CLUSTERING GENE EXPRESSION DATA USING SELF-ORGANIZING MAPS

Dr R.M. Suresh, K. Dinakaran, P.Valarmathie

Abstract

An analysis of clustering gene expression data is to identify groups of co-expressed genes recognized coherent expression patterns. In this paper, we propose a new approach named Self Organizing Maps (SOM) to organize gene expression data into clusters. This method is superior to the standard unsupervised approach of grouping genes based on gene expression data. One of the main goals of clustering gene expression data is to discover the function of unknown genes. The output is displayed graphically in order to understand the biological processes.

Keywords : Self Organizing Map, Clustering, Patterns, Gene Expression, Unsupervised.

1. Introduction

Functional genomics involves the analysis of large datasets of information derived from various biological experiments. It involves monitoring the expression levels of thousands of genes simultaneously under a particular condition called gene expression analysis. Although a broad variety of approaches are available for gene expression analysis, a microarray technology has become one of the indispensable tools which may be used to measure gene expression in many ways[8][9].

A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called spots. It contains thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene [9]. While the initial intent was to profile the expression patterns of individual genes with microarrays, the ability to cluster these patterns on a genomic wide scale has expanded the utility of microarrays to infer the function of specific genes [6]. In genomic and proteomic studies, there is a need to analyze large amount of data in order to extract information from it. A dominant technique called cluster analysis aims to identify a large set of genes that show a similar regulation to different types of drastic environmental changes. Such analysis may be useful in medical diagnostics in order to predict gene function. Clustering methods have been used on microarray data to distinguish tumor types and to predict clinical outcomes. The most widely used clustering methods are hierarchical clustering and partition based clustering [1] [7].

2. Proposed method

Clustering methods used in analyzing gene expression data. Typically, some standard clustering methods such as hierarchical and K-means for grouping genes based on gene expression data. The SOM is one of the best known unsupervised neural learning algorithms. The SOM layer is a low dimensional array of neurons [5]. It is defined as a discrete grid of map units and each map unit can represent certain kinds of data. Here, the map units represent genes expressed in a chosen set of conditions. The SOM finds prototype vectors to two dimensional spaces [3]. The mapping is defined by associating an n-dimensional model vector m_I with each map unit I, and by mapping each expression profile to the map unit having the closest model vector. Each map unit is directly associated with a weight vector [4]. The weights of the winning unit and its neighboring units are updated. At iteration of t in the training process, a distance $d(X_i, W_j)$ is defined and used for the measure of similarity between X_i and W_j[10]. SOM have the advantage that it is possible to easily display the output as a two dimensional grid of samples.

3. Similarity Measures

For similarity measure, here we are using Euclidean distance metric as given below.

$$d = \sqrt{\sum(Xi - Wj)2}$$

The neuron closest to the input vector is selected as a winning neuron with the corresponding weight vector W_k, where k is the index of the winning neuron as given in the following equation.

$\| X_i(t) - W_k(t) \| = \min\{ \| X_i(t) - W_j(t) \| \}$

When the winning neuron is selected, the weights of the neurons at iteration of (t+1) with two parameters of $N_c(i,t)$ and α are updated by using the following equation.

$W_j(t+1) = W_j(t) + \alpha(t) N_c(i,t) [X_i(t) - W_j(t)]$ for $I \leq N_c(i,t)$.

$W_j(t+1) = W_j(t)$ for all other indices of i.

Where $N_c(i,t)$ is a parameter which indicates neuron as neighborhood around the winning neuron. Here, the neurons that belong to $N_c(i,t)$ are updated while other neurons are not updated as in the above two equations. The parameter $\alpha(t)$ is the learning rate, which is a monotonically decreasing function of t. The range of applications includes pattern recognition, monitoring and data mining and image processing [2].

