. Clustering would provide the grouping of similar expression patterns, and projection would enable studying their interrelationships. This combination has, however, not been used so far in gene expression analysis.

3.3. Self-organizing maps in exploratory data analysis

SOM is unique in the sense that it combines both aspects. It can be used at the same time both to reduce the amount of data by clustering, and to construct a nonlinear projection of the data onto a low-dimensional display [9]. In brief, the SOM is a grid of ‘units’ or ‘nodes’ each similar to the data vectors under analysis. These units become cluster centers; thus each represents a number of original data items. In addition, and in contrast to other clustering methods, the units become organized in such a way that nearby units on the grid are similar to another. Such a smoothly ordered continuous display of the data items facilitates understanding of the structures in the data set [6]. The topology of the grid can be almost anything, but in practice rectangular two-dimensional grids are preferred since they are easy to display.

On a two-dimensional grid topology, each unit has now 4–8 neighbors. This is a 2–4 fold increase to the two neighbors in the dendogram display. There are now 2–4 times more degrees of freedom in organizing the data onto the grid. In fact, a dendogram-like display as in [2] can be produced with a SOM by using a
one-dimensional grid topology, i.e., a string of units. However, in this work we are concerned with rectangular two-dimensional SOM grids.

The number of data items mapped to each unit (acting as a cluster center) is called the number of ‘hits’ for the particular unit. If the number of units on the grid approaches the number of available data items, the result of the self-organization is more akin to two-dimensional sorting than clustering. In this case many units will receive no hits and they represent the borders between more densely populated areas.

The display is especially amenable to interactive work. A user can naturally group the units into ‘superclusters’ by viewing one vector component of the units at a time displayed on the grid with a color or a gray level encoding. To better delineate these superclusters, distances between neighboring units can be overlaid on the display. This so-called U-matrix method results in “walls” that separate contiguous areas (superclusters) from each other [17,18]. Combining this further with a graphical display of hits, results in a clear display of the borders between superclusters. As described in [17], it is essential to use a high-resolution SOM to this end.

SOMs have been used with gene expression data earlier by Tamayo et al. [15]. However, they used only SOMs of very small size, between 3 × 4 and 5 × 6 units. Only the clustering property of the SOMs was made use of in the analysis. The display and its potential for interactive analysis were not utilized. In the following section we describe how they can be used in gene expression data analysis.

4. Analysis with self-organizing maps

4.1. Data acquisition and formatting

The Stanford yeast gene expression data set used by DeRisi et al. is available in the Internet. Data is represented as ratios of red fluorescence to green fluorescence, as R/G ratios for about 6400 genes over seven different points in time. Since there are genes undergoing induction and repression, the value of the R/G ratio can span from 0 to ∞. However, the numerical range between induction and repression is unbalanced. The induction ratio covers the range from 1 to ∞, and the repression ratio spans only from 0 to 1. This is not ideal for the SOM that uses Euclidean distance to measure the similarities between data items, if induction and repression are desired to have similar weights in the analysis. We chose to represent repression in the negative range by converting all values between 0 and 1 to the negative inverse of the value (or just take the

---

2 http://cmgm.stanford.edu/pbrown/explore/index.html
negative of the G/R ratio). Hence, the ‘new’ R/G ratios for the repression process would cover the range from −1 to −∞. Furthermore, the gap between −1 and +1 was closed by adding/subtracting one resulting in fluorescence fraction (ff) as indicated below.

Fig. 1. Relative fluorescence fraction level at time 2 (left) and at time 7 (right) on the same organized SOM. There are only two gray levels associated with repression: White denotes relative fluorescence fraction values less than −4, and the lightest gray denotes the values between 0 and 4. Darker shades of gray are associated with induction.

Fig. 2. Line plots of the data for an enlarged region of the map containing a cluster (upper right-hand corner of Fig. 1). As illustrated, all of the neighboring cells contain records that have similar data trends. Each little graph contains the relative fluorescence fraction at seven different time points.
Fig. 3. Relative fluorescence fraction level at time 7 (left), and cells that exceed the thresholds, and are thus grouped into superclusters (right).

