Comparison of Cluster Identification Methods for Selection of GO Terms related to Gene Clusters

YOICHI YAMADA†, YUKI MIYATA†, MASANORI HIGASHIHARA*, KENJI SATOU †

† Graduate School of Natural Science and Technology
Kanazawa University
Kakuma-machi, Kanazawa 920-1192
JAPAN
youichi@is.t.kanazawa-u.ac.jp http://bioinfo.ec.t.kanazawa-u.ac.jp

* Graduate School of Knowledge Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292
JAPAN

Abstract: - The hierarchical clustering algorithm has frequently been applied to grouping genes sharing a certain characteristic from a microarray data set. Identification of clusters from a hierarchical cluster tree is generally conducted by cutting the tree at a certain level. In this method, the most parental clusters are identified as mutually correlated gene groups and their sibling clusters are ignored. However the sibling clusters have a possibility to show more significant GO term annotation than their parental clusters. To overcome this problem, Toronen developed a novel algorithm based on the calculation of each GO annotation in all the clusters that satisfy a threshold of correlation distance. However comparison of the algorithm and the general method has not been done enough yet. Therefore we compared the general method with Toronen's proposed algorithm for identifying gene cluster-relevant GO terms. Moreover, we compared the hierarchical clustering with fuzzy k-means clustering which can group a object into more than one cluster and permit a object not to belong to any clusters. Consequently, we confirmed that Toronen's algorithm is more available for identification of gene clusters and their relevant GO terms from a microarray data set than the other methods.

Key-Words: - Hierarchical clustering, Gene Ontology, Microarray data, Yeast cell cycle

1 Introduction

Genome projects in several organisms have revealed almost a complete set of genes on the genome of each organism. As a result, it is uncovered that *Homo sapiens* possesses genes of approximately 30,000, and *Saccharomyces cerevisiae*, which is a representative of unicellular eukaryote, has genes of about 6,000 [1]-[3]. At present, we can freely obtain the sequence of the genes in each organism from public databases.

When a gene functions in a cell, it is transcribed to messenger RNAs (mRNA) and often translated to proteins. Genes do not necessarily work all the time in a cell, and their expression is regulated by transcription factors, noncoding RNAs, epigenetic modifications of DNA or histone proteins, etc [4],[5]. Transcription factors often bind to the promoter regions of genes and induce or suppress the formation of the transcription initiation complex. Epigenetic modification is a chemical modification (i.e., methylation, acetylation, phosphorylation, etc) of DNA or histone proteins. For instance, when cytocines of CpG dinucleotides in the promoter region are methylated, transcription factors can not directly or indirectly bind to the promoter region. On the other hand, histone acetylation cause the decondensation of the chromatin, and transcription factors can access to the regions. Consequently, these expression regulation mechanisms allow genes to be expressed or suppressed in response to changes of internal and external environments of a cell or during the development.

A protein generally cooperates with several other proteins on a task within a cell or between cells. Therefore a protein often binds to other proteins and

works as a member of one protein complex. Because of this, when a protein complex works on a task, all members of the protein complex are required for completeness of the task. In this point, genes from one protein complex are coordinately expressed in time course (e.g., process of cell division cycle). Therefore if identify coordinately expressed genes in time course, we can predict the biological role of function-unknown genes from function-known genes within those genes.

Microarray technology allows us to monitor the expression of thousands of genes simultaneously [6]. As a first step for analyzing microarray data, a clustering algorithm is often applied and yields gene clusters sharing a certain characteristic [7]-[11]. Cluster analysis divides objects into groups so that similar objects belong to the same cluster and dissimilar objects to different clusters. For instance, time-course microarray experiments result in microarray data from a series of time points, and then the hierarchical clustering algorithm is applied to the microarray data and gene groups showing a similar expression pattern across a set of time points are identified. Since genes grouped by the clustering algorism show similar expression patterns across a series of time points, they have a possibility to perform functionally related tasks: they may be functionally correlated.

However, even if we know only the name of genes within the group, we will not understand what functional characteristics they share. In such cases, Gene Ontology (GO) is frequently used to give the genes any biological annotations [12]-[14]. The biological annotations are called GO terms and curated by GO Consortium. The GO terms annotate a number of genes from organisms of more than 50 species.

Accordingly, GO terms commonly associated with genes in a cluster are biological characteristics which they share. The significance of GO term annotation to genes is statistically tested [15]-[17].

