Analyzing Fuzzy Partitions of *Saccharomyces Cerevisiae* Cell-Cycle Gene Expression Data by Bayesian Validation Method

Si-Ho Yoo, Chanho Park and Sung-Bae Cho, *Member, IEEE*

Abstract— **Clustering of gene expression profiles has been used for gene function identification. Since the genes usually belong to multiple functional families, fuzzy clustering methods are appropriate. However, a natural way to measure the quality of the fuzzy cluster partitions is still required. In this paper, a Bayesian validation method for fuzzy partition selection with the largest posterior probability given the dataset is proposed. This method is compared to four representative fuzzy cluster validity measures using fuzzy c-means algorithm on four well-known datasets in terms of the number of clusters predicted in the data. An analysis of *Saccharomyces cerevisiae* cell cycle gene expression data follows to show the usefulness of the proposed method.**

Index Terms—**clustering, gene expression profiles, fuzzy clustering, Bayesian validation method, *Saccharomyces cerevisiae* cell cycle gene expression data**

I. Introduction

CLUSTERING groups thousands of genes by expression similarity can be used to identify gene function and assign putative function to genes with unknown function but similar expression to genes with known function [1]. Hard clustering, a hard partitioning method, assigns a sample to only one group. However, real world data such as gene expression profiles do not have clear boundaries and cannot therefore be easily partitioned. Since some genes also belong to multiple functional families, analyzing the genes by hard clustering method has limitations. Unlike hard clustering, fuzzy clustering assigns a sample to multiple groups by membership value [2]. Fuzzy clustering method is more robust to noise and more appropriate in analyzing gene expression profiles than hard clustering method [3].

The most important issues with any clustering method are the identification of the actual number of clusters as well as how "good" these clusters are. Thus, it is necessary to evaluate each fuzzy partition in terms of cluster validity. Various investigations about these matters have been conducted. Partition coefficient and classification entropy were first proposed by Bezdeck [4]. These two cluster validity indexes are widely used for the fuzzy clustering validity. These indexes decide the number of optimal partitions at maximum validity measures. The validation indices poposed by Xie-Beni [5] and Sugeno [6] are popular in the field of fuzzy clustering. The Xie-Beni index is a ratio of the within cluster sum of squared distances to the product of the number of elements and the minimum between cluster separations, and the Fukuyama Sugeno index measures the compactness and separation of the resulting fuzzy partition after a dataset has been separated into several clusters. However, since the conventional validity indexes are based on the distance between the clusters, we cannot fully represent the structure of the dataset [7].

In this paper, we propose a Bayesian validation method, which evaluates the result of clustering by posterior probability of the fuzzy partitions of given dataset. Unlike conventional validity measures, Bayesian validation never makes use of the inter-cluster distance. Instead, this approach selects the partitioning with the largest posterior probability in a given dataset. The usefulness of the proposed method is demonstrated with four well-known datasets: Wine, Image, Iris, and Small Round Blue Cell Tumor (SRBCT) dataset where the fuzzy c-means algorithm is used as the clustering algorithm. Especially, it is compared with others in terms of the number of clusters predicted in the data. After, *Saccharomyces cerevisiae* cell-cycle gene expression data is analyzed by the proposed method. The fuzzy genes of this data are analyzed through the PCA 3D plot.

II. Background

A. Fuzzy C-means Algorithm

The most widely used fuzzy clustering algorithm is fuzzy c-means algorithm proposed by Bezdeck [8]. It generates a fuzzy partition providing a degree of membership of each data to a given cluster. The procedure of the algorithm is shown in Figure 1. The membership values lie between 0 and 1. Values close to 0 indicate the absence of a strong association to the corresponding cluster. Similarly, values close to 1 indicate a

Manuscript received July 9, 2004. This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University.

S.-H. Yoo is with the Department of Computer Science, Yonsei University, Seoul, Korea. (e-mail: bonanza@sclab.yonsei.ac.kr).