Fig. 4. Repression Group R_1 (defined in the right side of Fig. 3). This should be compared to Fig. 5(E) in [1].
\[ ff = RGratio - 1 \quad \text{if} \ (RGratio > 1) \quad \text{[Induction]} \]
\[ ff = -1/RGratio + 1 \quad \text{if} \ (0 < RGratio < 1) \quad \text{[Repression]} \]

Since the objective of the analysis is to group the genes by similarity found in the data over time, the rate of ratio change could also be added as input to provide the second-order effect (slope) of the induction or repression process. Hence six additional columns were created by finding the difference of the fractions between the seven time readings.

To give equal importance to the fluorescence fraction and the rate change, data at each time period (T1–T7) was normalized to unit variance as a pre-processing step.

4.2. Creating and displaying the SOM

For this analysis, an 80 × 80 SOM (6400 cells) was used. Note that the size in this case was chosen to match the size of the small database to achieve a fine

![Repression Group R_2](image)

Fig. 5. Repression Group R_2 (defined in the right side of Fig. 3). This should be compared to Fig. 5(F) in [1].
resolution in interactive analysis. After training [9], the grid can be visualized in several ways. As each cell in the grid is a vector (with 13 components), we can select a particular component of interest and color code the cells according to the value of that component. Fig. 1 shows the SOM for the relative fluorescence fraction level at time T7 (right) with gray level coding. Darker shades represent the records in the highest fluorescence fraction level range at T7 (indicating highly induced genes), and light shades represent genes in the lowest range at T7 (indicating highly repressed genes).

Note that there are several groups of both highly induced and repressed genes. This indicates that there are differences in the expression patterns that drove them to cluster differently although they share the common trait of either high or low fluorescence fraction level at T7.

Viewing the fluorescence fraction level at different time periods will help us understand how the other data might have affected how the gene records were organized on the map. The left side of Fig. 1 displays the same information at time T2. Hence, the different views of the map (from looking at different

![Induction Group I_1](image)

Fig. 6. Induction Group I_1 (defined in the right side of Fig. 3). This should be compared to Fig. 5(C) in [1].
variables) will reveal mechanisms and differences in the data, and will help indicate how the data might have affected the way the genes were organized on the map.

Another example of visualizing a trained SOM is presented in Fig. 2. Zooming into the area of interest, one can inspect the data in the form of a line plot. This figure depicts the expression curves of the upper right 1/100th of the map. These curves are the mean values of the hits of each map unit.

4.3. Identifying gene clusters by thresholding

Additional relationship information can be generated from the SOM by observing which gene records were assigned to each cell, and if there are other gene records that are placed in the same cell or nearby cells because of their similarity. This is useful for identifying records with similar trends.

Since a part of the goal of this analysis is to evaluate the genes with high induction or repression at the end of the experiment at T7 [1], simple

![Induction Group 1_2](image)

Fig. 7. Induction Group 1_2 (defined in the right side of Fig. 3). This should be compared to Fig. 5(D) in [1].
thresholding was used to locate those records; that is, we set a minimum or maximum threshold for fluorescence at a particular time point. Viewing the SOM at T7, records with fluorescence fractions greater than 4.0 and less than −4.0 (meaning induction or repression of 5-fold or more) are shown at T7 in Fig. 3. Regions of similar records were clearly delineated – hence they were selected and grouped together as superclusters R1, R2, I1, I2, and I3.

The fluorescence fractions of the records contained in these clusters were then plotted. Figs. 4–8 show the resulting plots. These figures should be compared to plots 5B–5F in [1]. We can see that the five clusters that stood out in the T7 map correspond exactly to the five groups found by DeRisi et al. SOM was able to group 91.3% (32/35) of the genes listed in the plots of DeRisi et al. into appropriately similar groups, that is, three of the genes in DeRisi’s plots were missed. Moreover, it is notable that the SOM provided a list of additional genes not found by DeRisi et al. for each of the clusters. The two repression groups were found with 100% accuracy.

