APPLICATION NOTE

Genomics & Informatics Vol. 5(3) 124-128, September 2007

PathTalk: Interpretation of Microarray Gene-Expression Clusters in Association with Biological Pathways

Tae Su Chung[1,2], Hee-Joon Chung[1] and Ju Han Kim[1,2]*

[1]Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Korea, [2]Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea

Abstract

Microarray technology enables us to measure the expression of tens of thousands of genes simultaneously under various experimental conditions. Clustering analysis is one of the most successful methods for analyzing microarray data using the assumption that co-expressed genes may be co-regulated. It is important to extract meaningful clusters from a long unordered list of clusters and to evaluate the functional homogeneity and heterogeneity of clusters. Many quality measures for clustering results have been suggested in different conditions. In the present study, we consider biological pathways as a collection of biological knowledge and used them as a reference for measuring the quality of clustering results and functional homogeneities. PathTalk visualizes and evaluates functional relationships between gene clusters and biological pathways.

Keywords: Microarray, Cluster analysis, evaluation, biological pathways

Introduction

Microarray expression data are incessantly accumulated with the aid of recent technological advances. It is widely believed that biologically meaningful interpretations can be extracted from these large-scale data using suitable and well-organized methods of analysis. Clustering analysis is one of the most prominent methods to analyze microarray data. It explores the internal structure of complex data by organizing them into meaningful groups or gene sets. Genes of a similar expression profile may share similar functions; clustering a gene-expression profile can be used for tentative assignment of functional annotation of the unknown genes based on the functional annotations of the known genes (Eisen *et al.*, 1998; Tamayo *et al.*, 1999; Ben-Dor *et al.*, 1999; Sharan and Shamir, 2000; Sharan *et al.*, 2003)

In microarray data analysis, extracting meaningful gene-expression clusters is important because the following steps often rely on the quality of the clustering result. It is also important to measure the functional homogeneity of the clusters (or gene sets) from the results.

Quality measures for gene-expression clusters have been proposed in a variety of conditions. Assuming that the true (i.e., the gold standard) clustering solution is known, one can use the Minkowski measure (Sokal, 1977) or the Jaccard coefficient to compare the quality of different results. When the true solution is not known, there is no best measure of the quality of the result. Some have evaluated clustering results in terms of intra-cluster homogeneity using within-group similarity of gene-expression profiles only (Hansen and Jaumard, 1997; Sharan *et al.*, 2003; Yeung *et al.*, 2001). Others evaluated clustering results based on other biological knowledge. Gat-Viks *et al.* (2003) suggested a statistical measure according to prior biological knowledge, and Gibbons *et al.* proposed a way of judging the quality of clustered data by evaluating the mutual information between one gene's membership in a cluster and the attributes it possesses, given the annotation from the Saccharomyces Genome Database (Gibbons *et al.*, 2002).

Biological pathways are regarded as one of the most valuable collections of molecular biological knowledge, providing key information about the organization of biological systems. Therefore, it is natural to consider biological pathway information as a valuable resource for measuring the quality of clusters and/or the degree of homogeneity of a cluster. We first created a pathway-by-pathway similarity matrix by calculating the co-membership of genes between each pair of pathways. We then created a reference pathway map by using a multi-dimensional scaling method. This map represents a universal topological structure of genes and pathways independent of the experimental condition in which a microarray dataset is obtained, and hence it can be used as a reliable frame of reference to evaluate cluster quality and homogeneity. Gene-gene or pathway-pathway association in a given microarray experiment may differ from conditions to conditions. Mapping dataset-specific clustering results onto a universal pathway map may help to understand the

*Corresponding author: E-mail juhan@snu.ac.kr
Tel +82-2-740-8320, Fax +82-2-747-4830
Accepted 20 July 2007

underlying context of a microarray experiment. PathTalk is a web-enabled software package visualizing the reference pathway map onto which the relationship of gene-expression clusters are mapped and analyzed in terms of clustering quality and homogeneity.