Grouping of genes using microarray data is often performed by the hierarchical clustering [7]. The identification of gene clusters by the hierarchical clustering is generally performed by a break of merging of sibling clusters according to a threshold of correlation distance between genes. However disregard of sibling clusters of the identified clusters prevents us from detecting gene clusters with important characteristics. For instance, although statistical significance of GO term annotation to genes is influenced by the number of genes within a cluster, parental clusters always contain larger number of genes than their sibling clusters. Accordingly, sibling clusters have a possibility to show statistically significant GO annotations which their parental clusters do not show. In this context, Toronen reported a method in which statistical significance of each GO term annotation is examined in all the clusters (i.e., including both parental and sibling clusters) which satisfy a threshold of correlation distance between clusters [18],[19]. However there are few reports including his report that show its availability based on comparison with the general method. We therefore compared the results from both methods using a yeast cell cycle microarray data set.

On the other hand, a clustering method different from hard clustering (e.g., hierarchical clustering and k-means clustering) is also applied to grouping of genes from microarray data. Fuzzy k-means clustering is classified as a soft clustering method in which each object can belong to plural clusters or to no cluster. Gasch et al. applied this method to gene clustering from microarray data and confirmed its availability. Therefore we also applied the fuzzy k-means clustering algorithm to grouping genes and identification of their related GO terms using the yeast cell cycle microarray data set. Finally, we compared results from the hierarchical clustering with those from the fuzzy k-means clustering.

2 Problem Formulation

2.1 Hierarchical clustering algorithm

Identification of clusters from a hierarchical cluster tree is generally performed by cutting the tree at a certain level. Fig. 1 shows an example of the process. The tree in Fig. 1 is cut at the level of the dotted line, and cluster 5, 8 and 9 are identified as clusters containing mutually correlated genes. In this case, other sibling clusters (1, 2, 3, 4, 6, 7) are ignored for the next analysis. In the next analysis, enrichment of each GO annotation to genes within the clusters is statistically estimated, and GO terms showing a significant *p*-value (e.g., $p < 0.05$) are assigned to the clusters.

Here it is possible that a GO term shows more statistically significant annotation in sibling clusters than in their parental clusters. For instance, although *p*-values of GO term ‘a’ annotation in all the clusters below the dotted line are shown in Fig. 1, cluster 2, 6 and 7 show lower p-values than their parental clusters

and also satisfy the criterion of under 0.05.

In this case, cluster 2, 6 and 7 should be identified for GO term 'a' than cluster 8 and 9. In this context, Toronen reported a method in which statistical testing of each GO term annotation is conducted for all the clusters at lower level than a threshold of correlation distance, and the cluster showing the most significant *p*-value in the same branch of the cluster tree is selected for the GO term.

It seems to be a reasonable method but there are few studies comparing the general method with Toronen's method. To examine availability of the method proposed by Toronen, we compared the general method with the proposed method using a yeast cell cycle microarray data set.

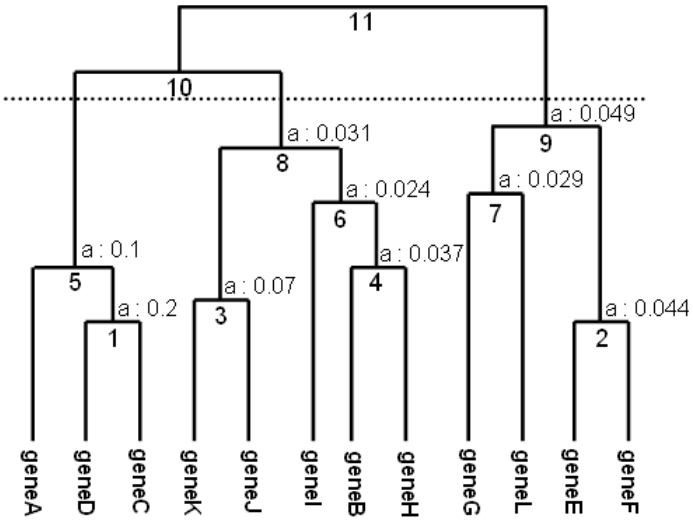

Fig. 1: A hierarchical cluster tree of genes constructed from a microarray data set

2.2 Fuzzy k-means clustering algorithm

Fuzzy k-means clustering is that introduce the conception of fuzzy in k-means clustering. It conducts a clustering that permit objects to ambiguously belong to groups [20],[21]. In hard clustering methods including the hierarchical clustering and the k-means clustering, a object certainly belongs to one cluster and can not belong to plural clusters.

On the other hand, fuzzy k-means clustering uses membership degrees of objects to clusters. Depending on the membership degree, fuzzy k-means clustering allows objects to belong to more than one group or not to belong any groups.