C. Park is with the Department of Computer Science, Yonsei University, Seoul, korea. (e-mail: cpark@sclab.yonsei.ac.kr).

S.-B. Cho is with the Department of Computer Science, Yonsei University, Seoul, Korea. (phone: 82-2123-2720; fax: 82-365-2579; e-mail: sbcho@sclab.yonsei.ac.kr).

0-7803-8728-7/04/$20.00 ©2004 IEEE

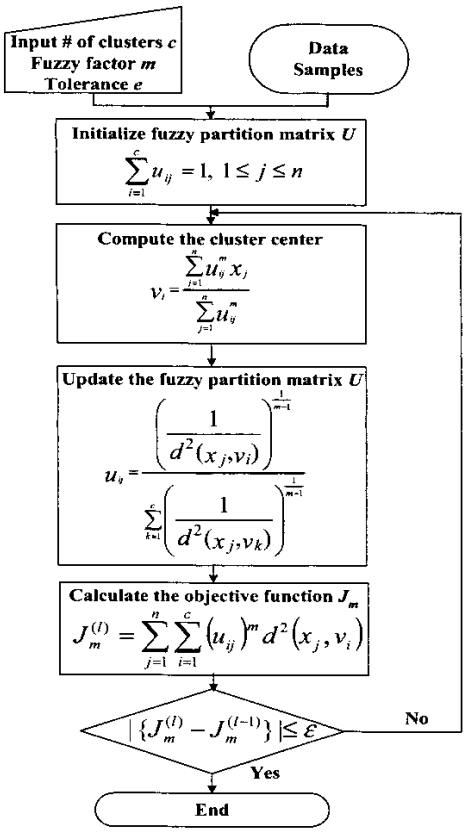


Fig. 1. Fuzzy c-means algorithm

strong cluster association. The objective of this algorithm was to minimize the objective function J_m to generate the optimal fuzzy partitioning for a given dataset $X=\{x_1, x_2, \ldots, x_n\}$, where m is a real-valued number which controls the 'fuzziness' of the resulting clusters and u_{ij} is the membership degree of data x_j to a cluster i, an element of a $(c \times n)$ pattern matrix $U=[u_{ij}]$. $d^2(x_j,v_i)$ corresponds to the square of the Euclidean distance between a data x_j and the cluster center v_i.

B. Conventional Cluster Validity Measures

Table I shows the four most widely used cluster validity measures. Bezdeck proposed Partition Coefficient (PC) and Classification Entropy (CE) for fuzzy cluster validation [4]. The optimal partitioning was obtained by maximizing the value of PC and minimizing the value of CE with respect to certain value of c (number of clusters). Sugeno tried to model the cluster validation by exploiting the compactness and separateness [6]. Smaller values of this model (FS) indicate better partitioning of a given dataset. Xie and Beni also proposed a validity index (XB) that focused on two properties: compactness and separateness [5]. The most desirable partitioning can be obtained when XB is minimized over all values of c. The term d_{min} indicates the minimum distance between the clusters. However, all of these indexes have focused primarily on the compactness and the variation within cluster. This approach fails to provide a correct representation

TABLE I
CONVENTIONAL CUSTER VALIDITY MEASURES

Validity Measures	Equation
Partition Coefficient [4]	$PC(U;c) = \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^2 / n$
Classification Entropy [4]	$CE(U;c) = -\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}\log_a u_{ij} / n$
Sugeno [6]	$FS(U,V;X) = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m\left(\|x_j - v_i\|^2 - \|v_i - \bar{v}\|^2\right)$ $\bar{v} = \sum_{i}^{n} x_i / n$
Xie-Beni Index [5]	$XB(U,V;X) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2\|v_i - x_j\|^2 / n d_{\min}^2$ $d_{\min} = \min_{i,j}\|v_i - v_j\|$

TABLE II
RELATED WORKS ON DNA MICROARRAY DATA

Authors	Algorithms	Validity Index	Dataset
Yeung et al. [9]	K-means Single-linkage	FOM	Yeast
Bolshakova and Azuaje [10]	K-means SOM	Dunn's based index	Leukemia Lymphoma
Gasch and Eisen [11]	Fuzzy k-means	N/A	Yeast
Dembele and Kastner [12]	Fuzzy c-means	Silhouette index	Yeast Human cancer Serum

of fuzzy partitioning.