Methods

Creating a reference pathway map

We collected 471 human biological pathways from the ArrayXPath knowledgebase (Chung *et al.*, 2003; Chung *et al.*, 2004), integrating pathway information from a variety of biological resources, including the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2004), GenMAPP (Dahlquist *et al.*, 2002), BioCarta (http://www.biocarta.com), and PharGKB Pathways. A pathway similarity matrix was created by calculating the following equation for each pair of pathways P_1 and P_2:

$$sim(P_1, P_2) = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|},$$

where G_1 and G_2 are gene sets in the pathways P_1 and P_2, respectively. Since this similarity table is independent of individual experimental conditions, it can be used as a reference map to measure clustering results.

Using the classical MDS (multi-dimensional scaling) algorithm, which minimizes the topological distortion, we can visualize a pathway map on 2-dimensional space. We also created a network of the pathways, representing pathway cross-talks. We defined the degree of link between a pair of pathway as the number of shared genes. When two pathways are regarded as linked if they share more than 3 genes, the whole network produces 92 connected components, including one giant component of size 374, 6 double-ton components, and 85 singletons.

Fig.1 shows a pathway cross-talk map on 2-dimensional space, where small circles represent a pathway, the color

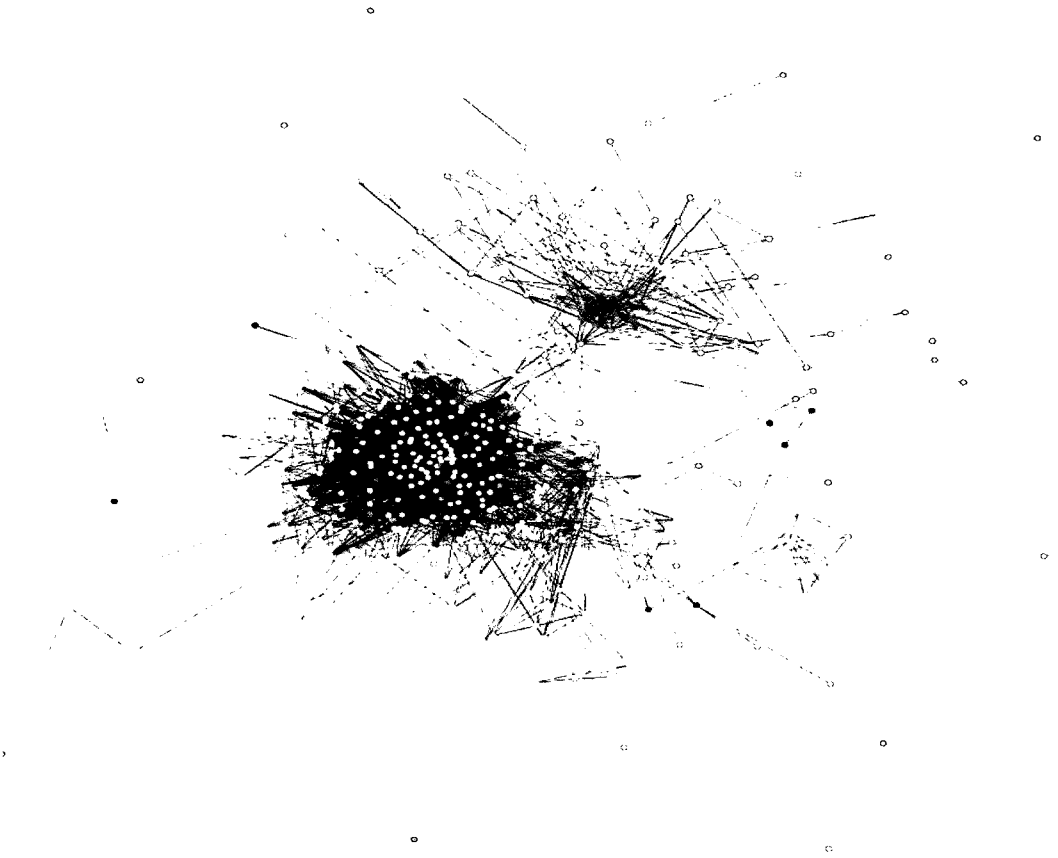

Fig. 1. Pathway cross-talk. Circles represent pathways, colors of the circles represent the sources of the pathways (red for KEGG, yellow for BioCarta, green for GenMAPP and blue for PharmGKB), and the lines joining two pathways represent those pathways sharing more than three genes. The two distinct structures in the upper right and in the lower left sides of the reference map represent the cytoplasmic and nuclear components, respectively (see method).

of a circle represents the source of the pathways (red for KEGG, yellow for BioCarta, green for GenMAPP, and blue for PharmGKB), and the line joining two pathways represents that the two pathways share more than 3 genes. Interestingly, the red KEGG circles seemed to be separately clustered from the yellow BioCarta circles, representing the separation of metabolic and signaling pathways in a cellular system. We named them as the cytoplasmic and the nuclear components, respectively.

