1 Introduction

A method to evaluate and analyze the massive data generated by series of microarray experiments is of utmost importance to reveal the hidden patterns of gene expression. To extract useful information from expression profiles, computational tools, for example, hierarchical clustering and self-organizing map (SOM) clustering, have been widely used to cluster and display data. Compared with hierarchical clustering, SOM has a number of features that make them particularly well suited to clustering and analysis of gene expression patterns [1]. Though SOM has good computational properties and are easy to implement, reasonably fast, and scalable to large data sets, there is a shortcoming to a conventional plane SOM (2D SOM) whose neighborhood relationship is defined by a 2D rectangular or hexagonal lattice. The grid units at the boundary of the SOM have fewer neighbors than the units inside the map, which often leads to the notorious “border effect” — the weight vectors of these units “collapse to the center of the input space” [2]. To solve the problem, Ritter [3] suggested that the spherical SOM would be very suitable for data with underlying directional structures.

In this study, a spherical SOM (sSOM) software developed by SOM Japan, Co. Ltd. was applied to analyze the data from cell surface marker DNA microarray and visualize the expression patterns. By creating a spherical Self-Organising-Map, the boundary effect can be removed so that the distance and direction actually convey some information. In addition, the sSOM software can also speed up the training process. Thus, by comparing the sSOM patterns of tumor cells with that of normal cells/tissues, we could search for some candidates of tumor-specific cell surface markers more easily.

2 Method and Results

2.1 Expression Analysis.

Total RNA of mouse brain tumor cell lines was isolated from cells with RNaseasy Mini kit (Qiagen, #74104). Total RNA of mouse brain was bought from Stratagen (Cat. #736001). Twenty micrograms of total RNA was used to synthesize cDNA. Cy3-labeled cDNAs were prepared by indirect labeling method adapted from the Brown web site (http://cmgm.stanford.edu/pbrown/protocols) and hybridized to cell surface marker DNA microarray in 5xSSC/0.5%SDS solution at 55°C for 15h. Arrays were designed to contain 1402 probes (1390 for mouse genes of cell surface proteins, 7 for housekeeping genes and 5 for exogenous controls), which are spotted at least in triplicate. After washing, arrays were scanned on a FLA8000 scanner (Fuji Film, Japan). Intensity for each feature of the array was captured by using GenePix® Pro 5.1 image analysis software (Axon Instruments) and a single raw expression level for each gene was derived from the 3 spots representing each gene by mean algorithm. Then, the raw fluorescent intensity of each gene was divided by that of control spot to calculate a percentage value of each gene referred to as “relative fluorescent intensity (RFI)”, which represented the
expression level of each gene. Gene expression levels were compared to one another by RFI value to identify differentially expressed genes.

2.2 Data Preprocessing and sSOM Clustering

In order to eliminate genes that did not change significantly between tumor cell lines and normal tissue, each gene is given a score by a formula $|A - G| - V$. “A”, “G”, and “V” denotes the expression level of a gene in normal brain, the average expression level of the gene in the 5 tumor cell lines, the standard deviation of the gene expression level among the 5 tumor cell lines, respectively. Genes were eliminated if they did not show $|A - G| - V > 0$, since only the genes with a score greater than a threshold (i.e., zero) are deemed potentially significant. The expression levels were then normalized first by the maximal value of each gene among the six samples, and secondly by the maximal difference value between maximum and minimum of each gene within all the remained genes. The normalized data were clustered and displayed using sSOM.

2.3 Results

As a result, 686 genes with a potentially significant difference were analyzed by sSOM. As shown in Fig.1, there were similar expression patterns in all the five tumor cell lines, in which red, blue and white indicate that the expression level for a single gene is above, below and same as the median of that gene across all samples. Based on the sSOM patterns, several genes that were highly expressed in the tumor cell lines were selected as candidates of tumor-specific cell surface markers. For example, there were very low density lipoprotein receptor (GenBank Accession No. NM_013703), MAS1 oncogene (GenBank Accession No. X67735), 5-hydroxytryptamine (serotonin) receptor 1A (GenBank Accession No. U39391), and membrane metallo endopeptidase (CD10) (GenBank Accession No. NM_008604).

Figure 1: Clustering by spherical Self-Organizing Map (sSOM)

3 Discussion

The results show that the sSOM software is a powerful tool for data mining or knowledge discovery and visualization of high dimensional data.

References