C. Related Work

Studies about cluster analysis of the DNA microarray data are summarized in Table II. Yeung analyzed yeast cell-cycle data by k-means and single-linkage algorithm [9]. Bolshakova and Azuaje used SOM and hard k-means algorithm for clustering and Silhouette index for cluster validation [10]. Also, Eisen analyzed yeast cell-cycle data by fuzzy k-means algorithm and k-means algorithm [11]. Dembele and Kastner used fuzzy c-means algorithm to analyze serum and yeast cell-cycle data [12]. Most of the validity indexes used in this prior research was based on the inter-cluster distance or the inner-cluster distance.

III. BAYESIAN VALIDATION METHOD

All the previous indexes including PC, CE, FS and XB focused on only the compactness and the variation within cluster. However, those indexes fail to provide a correct representation of fuzzy partition in the data since the separation is simply computed by considering only the distance between cluster centroids.

$$\lim_{c \to n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m\|x_j - v_i\|^2 = 0 \qquad (1)$$

As shown in (1), if the number of clusters c approaches the number of samples n, the distance between the cluster centroid and a sample becomes 0. Thus, the traditional indexes lose their

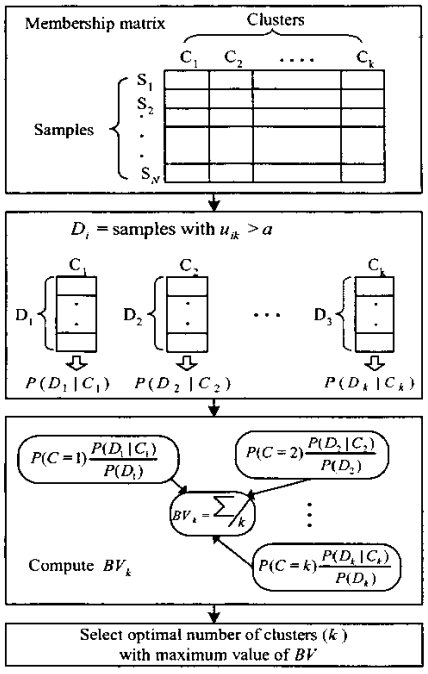


Fig. 2. Bayesian validation method

ability to validate fuzzy partition for large values of c [7]. Bayesian validation method is a probability-based approach, selecting a fuzzy partition with the largest posterior probability given the dataset. It chooses a partition which has maximum posterior probability given the dataset as an optimal cluster partition. Using Bayes theorem, the posterior probability given the *Dataset*=$\{d_1, d_2, \ldots, d_N\}$, could be obtained by multiplication rule and independence rule as follows:

$$P(Cluster \mid Dataset) = \frac{P(Cluster)P(Dataset \mid Cluster)}{P(Dataset)} \tag{2}$$

$$\begin{aligned} P(Cluster \mid Dataset) &= P(Cluster \mid d_1, d_2, \ldots, d_N) \\ &= P(Cluster \mid d_1) \times P(Cluster \mid d_2) \times \ldots \times P(Cluster \mid d_N) \end{aligned} \tag{3}$$

The sum of $P(Cluster|Dataset)$ for all c is calculated using (4) and (5) and this value is defined as Bayesian Value (BV). This value indicates how well the fuzzy partition represents the dataset by the posterior probability. Larger values of BV indicate more appropriate cluster partitions.

$$\begin{aligned} BV &= \frac{\sum_{i=1}^{c} P(C_i \mid D_i)}{C} = \frac{\sum_{i=1}^{c} P(C_i \mid d_{i1}, d_{i2}, \ldots, d_{iN})}{C} \\ &= \frac{\sum_{i=1}^{c} P(C_i \mid d_{i1}) P(C_i \mid d_{i2}) \ldots P(C_i \mid d_{iN})}{C} \\ &= \frac{\sum_{i=1}^{c} \prod_{j=1}^{N_i} P(C_i) P(d_{ij} \mid C_i) / P(d_{ij})}{C}, \\ D_i &= \{d_{ij} \mid u_{ij} > \alpha,\ 1 \le j \le n\},\ N_i = n(D_i) \end{aligned} \tag{4}$$