For each given clustering result, we visualized associated pathways on the cross-talk map of Fig. 1.

Measuring pathway-homogeneity

In this subsection, we suggest a method to measure the homogeneity of a given gene set based on the topology of biological pathways. First, for a given gene set G, we define P_G by a set of pathway p such that p is significantly related with the gene set G. The Fisher's exact test is used to calculate the statistical significance. The pathway-homogeneity $Hom(G)$ of a gene set G is defined by

$$Hom(G) = Hom(P_G) = \frac{1}{|P_G|(|P_G|-1)/2} \sum_{p,p' \in P_G} sim(p, p')$$

We also define the p-value of the measure of a gene set, G, by the relative frequency of random homogeneity $Hom(G')$ being bigger than $Hom(G)$. Here the gene set G' is taken randomly more than 1,000 times, having the same size of G. We here recall that a clustering result is a list of gene sets. PathTalk provides a table of two indices of each gene set for the clustering result and hence provides a guide to select more informative gene sets for the following analysis. We use the average of pathway-homogeneities for the measure of clustering quality.

Outlier problem and dumbbell problem

In the definition of pathway-homogeneity, the gene space was transformed into pathway space. Thus, it is needed to overcome the classical problems in clustering algorithms like the outliers and the dumbbell-shaped cluster problem in the pathway space. For the outlier problem, we define a singletone index $SI(G)$ of a gene set by the following procedure

1) Define $dist(p, P_G)$ by

$$dist(p, P_G) = \frac{1}{|P_G| - 1} \sum_{p' \neq p, p' \in P_G} 1 - sim(p, p')$$

2) Let p^* be the pathway which maximize the $dist(p, P_G)$

3) Define $SI(G)$ by

$$SI(G) = \frac{Hom(P_G - \{p^*\})}{Hom(P_G)}.$$

We define a dumbbell index $DI(G)$ of a gene set by the following procedure:

1) Let p_1^* and p_2^* be the pathways that maximize $dist(p_1, p_2)$, where $dist(p_1, p_2) = 1 - sim(p_1, p_2)$
2) Divide P_G into P_1 and P_2 such that p is in P_1 if $dist(p, p_1^*) < dist(p, p_2^*)$ and p is in P_2 if $dist(p, p_1^*) > dist(p, p_2^*)$
3) Define $DI(G)$ by

$$DI(G) = \frac{Hom(P_1, P_2)}{Hom(P_G)} = \frac{(|P_1| \cdot Hom(P_1) + |P_2| \cdot Hom(P_2))/|P_G|}{Hom(P_G)}.$$

Results and Discussion

We used a human HeLa cell-cycle dataset containing 2252 genes. We clustered the gene expression vectors into 10 clusters for the evaluation of the clustering result. This clustering result was input to PathTalk. The distribution of each cluster is shown in Table 1. Fig. 2 shows a qualitative visualization of homogeneities for 10 clusters, and Fig. 3 shows quantitative pathway-homogeneities and p-values of 10 clusters.

Table 1 and Figs. 2 and 3 are the example output of PathTalk. Any clustering result written in usual tab-delimited text file can be an input of PathTalk.

Table 2 shows that clusters 1 and 9 have outlier pathways. The outlier pathways are related to two genes in each cluster. Table 2 suggests that it may return better clustering results to split cluster 1 into two sub-clusters with sizes 7 and 10, respectively, and cluster 3 into two sub-clusters of sizes 257

Table 1. Distribution of genes and their associated pathways of 10 clusters in the human HeLa cell dataset.