Gasch et al. applied fuzzy k-means clustering to grouping of genes from microarray data and confirmed availability of its method.

However few reports compare fuzzy k-means clustering with hierarchical clustering for identification of gene clusters and their relevant GO terms from microarray data. Therefore we also compared Fuzzy k-means clustering with hierarchical clustering for identification of gene clusters and their relevant GO terms using the microarray data set of the yeast cell cycle.

3 Materials and methods

3.1 Microarray data set

The microarray data set used in this study was produced by Spellman et al [22]. In the data set, yeast cells were alpha factor-arrested and synchronized, and periodically recovered in a series of time points after release. Then RNAs were extracted from recovered yeast cells. Control cells were also recovered from asynchronous yeast cells growing in the same culture condition at the same time points and their RNAs were extracted in the same way. Fluorescently labeled cDNA was synthesized from each extracted RNA and the ratio of experimental to control cDNA was measured every recovery time point. The expression ratio of each gene in obtained data was subject to logarithmic conversion.

3.2 Construction of a hierarchical cluster tree from a microarray data set

Cosine coefficient distances were calculated in all the possible gene pairs from the yeast cell cycle microarray data set using the open source clustering software, Cluster3 [23]. Cluster3 calculates the cosine coefficient using the following equation 1,

$$\text{cosine distance} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i y_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2}} \qquad (1)$$

where n is the number of microarray data in the yeast cell cycle microarray data set, x_i and y_i depict expression ratios of two distinct genes in each microarray data. Centroid linkage method was applied to the formation of a hierarchical cluster tree using the Cluster3.

3.3 Identification of clusters and their relevant GO terms from a hierarchical cluster tree by the general method

A hierarchical cluster tree was cut at the level of each threshold (0.7, 0.75, 0.8, 0,85, 0.9) of correlation

distance. Then the most parental clusters at the lower part of separated trees were identified. For instance, the hierarchical cluster tree of Fig. 2 was cut at the dotted line, and cluster 5, 8 and 9 are identified. Next, enrichment of GO term annotations to the identified clusters was calculated according to the following equation 2:

$$p-\text{value} = \sum_{j}^{\min(n,M)} \frac{{}_{M}C_{j} \cdot {}_{N-M}C_{n-j}}{{}_{N}C_{n}} \qquad (2)$$

where N is the number of genes examined by the microarray experiment which we refer to as "population gene set", M is the number of genes annotated to a GO term in the population gene set, n is the number of genes within a cluster, and j is the number of genes assigned to the GO term in the cluster. GO term annotations (category of biological process) of about 15,000 are calculated in each identified cluster. GO terms showing p-values under a threshold (e.g., 0.05) are identified as common characteristics among genes in each cluster. In the example of Fig. 2, annotation significance of GO term 'a' in identified cluster 5, 8 and 9 is calculated and cluster 8 and 9 showing statistical significance of <0.05 are identified as clusters of GO term 'a'. Although results of multiple comparisons require to be corrected, no correction is applied to those results because Bonferroni correction is so strict that a few GO terms show a statistical significance.