In Eq (4), d_{ij} is the jth sample that belongs to the ith cluster. $n(D_i)$ is the number of D_i's and we select only a sample that has larger membership value (u_{ij}) than certain threshold α for calculation. Since the fuzzy clustering aims mainly to analyze the samples that belong to multiple classes, evaluating the partition with samples whose membership values are larger than a certain threshold is more appropriate to group samples by fuzzy clustering method. This threshold is defined as α-cut. Since each membership value u_{ij} represents the degree to which data x_i belongs to a certain cluster c, u_{ij} can be substituted for $P(d_{ij}|C_i)$. $P(C_i)$ and $P(d_{ij})$ are calculated as follows:

$$\begin{aligned} P(C_i) &= \sum_{j=1, u_{ij} > \alpha}^{n} u_{ij} \Big/ \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} \\ P(d_{ij}) &= \sum_{i=1}^{c} P(C_i) P(d_{ij}) = \sum_{i=1}^{c} P(C_i) u_{ij} \end{aligned} \tag{5}$$

Figure 2 shows the outline of the proposed method. D_1 includes the samples in cluster C_1 whose membership values are larger than α. Finally, BV is obtained and used to select the optimal fuzzy partition.

The algorithm of Bayesian validation method is as follow:

- Step 1: Compute the membership matrix u_{ij},
- Step 2: Construct D_i by selecting samples ($u_{ij} > \alpha$) in each cluster,
- Step 3: Compute $P(D_j|C_j)$, $P(D_j)$, and $P(C_j)$ of D_i,
- Step 4: Compute Bayesian Score using the calculated values at step 2,
- Step 5: Evaluate the fuzzy partition with the maximum value of BV as optimal one.

IV. Experimental Results

A. Datasets

To show the usefulness of the proposed method, comparisons with four fuzzy cluster validity indexes (PC, CE, FS, XB) for the fuzzy partitions obtained from FCM are conducted on four datasets: Iris, Wine, and Image datasets downloaded from Univ. of Calif., Irvine (UCI) machine learning repository (http://www.ics.uci.edu/~mlearn/MLSummary.html) and SRBCT [15] which is gene expression dataset. The Iris dataset contained 150 samples in 4 dimensional measurement spaces, and consists of two or three clusters because of the substantial overlap of two of the clusters. Wine dataset includes 178 samples in 13 dimensional measurement spaces and has three clusters. Image dataset contains 210 samples in 19 dimensional measurement spaces where seven clusters are considered optimal. The SRBCT consists of four types of cancer (RMS, NB, BL, EWS) and has 63 samples in 96 dimensional measurement spaces [15].

After comparing the proposed method to four conventional validity indexes with the four datasets, *Saccharomyces*