Cluster ID	1	2	3	4	5	6	7	8	9	10
#(genes)	17	342	288	283	231	392	243	272	21	163
#(pathway)	5	233	231	238	144	257	194	205	12	153

Table 2. Isolated genes and dumbbell index of 10 clusters in human HeLa cell data

Cluster ID	1	2	3	4	5	6	7	8	9	10
SI(G)	**1.6**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	**1.4**	1.0
outlier genes	**2**	0	0	0	0	0	0	0	**2**	0
DI(G)	**2.2**	1.1	**1.4**	1.2	1.2	1.1	1.2	1.1	1.0	1.0
$\|P_1\|$	**7**	115	**257**	172	180	190	121	133	11	45
$\|P_2\|$	**10**	227	**31**	111	51	202	122	139	10	118

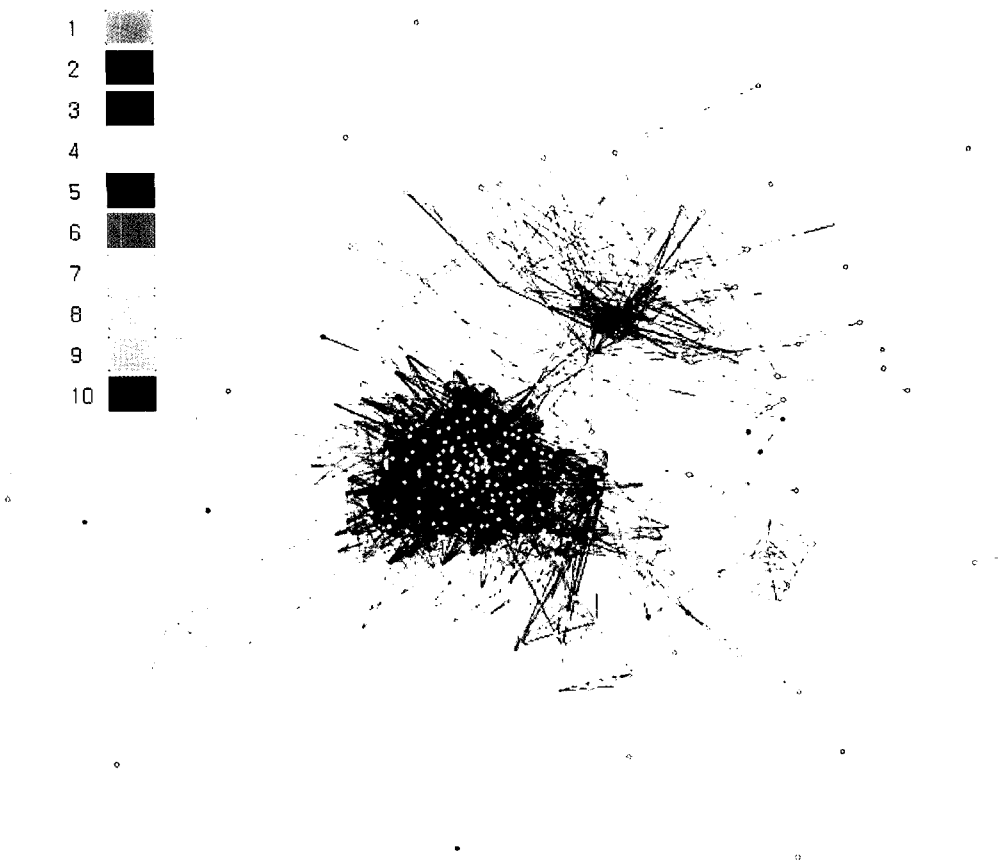

Fig. 2. Clusters superimposed on pathway cross-talk reference map. For the purpose of illustration, we highlighted the members of cluster 7 in bright green edges and those of cluster 9 in purple edges. Both clusters demonstrate good clustering of the member genes mapped onto the pathway space. Moreover, the two clusters are clearly separated even within the small area of left lower nuclear component composed mainly of BioCarta yellow circles.