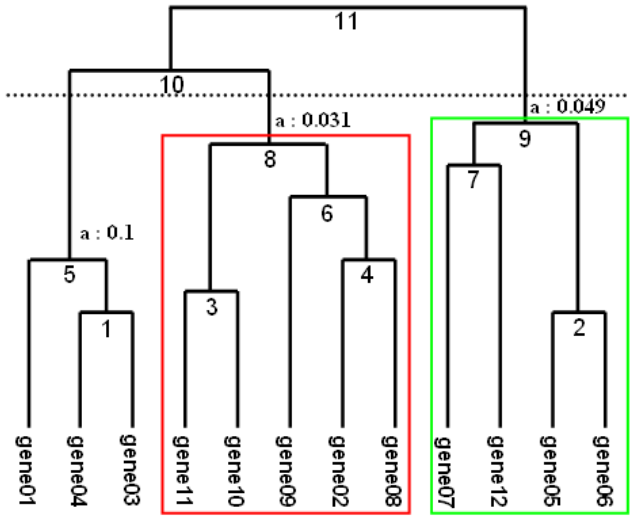

Fig. 2: Identification of clusters for GO term 'a' by the general method

3.4 Identification of clusters and their relevant GO terms from a hierarchical cluster tree by Toronen's method

A hierarchical cluster tree was cut at the level of each threshold (0.7, 0.75, 0.8, 0,85, 0.9) of correlation distance. The statistical enrichment of each GO annotation in all the clusters that fulfill each threshold (0.7, 0.75, 0.8, 0,85, 0.9) are calculated using the equation 1. GO term annotations (category of biological process) of about 15,000 are calculated in each cluster. GO terms showing p-value below 0.05 are assigned to the clusters. Repeated hits in the same branch of the cluster tree are discarded and the most statistically significant GO term is assigned to each branch of the cluster tree. For instance, cluster 4, 6 and 8 on the same branch in Fig. 3 display statistical significance of <0.05 and cluster 6 shows the lowest p-value. In this case, cluster 6 with lowest p-value is assigned to GO term 'a'. Note that discarded GO terms can still be assigned to other parts of the cluster tree. For instance, cluster 2 and 7 besides 6 are assigned to GO term 'a' in Fig. 3. No correction is applied to the results of these multiple comparisons due to the same reason as above description.

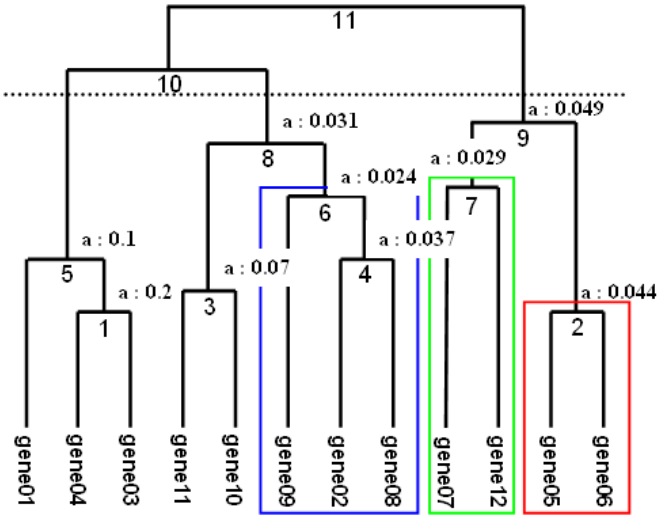

Fig. 3: Identification of clusters for GO term 'a' by Toronen's method

3.5 Identification of clusters and their relevant GO terms by fuzzy k-means clustering

Fuzzy k-means clustering is conducted on the basis of previous method [24]. Briefly, the following steps are carried out in fuzzy k-means clustering.