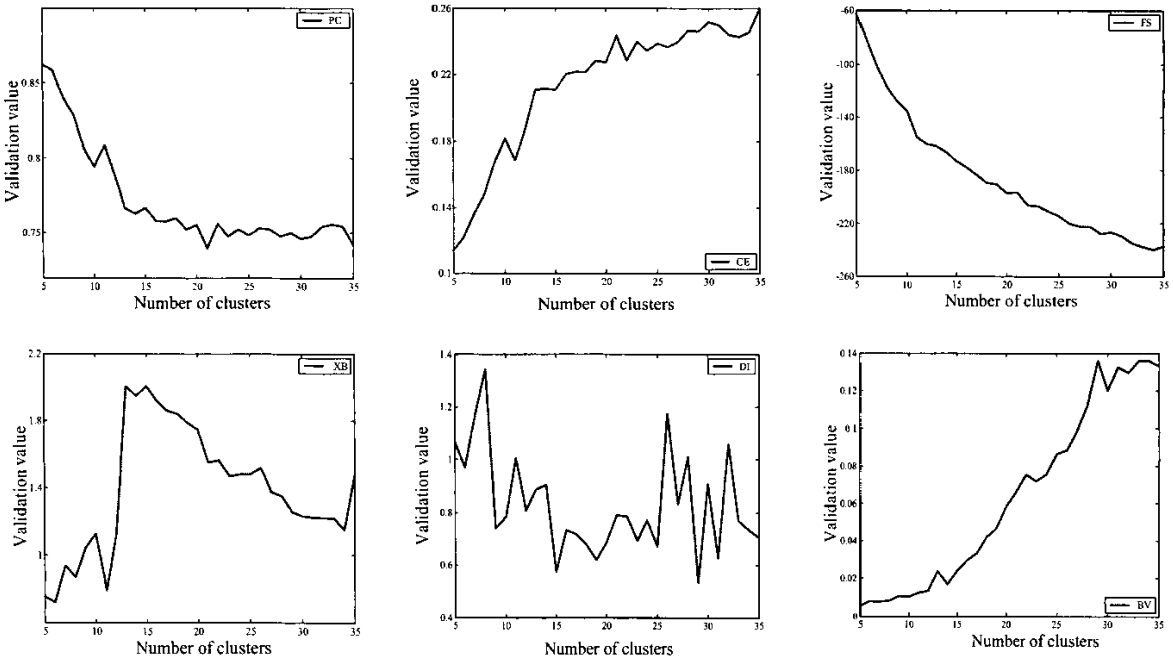


Fig. 3. Preferable values of c for *Saccharomyces cerevisiae* cell-cycle gene expression data by each cluster validity measure

TABLE III
CLUSTER VALIDITY VALUES ON THE WINE DATASET (C=3)

c	PC	CE	FS	XB	DI	BV
2	**0.9358**	**0.0476**	29.3216	0.4516	**1.4520**	0.1583
3	0.9258	0.0564	-1.5084	**0.4312**	1.3314	0.2707
4	0.8814	0.0932	-8.9773	0.8636	0.9450	**0.2773**
5	0.8308	0.1344	-12.293	1.3137	0.7487	0.2556
6	0.8180	0.1494	-15.310	1.5032	0.7254	0.2477
7	0.7964	0.1673	**-18.470**	1.3638	0.6162	0.2160

TABLE IV
CLUSTER VALIDITY VALUES ON THE IMAGE DATASET (C=7)

c	PC	CE	FS	XB	DI	BV
2	0.9468	**0.0379**	35.2037	**0.2617**	**1.8185**	0.2190
3	0.9270	0.0550	-22.327	0.5519	1.2417	0.3846
4	**0.9539**	0.0384	-65.149	0.3935	1.4623	0.4120
5	0.9448	0.0464	-74.462	0.3895	1.0081	0.3918
6	0.9292	0.0599	-85.939	0.7165	0.5850	0.4672
7	0.8980	0.0866	**-91.060**	0.8023	0.5734	**0.5490**

TABLE V
CLUSTER VALIDITY VALUES ON THE IRIS DATASET (C=2 OR 3)

c	PC	CE	FS	XB	DI	BV
2	**0.9916**	**0.0060**	-311.72	**0.0619**	**3.9295**	**0.7512**
3	0.9781	0.0156	-426.90	0.1539	2.4267	0.5825
4	0.9704	0.0226	-459.02	0.2189	1.8279	0.4494
5	0.9569	0.0331	-459.62	0.5045	1.1272	0.3046
6	0.9560	0.0333	-462.48	0.9038	1.3143	0.4712
7	0.9510	0.0366	**-481.40**	0.6820	1.3141	0.4610

cerevisiae cell cycle gene expression data is analyzed with the proposed method. This set contains time-course expression profiles for more than 6000 genes, with 17 time points for each gene taken at 10-min intervals covering nearly two yeast cell cycles (160min). This dataset is very attractive because a large number of genes contained in it are biologically characterized and have been assigned to different phases of the cell cycle. We used the same 421 genes that have been used in previous experiments due to their known utility in clustering [16].