**Induction Group I_3**

![Induction Group I_3](image)

Fig. 8. Induction Group I_3 (defined in the right side of Fig. 3). This should be compared to Fig. 5(B) in [1].
5. Results

5.1. Fast formation of gene family clusters

The use of SOM for analysis of array data, as presented here, yields many benefits. It is labor intensive and time-consuming to plot approximately 6400 repression–induction graphs to be able to compare all of the resulting patterns to each other to find similarities and differences. SOM very rapidly discerns similarly expressed genes and clusters them, with the ability of the user to view the clusters by gene name (or other input data), or by line graph. These options quickly give the user information on the type of pattern recognized by the SOM algorithm. Thus, a particular researcher could scan the clusters by an expression pattern of interest, such as the group of genes expressed early and remaining high until glucose depletion, or those expressed very late in the oxidative phase. This rapidity of analysis and identification allows the researcher to spend less time in analysis and more time generating experimental data. This is especially important as many array experiments are initial screening or association analysis tools meant to lead to further investigation, such as in drug discovery.

5.2. Identifying additional genes for each family

In our analysis of the data of DeRisi et al. [1], we discovered five clusters of genes in the T7 threshold map which correspond to the five clusters published

![Fig. 9. Repression cluster at T1 (left) and repression clusters at T2 (right). Same gray scale is used as in Figs. 1 and 3 although higher contrast would make the desired cluster stand out better.](image-url)
earlier in [1]. More members of these families were found, with many of the members having no gene name or known function. For example, SOM discovered 71 genes belonging to cluster R2 in Fig. 4, including the seven listed in the paper. This kind of information will help lead to results on the expression pattern of previously unknown genes, and give researchers a clue as to their function and potential partner interactions. There are also some cells showing high repression or induction which do not become a part of clusters. These may be analyzed as well, and may turn out to be specific regulators involved in the precipitation of a cascade of events, or other interesting genes.

5.3. Identifying additional clusters of genes and temporal relationships

Lastly, through viewing cluster maps at each time point, not only at T7, and by using thresholds to guide apparent clustering, it will be possible to identify many

Repression Group for Time Period 1

Fig. 10. Expression patterns of genes falling into repression cluster at T1 (left side of Fig. 9).
additional clusters of genes which, perhaps, have peak induction or repression at very early time points and then remain static throughout the remainder of the experiment, or which peak for only a very short time period. Two examples are presented in Figs. 9–12. Fig. 9 depicts repression clusters at T1 and T2. There appears to be one major repression cluster at T1, and the corresponding line plot of expression patterns is presented in Fig. 10. The two clusters at T2 are depicted in Figs. 11, 12. All these genes within a cluster have very similar expression patterns and are probably involved in similar mechanisms.

By viewing the seven cluster maps sequentially, one can see the clusters appear and disappear over the range of the experiment. The ultimate goal of the array data is to discover new and novel genes and to define pathways within the cell. By observing the temporal induction–repression patterns of the clusters relative to each other, possible causal relationships can start to be defined, i.e., inducer/induced family identities, which will lead us to a greater understanding of the mechanisms of control in operation in yeast, and ultimately in humans as well.

![Repression Group 1 for Time Period 2](image)

Fig. 11. Expression patterns of genes falling into a repression cluster at T2 (right side of Fig. 9, the cluster closer to the middle of the map).
6. Discussion

We have demonstrated how SOMs can be applied to exploratory analysis of gene expression data. This method provides simple and efficient means to quickly find answers to exploratory questions like “Which uncharacterized genes behave similarly as the known gene xxx?”, or “Which groups of genes express high levels of induction/repression at time point xxx?”. Compared to previous work of DeRisi et al. [1] our approach seems to be capable of finding the same clusters but with additional members. In contrast to the work of Tamayo et al. [15] who also used SOMs to cluster expression patterns, we made use of the high-resolution SOM display and its aptness to interactive exploratory analysis [18].

How does the approach scale to larger databases, for example, to a database generated from a $1000 \times 1000$ microarray? Obviously the ‘2-d sorting’ approach we used with as many SOM units as there are data will not be quite practical. However, with advanced SOM training methods for speedup [10], or
with hierarchically proceeding training algorithms as the Tree-SOM [7,8], very large maps can be created if finer detail is desired from the analysis. By always using a standard initialization procedure, for example, initializing the two axes of the map with two principal components of the data instead of a random initialization, deterministic maps can be produced.

As another strength of the method we can mention that missing values in the data do not need to be manufactured. The SOM handles these naturally. The subjectivity in the process of picking the clusters from the display can be considered as a weakness by those who prefer rigid, deterministic analysis, and as a strength by those who are fond of interactive exploratory analysis.

References