1. Prototype centroids of k/3 (large colored circles in Fig. 4) were identified as the most informative

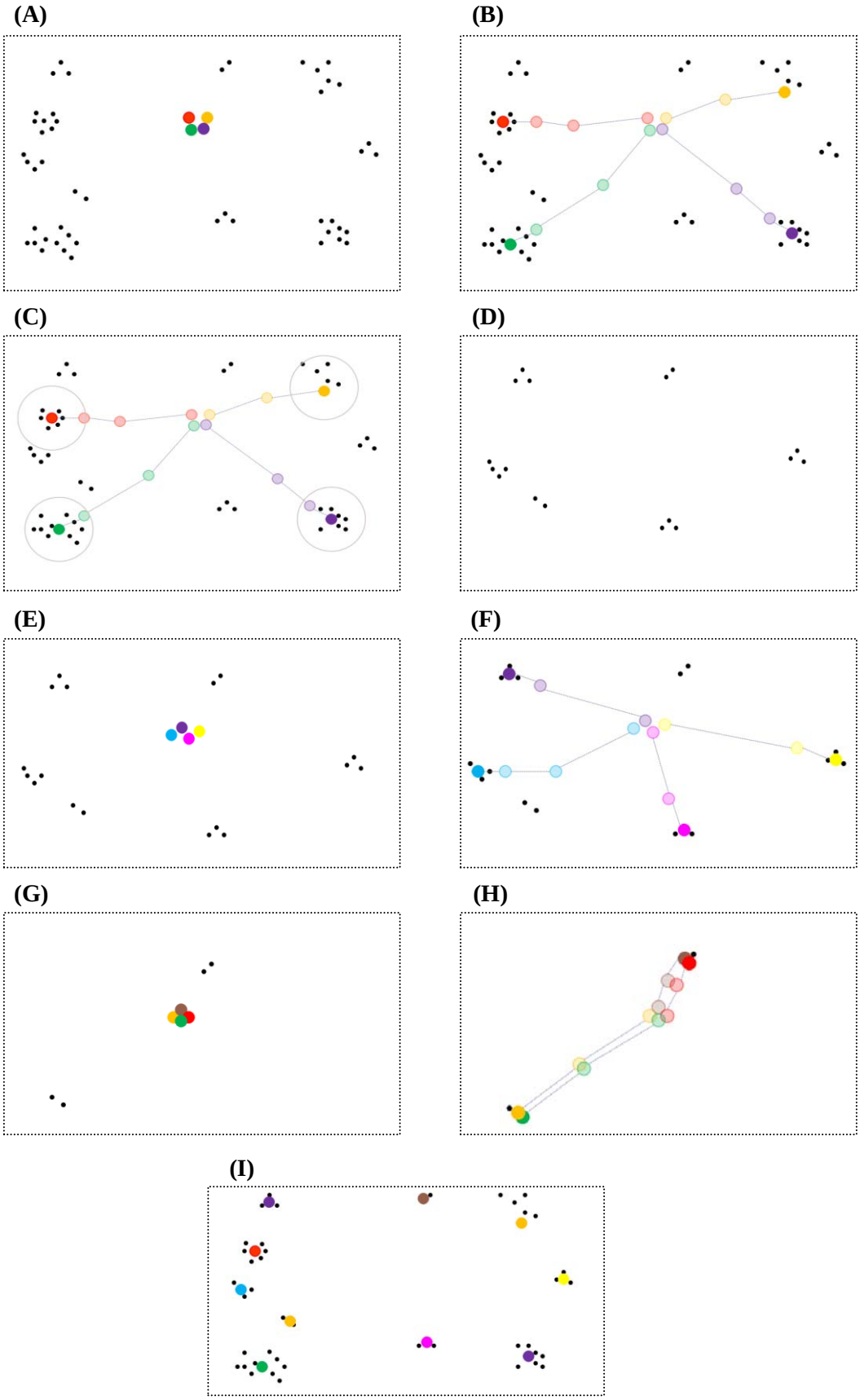

Fig. 4: Process of fuzzy k-means clustering

eigen vectors by principal component analysis [25] in the first round clustering. K is the number of clusters that finally generate, and 3 is the number of clustering cycles (Fig. 4 A).

2. Pearson correlation coefficient is calculated between each prototype centroid and any genes. Membership degrees to each centroid were calculated in each gene using the following equation 3:

$$m_{x_i v_j} = \frac{\frac{1}{d^2_{x_i v_j}}}{\sum_{j=1}^{K} \frac{1}{d^2_{x_i v_j}}} \qquad (3)$$

where m_{xivi} is the membership degree of gene xi to prototype centroid vj, d_{xivi} is the Pearson correlation distance between x_i and v_j. Each gene weight w_{xi} is also calculated as described in [24].

3. New centroids v'_j are calculated as

$$V'_j = \frac{\sum_{i=1}^{N} m^2_{x_i v_j} W_{x_i} X_i}{\sum_{i=1}^{N} m^2_{x_i v_j} W_{x_i}} \qquad (4)$$

where m_{xivi} is the membership degree of gene x_i to prototype centroid v_j, w_{xi} is the gene weight. The centroid refinements are continued until the average change of gene memberships becomes < 0.001 (see Fig. 4 B).

4. Genes showing correlation coefficient of >0.7 to any centroids are removed from the dataset and all centroids are also removed (see Fig. 4 C,D).

5. Process of 1-4 is also repeated in the next clustering cycle. Consequently, gene clusters showing correlation coefficient of >0.7 to each centroid are obtained (see Fig. 4 E-I).

The statistical enrichment of each GO annotation in all the identified clusters is calculated using the equation 1. GO term annotations (category of biological process) of about 15,000 are calculated in each cluster. GO terms showing p-value below 0.05 are assigned to the clusters. No correction is applied to the results of these multiple comparisons due to the same reason as above description.

4 Results

4.1 Comparison of the general method and Toronen's method in hierarchical clustering

To compare the general method and Toronen's method in the hierarchical clustering, gene clusters showing a wide variety of distances between genes were used. Eight clusters in Table 1 were identified by Spellman et al. using a Fourier algorithm and a correlation algorithm [22]. Genes within the clusters show periodically oscillated expression during the yeast cell cycle (see Fig. 5). We calculated the correlation distance between genes within each cluster using the centroid linkage method in the hierarchical clustering (see Table 2 and 3). For instance, since genes in 'Histone cluster' show uniform expression patterns during the cell cycle, they show high correlation distance of 0.916 (see Fig. 5 and Table 1). In contrast, since genes in MAT cluster show inconsistent expression patterns, they show low correlation distance of -0.495 (see Fig. 5 and Table 1).