TABLE VI
CLUSTER VALIDITY VALUES ON THE SRBCT DATASET (C=4)

c	PC	CE	FS	XB	DI	BV
2	0.8758	0.0969	164.407	**1.1294**	0.8338	0.1918
3	0.9205	0.0709	86.6529	0.8127	0.9454	0.5612
4	**0.9393**	**0.0616**	27.3224	**0.5657**	**1.1721**	**0.7073**
5	0.9100	0.0850	-0.9891	0.8487	0.7477	0.6731
6	0.8922	0.0977	-22.604	0.7798	0.7405	0.6411
7	0.8989	0.0979	**-34.773**	0.8670	0.6908	0.6852

B. Performance Comparison

All the experiments are repeated six times on each dataset by increasing the α-cut value from 0.1 to 0.6 by 0.1 and the average of six results is used as BV. For other validity measures, we have repeated ten times on each dataset and presented its average value. We have used m=1.2 for the fuzziness parameter which is the same value in previous experiments [12]. Table 3 shows the results of Wine dataset. PC and CE produce the optimal fuzzy partition at c=2, FS at c=7, and XB at c=3, whereas BV yields c=4 as the optimal fuzzy partition. XB was the only method that produced the correct number of clusters (c=3) for wine dataset. Since the difference between the BV at c=3 and at c=4 is very small (0.0066) compared to other margins, it can be said that the proposed method makes the optimal fuzzy partition at c=3 or c=4. DI is a Dunn's index in Table III and we have done extra experiments to compare DI with the proposed method. DI shows similar pattern with PC and CE selecting c=2 for optimal cluster partitions.

Table IV shows the results from experiments with the Image dataset. With the exception of the proposed new method, all other methods arrive at an incorrect optimal number of clusters,

given that the known optimal value is c=7: PC at c=4, CE, XB, DI at c=2, and FS at c=8, respectively. Table V and Table VI show the results of Iris dataset and SRBCT dataset. In Table V, all the methods select c=2 as an optimal fuzzy partition except FS that has the optimal value at c=7. In the case of SRBCT, 4 clusters are known as the optimal number of clusters, and PC, CE, XB, DI, and BV find out the optimal fuzzy partition at c=4, whereas FS finds at c=7.

We found that FS was the most unreliable index since it cannot yield the correct number of clusters for all the four datasets. The worth of PC, CE, XB, and DI tends to decrease with increasing cluster numbers on Wine and Image datasets. BV makes the correct number of clusters except Wine dataset and does not show monotonic decreasing tendency as c increases.

C. *Analysis of Saccharomyces cerevisiae cell cycle gene expression data*

Figure 3 shows the results of all the validation methods including the proposed one, where the x-axis represents the number of clusters and the y-axis represents the evaluation value of each validation method. PC and CE have determined the optimal fuzzy partition at c=5, FS at c=35, XB at c=13, and DI at c=7 respectively. Unlike the other methods, BV leads to the optimal value at c=29. All validity measures show different results and we analyzed biological functions of the cluster partition and its members (genes) that belong to multiple clusters.

We have compared the result of BV which produces the optimal fuzzy partition at c=29 with biological knowledge of yeast cell-cycle data [16]. Yeast cell-cycle data represents expression levels of the genes in each of the five cell cycles (Early G_1 – Late G_1 - S - G_2 - M). Each cell cycle includes the genes that show higher expression levels at that cycle time than other cycle times.

By finding clusters that show high peak point in expression levels at certain time in the cycle, we have assigned the cluster to that cycle. Table VII shows the assigned cluster number and the cycles that they belong to. Clusters that have high expression levels at certain cycle time show low expression level at the other cycle times. Genes assigned between the cycles play a role in regulating the genes that lie in the next cell cycle.

The next step of the analysis is to verify known biological information that the proposed method is indeed able to extract correct information that corresponds to different phases of the yeast cell-cycle data.

