

# A New Clustering Method Starting from Seed Genes in Expression Profile Analysis

Ho-Youl Jung,<sup>1</sup> Ji-Eun Kim,<sup>2</sup> Miyoung Shin<sup>3</sup> Seon-Hee Park<sup>4</sup>

**Keywords:** gene expression analysis, clustering, seed-based method

## 1 Introduction

With the advances in genomics and microarray technologies and large amounts of microarray data produced, clustering has been applied to identifying groups of genes. For example, Eisen et al. applied a hierarchical clustering algorithm to identify groups of co-regulated yeast genes [1]. Tamayo et al. used SOM (self-organizing maps) to identify clusters of genes with similar expression patterns [2]. Tavazoie et al. applied  $K$ -means method to identify clusters of genes [3]. However, major problem of conventional clustering methods is that the clustering results are dependent on which clustering method or parameter is used. Users could be confused by obtaining various results according to different clustering methods.

## 2 Method: Seed Clustering Algorithm

Genes having strongly similar expression patterns would be grouped together regardless of clustering methods to be used. A set of genes grouped together by several methods is more likely to be stable and reliable. We consider this set of genes as *seed*. In our method, we firstly take the clustering results from conventional methods, e.g. hierarchical method,  $K$ -means, and SOM. Next, we compute the seed sets from the clustering results. In Fig. 1, two sets of genes -  $\{G1, G2, G8\}$  and  $\{G6, G7\}$  can be used as initial seeds. And then, clustering method can start with the seed information.

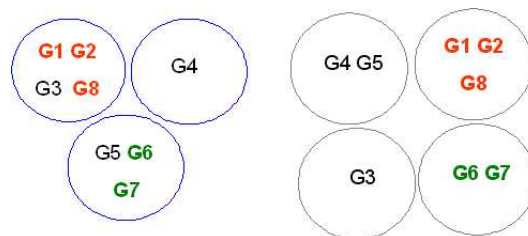


Figure 1: Extracting seed sets from clustering results:  $\{G1, G2, G8\}$  and  $\{G6, G7\}$  can be used as initial seeds.

<sup>1</sup>Bioinformatics Research Team, Electronics and Telecommunications Research Institute, 161 Gajeong-dong, Yuseong-gu, Daejeon 305-350, Republic of Korea. E-mail: hyj@etri.re.kr

<sup>2</sup>E-mail: jekim@etri.re.kr

<sup>3</sup>E-mail: shinmy@etri.re.kr

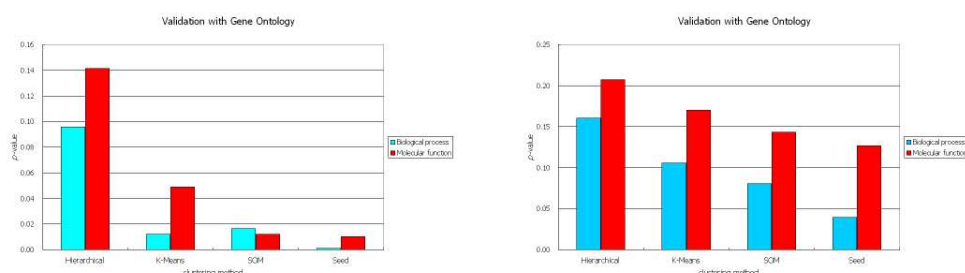
<sup>4</sup>E-mail: shp@etri.re.kr

### 3 Experimental Results

We used the hypergeometric distribution to model the probability of observing at least  $k$  genes from a cluster of size  $n$  by chance in a GO (Gene Ontology) term containing  $c$  genes from a total number  $g$  of genes which are used for clustering. For each cluster, we compute the  $p$ -values as follows:

$$p = 1 - \sum_{i=0}^{k-1} \binom{c}{i} \binom{g-c}{n-i} / \binom{g}{n}$$

In order to compare with several clustering method, we average the  $p$ -values of each clusters. Fig. 2 shows the clustering results for the two gene expression data sets- (a) is obtained from Gasch et al. [4] and (b) from Ogawa et al. [5], respectively.



(a) evaluation with data from Gasch et al. [4] (b) evaluation with data from Ogawa et al. [5]

Figure 2: Comparison with conventional clustering methods using statistical validation.

### 4 Discussions

Here we presented the usefulness of the seed clustering method in identifying biologically relevant groups of genes. Furthermore, it is possible to extract the seed genes using the GO terms in our system. In the immediate future, we will implemented the module for extracting seeds from genes which are belong to the same pathway.

### References

- [1] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *PNAS*, 95(25):14863–14868, 1998.
- [2] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R., Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *PNAS*, 96(6):2907–2912, 1999.
- [3] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M., Systematic determination of genetic network architecture, *Nature Genetics*, 22(3):281–285, 1999.
- [4] Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O., Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes, *Mol. Biol. Cell*, 11(12):4241–4257, 2000.

- [5] Ogawa, N., DeRisi, J., and Brown, P. O., New Components of a System for Phosphate Accumulation and Polyphosphate Metabolism in *Saccharomyces cerevisiae* Revealed by Genomic Expression Analysis, *Mol. Biol. Cell*, 11(12):4309–4321, 2000.