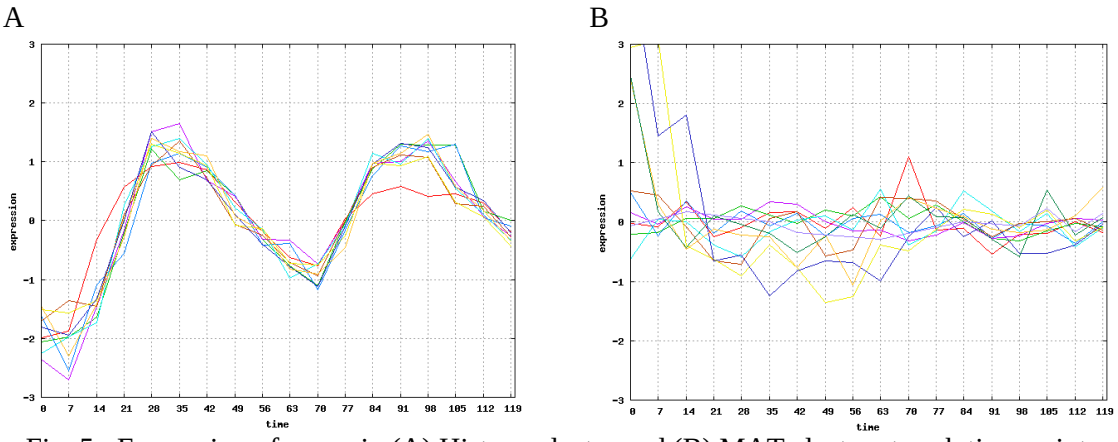

Fig. 5: Expression of genes in (A) Histone cluster and (B) MAT cluster at each time point.

Table 1: Comparison of the general method and Toronen's proposed method

Gene cluster	Distance between genes	Overrepresented GO terms	Threshold of correlation distance in hierarchical clustering				
			0.7	0.75	0.8	0.85	0.9
CLN2 cluster	0.589	145	61/44	61/40	61/61	62/44	61/45
Y' cluster	0.732	6	1/0	1/0	1/0	1/1	1/1
MAT cluster	-0.495	17	0/0	0/0	0/0	0/0	9/9
MCM cluster	0.364	42	30/17	30/17	21/16	6/4	2/2
SIC1 cluster	0.474	12	12/7	12/8	12/9	12/12	12/12
CLB2 cluster	0.615	71	28/0	28/0	34/25	34/25	32/21
Histone cluster	0.916	37	30/30	32/27	32/27	32/27	32/30
MET cluster	0.582	32	19/14	19/18	7/7	0/0	0/0

If Toronen's proposed method is more suitable for identification of clusters and their relevant GO terms from microarray data than the general method, it will identify more GO terms significantly assigned to clusters with a high correlation distance than the general method but not with a low correlation distance. To examine this supposition, we first determined significantly overrepresented GO term annotations ($p < 0.05$) in the clusters identified by Spellman et al., as shown in Table 1. 'Overrepresented GO terms' in Table 1 means the number of GO terms significantly assigned to each 'Gene cluster'. Numbers to the left and the right of each slash in Table 1 represent the numbers of 'Overrepresented GO terms' identified by Toronen's method and the general method, respectively. Note that not only 'Overrepresented GO terms' but also genes annotated to those GO terms overlapped between 'Overrepresented GO terms' in each 'Gene cluster' and those identified by both Toronen's method and the general method. In most of thresholds of correlation distance, Toronen's method could identify more overrepresented GO terms in gene clusters showing positive correlation distances (CLN2, SIC1, CLB2, Histone and MET) than general method (see Table 1). In contrast, MAT cluster showing the negative correlation distance displayed no difference between Toronen's method and the general method in all thresholds of correlation distance (see Table 1). This suggests that identification probability of false positive is identical between both methods. Moreover, although the MAT cluster shows low correlation distance of -0.495, both methods identified 9 overrepresented GO terms of MAT cluster in the threshold of 0.9. Calculation results of correlation distances between genes within MAT cluster are shown in Table 2. Since the correlation distance between two genes (YCL055W and YCL027W) was high correlation distance of 0.927, the general method and Toronen's method seemed to identify 9 GO terms associated with these two genes. In contrast, correlation distances between genes in histone cluster were high (>0.9) in all nodes (see Table 3).