Table VIII arranges the genes whose biological functions are known and their cluster number in bracket. Each cycle includes the detailed function groups like DNA replication, biosynthesis, mating pathway and so on. We have confirmed that the results produced by the proposed method are reliable according to the biological knowledge of the genes.

We have chosen special genes whose 1st membership values lie between 0.35 and 0.7, and 2nd membership values are larger than 0.3. These fuzzy genes are belonged to multiple clusters

TABLE VII
ANALYSIS OF CELL CYCLE AND CLUSTERS

Time (×10min)	Cell Cycle	Clusters showing peak expression levels on corresponding cell cycle
0~3	G_1 phase	Cluster5, Cluster6, Cluster4, Cluster24
	G1/S phase	Cluster2, Cluster12, Cluster26, Cluster28
3~5	S phase	Cluster8, Cluster13, Cluster14, Cluster16
	S/G2 phase	Cluster11
5~7	G_2 phase	Cluster13
	G2/M phase	Cluster18
7~9	M phase	Cluster7, Cluster17
	M/G1 phase	Cluster10, Cluster21, Cluster3, Cluster20, Cluster19
9~11	G_1 phase	Cluster5, Cluster6, Cluster4, Cluster24
	G1/S phase	Cluster2, Cluster12, Cluster26, Cluster28
11~13	S phase	Cluster8, Cluster13
	S/G2 phase	Cluster11
13~15	G_2 phase	Cluster0, Cluster13
	G2/M phase	Cluster18
15~17	M phase	Cluster7, Cluster17

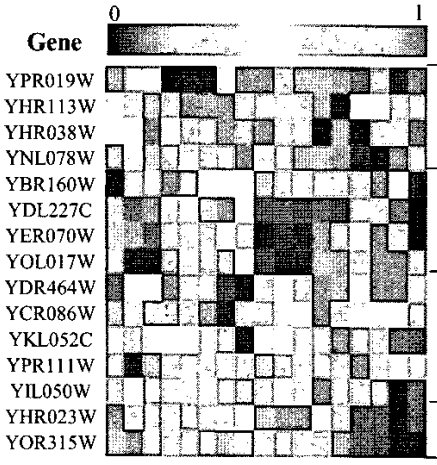
TABLE VIII
ANALYSIS OF CELL CYCLE AND FUNCTIONAL GOUPS WITH GENES

Cell Cycle	Functional Groups	Genes
Early G_1 phase	DNA Replication	YBL023C(10) YEL032W(10) YPR019W(10)
	Mating Pathway	YJL157C(3) YKL185W(3)
	Glycolysis, Respiration	YCR005C(20) YCL040W(20) YLR258W(20)
	Biosynthesis	YIL009W(21) YLL040C(21)
Late G_1 phase	Cell Cycle Regulation	YBR160W(12) YDL127W(12) YGR109C(12) YPR120C(12)
	Chromosome Segregation	YDL003W(26) YFL008W(26) YJL074C(26) YKL042W(26) YMR076C(26) YMR078C(26)
	DNA Replication	YBR278W(24) YKL045W(24) YLR103C(24) YPR018W(24)
S phase	Chromosome Segregation	YDR113C(16) YGR140W(16) YHR172W(16)
	DNA Replication	YBL002W(8) YBL003C(8)
	Miscellaneous	YCR035C(14) YER016W(14) YJR137C(14)
G_2 phase	Directional Growth	YJL099W(11) YJR076C(11)
	DNA Replication	YDR224C(27) YDR225W(27)
M phase	Cell Cycle Regulation	YGL116W(7) YPR119W(7)
	Transcriptional Factor	YDR146C(18) YLR131C(18)
	Directional Growth	YCL037C(17)

and they provide useful information in gene analysis. Figure 4 shows these "fuzzy genes" with their biological descriptions and assigned cluster numbers.