4. Experiments and results

We applied the method of SOM to the dataset (Fibroblast) selected from an experiment studying the response of human fibroblasts to serum. The results are obtained by manipulating the microarray gene expression data using a clustering tool. Clustering tool provides a computational and graphical environment for analyzing data from microarray experiments. In order to obtain the perfect clustering, the large amount of data from microarray experiment is normalized by adjusting and filtering noisy data. The presence of large contiguous patches of color in the fig.1 represents the group of genes that share similar expression patterns over multiple conditions. When large groups of clustered genes for example, we observed the strong tendency for these genes to share common characters in cellular processes. Clustering similar genes helps biologists to find the function of unknown genes.

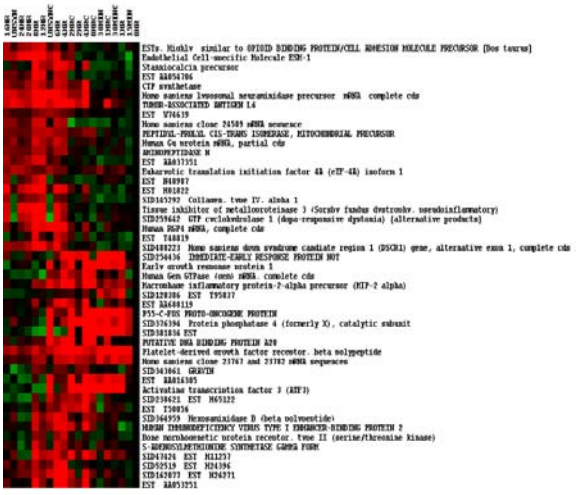

Figure 1.The SOM of fibroblast serum. The red color shows large content of a particular class of genes.

5. Conclusion

As the growth of gene expression data have been increased, there is a need to use new computational tools that help to organize and analyze these data are critical. The SOM approach presented here makes the analysis as easy and also achieved better performance. In this paper, the problem related to data dimensionality was solved and is therefore well suited for DNA microarray data analysis. In future work, we go for finding clear clustering boundaries from the results of SOM as an easy manner.

6. References

[1] Janne Nikkila., Petri Toronen., Samuel Kaski., Jarkko Venna., Eero Castren., Garry Wong: Analysis and visualization of gene expression data using Self-Organizing maps, Elsevier Science Ltd.

[2] Lalinka de Campos Teixeira Gomes., Fernando J. Von Zuben., Pablo Moscato: A proposal for direct-ordering gene expression data by self-organizing maps, Elsevier B.V.

[3] Akinobu Sugiyama., Manabu Kotani: Analysis of gene expression data by using Self-Organizing Maps and K-means Clustering, IEEE.

[4] Dr.S.N.Sivanandam and Dr.M.Paulraj: Inroduction to artificial neural networks, Vikas publications. pp 93-95.

[5] Parmigiani G., Garrett E.S., Irizzary R.A., Zeger S.L, The Analysis of Gene Expression Data Technometrics, Volume 45.

[6] Eisen. M.B., Spellman P.T., Brown P.O., and Bostein D: Cluster analysis and display of genome-wide exoression patterns,Proc. Natl. Acad.Sci.USA,Vol.95.

[7] Yong Wang., Chengyong Yang., Kalai Mathee and Giri Narasimhan: Clustering using adaptive Self-organizing Maps and Applications, Miami FL 33199, USA.

[8] Hans Peter saluz., Javeed Iqubal., Gino V Limmon., Andre Ruryk and Zhihao Wu:
Fundamentals of DNA-Chip/array technology for comparative gene expression analysis. Current science, vol. 83, no. 7, 10 October 2002 833

[9] M.Madan Babu: An introduction to Microarray Data analysis, MRC Laboratory of molecular biology, Hills road, Cambridge, United Kingdom.

[10] Jihua Huang., Hiroshi Shimizu., Suteaki Shioya: Clustering Gene Expression Pattern and Extracting Relationship in Gene Network Based on Artificial Neural Networks, Journal of Bioscience and Bioengineering Vol.96.

[11] Fang-Xiang Wu., W.J. Zhang., and Anthony J. Kusalik: A Genetic K-means Clustering Algorithm Applied to Gene Expression Data, Springer-verlag Berlin Heidelberg.