Table 2: Calculation results of correlation distances between genes in MAT cluster

Node	Component	Correlation distance
Node 1	YCL055W, YCL027W	0.927
Node 2	Node 1, YLR452C	0.87
Node 3	Node 2, YNR044W	0.817
Node 4	YDR493W, YCR018C	0.594
Node 5	YJR004C, YLR040C	0.52
Node 6	Node 5, YKL177W	0.492
Node 7	YGL090W, YKL178C	0.452
Node 8	Node 6, Node 3	0.344
Node 9	Node 4, Node 8	0.039
Node 10	Node 7, Node 9	−0.495

Table 3: Calculation results of correlation distances between genes in Histone cluster

Node	Components	Correlation distance
Node 1	YNL030W, YDR224C	0.985
Node 2	Node 1, YNL031C	0.985
Node 3	Node 2, YBR010W	0.979
Node 4	Node 3, YDR225W	0.978
Node 5	YBR009C, Node 4	0.97
Node 6	YBL002W, YBL003C	0.966
Node 7	Node 5, Node 6	0.965
Node 8	Node 7, YPL127C	0.916

Table 4: Number of 'Overrepresented GO terms' in Table 1 identified by fuzzy k-means clustering

Gene cluster	Threshold of membership degree											
	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.15	0.2	0.25	0.3	0.35
CLN2 cluster	60	46	48	44	29	25	30	29	22	22	36	35
Y' cluster	5	5	5	5	0	0	0	0	0	0	0	0
MAT cluster	0	0	0	0	0	0	0	0	0	0	0	0
MCM cluster	22	14	8	8	8	8	8	8	8	0	0	0
SIC1 cluster	7	6	7	9	8	10	11	13	13	13	13	13
CLB2 cluster	22	23	24	29	36	38	38	31	25	23	23	23
Histone cluster	15	21	24	21	23	20	24	27	25	27	30	32
MET cluster	14	3	6	0	0	0	0	0	0	0	0	0

Although Y' cluster showed comparatively high correlation distance of 0.735, a few overrepresented GO terms were identified in that cluster. The Y' cluster contained 31 ORFs sharing high DNA sequence similarity, and the ORFs were also repeated sequences found subtelomeric regions on many chromosomes. Moreover, they did not strictly form a functional category but had similarity to RNA helicases. Since a few GO terms were assigned to genes within Y' cluster, few overrepresented GO terms seemed to be identified in this cluster.

4.2 Fuzzy k-means clustering

Fuzzy k-means clustering is classified into a soft clustering because not only can group a object into more than one cluster but also permit a object not to belong to any clusters. Therefore we thought it may be more suitable for identifying clusters and their related GO terms than the hierarchical clustering.

Fuzzy k-means clustering requires determining a threshold of membership degree. It discards genes showing membership degrees of under a threshold to any centroids, and groups genes showing membership degrees of more than the threshold into clusters. Then GO terms showing significant annotation to identified gene clusters are identified as the cluster-relevant GO terms.

We examined how many 'Overrepresented GO terms' of Table 1 the fuzzy k-means clustering identify every threshold of the membership degree (see Table 4). We changed the threshold of membership degree from 0.04 to 0.35 and identified gene clusters satisfied with each threshold. Then statistical significance of GO term annotations to the clusters was examined, and GO terms with p-value of <0.05 were identified as cluster-related GO terms.

Table 4 shows the total number of 'Overrepresented GO terms' of Table 1 identified by the fuzzy k-means clustering. The first row index in Table 4 shows the thresholds of the membership degree of genes to centroids. In membership degrees of more than 0.3, genes within clusters identified by fuzzy k-means clustering showed correlation distance of more than 0.85. Compared to numbers of 'Overrepresented GO terms' of Table 1 identified by Toronen's method, fuzzy k-means clustering identified less 'Overrepresented GO terms' in highly correlated clusters (correlation distance of >0.85). This might result from the difference of calculation methods of correlation distance between hierarchical clustering

and fuzzy k-means clustering. Although the hierarchical clustering calculates the correlation distance between genes, fuzzy k-means clustering calculates the correlation distance between centroids and genes. Therefore, the fuzzy k-means clustering might tend to identify clusters with lower correlation distance between genes than the hierarchical clustering. Actually, fuzzy k-means clustering could identify as many 'Overrepresented GO term' as hierarchical clustering in lower thresholds of the membership degree. Another interpretation is that cluster size identified is different between the hierarchical clustering and the fuzzy k-means clustering. The size of clusters identified by fuzzy k-means clustering tended to be larger than those identified by hierarchical clustering because it classified one gene into plural clusters. This effect may increase p-values of GO term annotations in clusters identified by the fuzzy k-means clustering.

Furthermore, fuzzy k-means clustering identified no 'Overrepresented GO term' that is significantly associated with MAT cluster in all the thresholds of membership degree (see Table 4). This consequence suggests that identification rate of false positive is almost identical between hierarchical clustering and fuzzy k-means clustering.