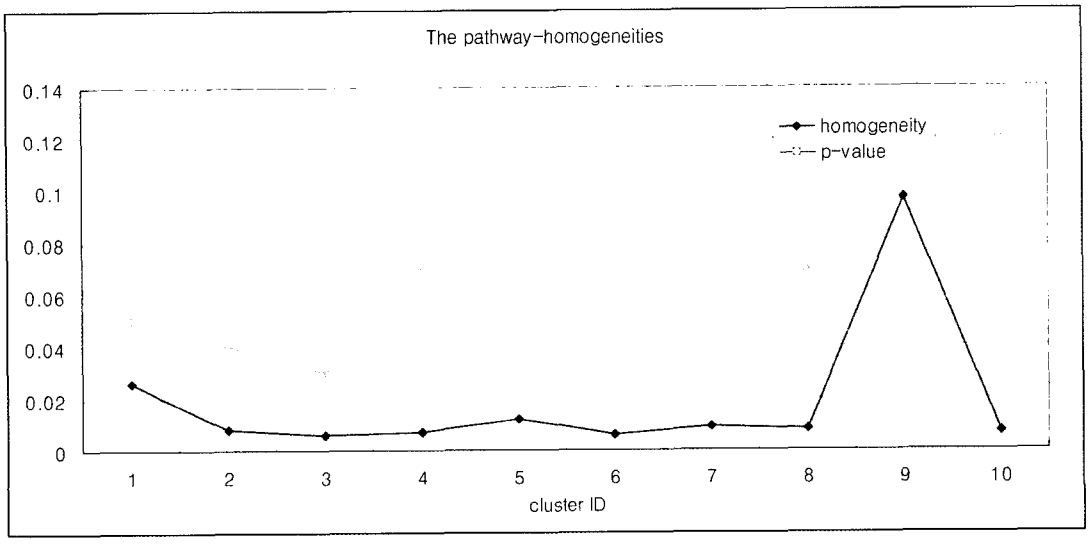

Fig. 3. Pathway-homogeneities and p-values of the 10 clusters in human HeLa cell dataset.

and 31.

In the present study, we developed PathTalk, which provides visualization and evaluation of the structural association among gene-expression clusters and biological pathways. PathTalk helps to extract high quality clusters for improved further analysis of gene-expression microarray data by visualizing the reference pathway map and systematic comparison of the clustering results.

Acknowledgements

This study was supported by a grant from Korea Health 21 R&D Project (A040163), and H.J.'s educational training was supported by a grant from the Korean Pharmacogenomics Research Network (A030001), Ministry of Health and Welfare, Republic of Korea.

References

Ben-Dor, A., Shamir, R., Yakhini, Z. (1999). Clustering gene expression patterns. *J. Comput. Biol.* 6, 281-297.

Chung, H.J., Kim, M., Park, C.H., and Kim, J.H. (2004). ArrayXPath: mapping and visualizing microarray gene expression data with integrated pathway resources using Scalable Vector Graphics. *Nucleic Acids Research* 1;32:W460-W464.

Chung, H.J., Park, C.H., Han, M.R., Lee, S., Ohn, J.H., Kim, J., Kim, J., and Kim, J.H. (2005). ArrayXPath II: mapping and visualizing microarray gene expression data with biomedical ontologies and integrated pathway resources using Scalable Vector Graphics. *Nucleic Acids Research* 1;33:W621-W626.

Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C., and Conklin, B.R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet.* 31, 19-20.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Clustering analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863-14868.

Gat-Viks, I., Sharan, R., Shamir, R. (2003). Scoring clustering solution by their biological relevance, *Bioinformatics* 19, 2381-2389.

Gibbons, F.D. and Roth F.P., (2002). Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. *Genome Research* 12, 1574-1581.

Hansen, P. and Jaumard, B. (1997). Cluster analysis and mathematical programming. *Math. Program.* 79, 191-215.

Sharan, R., Maron-Katz, A., and Shamir, R. (2003). Click and expander: a system for clustering and visualizing gene expression data. *Bioinformatics* 19, 1787-1799.

Sharan, R. and Shamir, R. (2000). CLICK: a clustering algorithm with applications to gene expression analysis. *In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)* 307-316.

Sokal, R.R. (1977). Clustering and classification: background and current directions. In Van Ryzin,J. (ed.), Classification and Clustering. Academic Press, London, pp. 1-15.

Spellman, P.T., Sherlock, G., Zhang, H.Q., Iyer, V.R., Andres, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* 9, 3273-3297.

Stephanopoulos, G., Hwang, D., Schmitt, W., Misra, J., and Stephanopoulos, G. (2002). Mapping physiological states from microarray expression measurments. *Bioinformatics* 18, 1054-1063.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* 96, 2907-2912.

Yeung, K., Haynor, D., and Ruzzo, W. (2001). Validating clustering for gene expression data. *Bioinformatics* 17, 309-318.