Consequently, Toronen's method in hierarchical clustering is suggested to be more suitable for identifying gene clusters and their relevant GO terms from a microarray data set.

5 Conclusion

In this study, we compared three methods (i.e., the general and Toronen's methods in hierarchical clustering and the general method in fuzzy k-means clustering) for identification of clusters and their relevant GO terms from a microarray data set. As a result, Toronen's method in hierarchical clustering could identify more GO terms significantly associated with gene clusters showing positive correlation distances than the other methods. Moreover, the number of GO terms significantly associated with gene clusters showing the negative correlation distance is almost identical between three methods: identification rate of false positive is identical between three methods. Consequently, our simulation confirmed availability of Toronen's method for identification of clusters and their relevant GO terms from a microarray data set.

References:

[1] E.S. Lander et al., Initial sequencing and analysis of the human genome, *Nature*, Vol.409, 2001, pp.860-921.

[2] J.C. Venter et al., The sequence of the human genome, *Science*, Vol.291, 2001, pp.1304-1351.

[3] A. Goffeau, B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S.G. Oliver, Life with 6000 genes, *Science*, Vol.274, 1996, pp.563-567

[4] O. Hobert, Gene regulation by transcription factors and microRNAs, *Science*, Vol.319, 2008, pp.1785-1786.

[5] R. Jaenisch, A. Bird, Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals, Nat. Genet., Vol.33, 2003. pp.245-254.

[6] D. Shalon, S.J. Smith, P.O. Brown, A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization, *Genome Res*., Vol.6, 1996, pp.639-645.

[7] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, Vol.95, 1998, pp.14863-14868.

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

[9] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, W.L. Ruzzo, Model-based clustering and data transformations for gene expression data, *Bioinformatics*, Vol.17, 2001, pp.977-87.

[10] S. Datta and S. Datta, Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes, *BMC Bioinformatics*, Vol.7, 2006, pp.397.

[11] F.D. Gibbons, F.P. Roth, Judging the quality of gene expression-based clustering methods using gene annotation, *Genome Res*., Vol. 12, 2002, pp.1574-1581.

[12] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L.

Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet.*, Vol.25, 2000, pp.25-29.

[13] Gene Ontology Consortium, Creating the gene ontology resource: design and implementation, *Genome Res*., Vo.11, 2001, pp.1425-1433.

[14] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N, Harte, R. Lopez and R. Apweiler, The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology, *Nucleic Acids Res*., Vol.32 (Database issue), 2004b, D262-266.

[15] T. Beissbarth, T. P. Speed, GOstat: find statistically overrepresented Gene Ontologies within a group of genes, *Bioinformatics*, Vol.20, 2004, pp1464-1465.

[16] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, G. Sherlock, GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *Bioinformatics*, Vol.20, 2004, pp.3710-3715.

[17] R.S. Sealfon, M.A. Hibbs, C. Huttenhower, C.L. Myers, O.G. Troyanskaya, GOLEM: an interactive graph-based gene-ontology navigation and analysis tool, *BMC Bioinformatics*, Vol.7, 2006, pp.443.

[18] P. Toronen, Selection of informative clusters from hierarchical cluster tree with gene classes, *BMC Bioinformatics*, Vol.5, 2004, pp.32.

[19] M. Kankainen, G. Brader, P. Törönen, E.T. Palva, L. Holm, Identifying functional gene sets from hierarchically clustered expression data: map of abiotic stress regulated genes in Arabidopsis thaliana, *Nucleic Acids Res*., Vol.34, 2006, e124.

[20] H.C. Liu, D.B. Wu, H.L. Ma, Fuzzy clustering with new separable criterion, *WSEAS Transactions on Biology and Biomedicine*, Vol.4, 2007, pp.99-102.

[21] H.C. Liu, D.B. Wu, J.M. Yih, S.W. Liu, Fuzzy possibility c-mean based on mahalanobis distance and separable criterion, *WSEAS Transactions on Biology and Biomedicine*, Vol.4, 2007, pp.93-98.

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

[23] M.J.L. de Hoon, S. Imoto, J. Nolan, S. Miyano, Open source clustering software, *Bioinformatics*, Vol.20, 2004, pp.1453-1454.

[24] A.P. Gasch, M.B., Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering, *Genome Biol*., Vol.3, 2002, RESEARCH0059.

[25] F. Shao, G. Jiang, M. Yu, Color correction for multi-view images combined with PCA and ICA, *WSEAS Transactions on Biology and Biomedicine*, Vol.4, 2007, pp.73-79.