We have classified 4 categories of genes by using the discovered knowledge from Table VII. The genes in cluster 3, cluster 10, cluster 20, and cluster 21 are related to Early G_1 phase. For example, YNL078W belongs to cluster 3 (0.4316) and cluster 19 (0.3139) simultaneously. Actually cluster 3 is related to mating pathway and cluster 19 is related to glycolysis respiration in the same Early G_1 cycle. YNL078W plays multiple roles in Early G_1 cycle. YPR019W, YHR113W, and YHR038W are also fuzzy genes that have multiple functions in cell's life.

Other fuzzy genes in second category (cluster 12, cluster 24, and cluster 26) are related to Late G_1 phase. Gene like



Gene	Gene description	Clusters
YPR019W	member of the Cdc46p/Mcm2p/Mcm3p family	10, 3
YHR113W	similarity to vacuolar aminopeptidase Ape1p	20, 21
YHR038W	killed in mutagen	20, 21
YNL078W	hypothetical protein	3, 19, 25
YBR160W	"g1,g2" CDC28 cyclin-dependent kinase	6, 12, 24
YDL227C	homothallic switching endonuclease	5, 12, 26
YER070W	ribonucleoside-diphosphate reductase, large subunit	12, 24, 26
YOL017W	similarity to YFR013w	12, 26
YDR464W	regulates spliceosome components	9 ,11
YCR086W	hypothetical protein	9 ,11
YKL052C	hypothetical protein	9 ,11
YPR111W	kinase involved in late nuclear division	9 ,13
YIL050W	cyclin like protein interacting with Pho85p	11, 13
YHR023W	myosin-1 isoform heavy chain	7, 18
YOR315W	hypothetical protein	7, 18

Fig. 4. Analysis of fuzzy genes (gene description and cluster number)

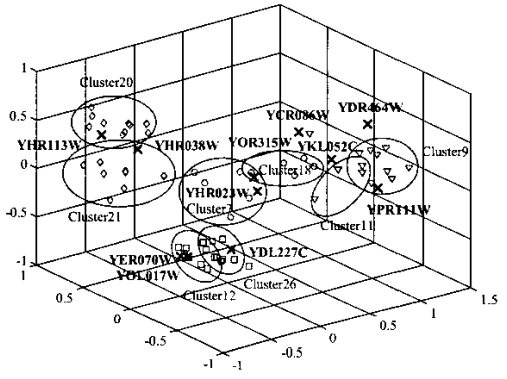


Fig. 5. 3D plot display of fuzzy genes

YBR160W, belongs to cluster 12 (0.3982) and cluster 6 (0.3464) simultaneously. Cluster 12 is related to cell cycle regulation and cluster 6 is related to chromosome segregation. Cluster 9, cluster 11, and cluster 13 are related to G_2 phase and cluster 7 and cluster 18 are related to M phase in cell cycle rotation as shown in Figure 4.

We have plotted the fuzzy genes which are analyzed in Figure 4 and their relations are shown in Figure 5. We have used principal component analysis (PCA) to reduce the dimensions of the genes to three and displayed all genes in 3-dimensional space. Fuzzy genes are represented as black cross (X) and rests of genes are represented as different shapes (diamonds, rectangle, triangle, and circle) according to their belonged clusters.

As shown in Figure 5, it is clear to see that YHR113W and YHR038W are located between cluster 20 and cluster 21 which are related to Early G_1 phase. Also YHR023 and YOR315W which belong to cluster 7 and cluster 18, are located between these two clusters. These two clusters are related to M phase in cell cycle rotation. Between the other clusters related to Late G_1 phase and G_2 phase, there exist fuzzy genes, providing useful information for further research about unknown genes. Fuzzy genes which have multiple functional families do not have clear boundaries and belong to multiple clusters simultaneously.

V. Concluding Remarks

In this paper, a new cluster validation method for the fuzzy partition has been proposed. Bayesian validation method evaluates the fuzzy partition by the posterior probability for the dataset at hand. The best fuzzy partition is obtained by finding the maximum BV with respect to the number of clusters.

We have established α-cut as threshold in computing the value of BV to evaluate various kinds of cluster partitions. The performance of the proposed method was tested on the four well-known datasets (Wine, Image, Iris, and SRBCT dataset) demonstrating its usefulness with fuzzy c-means algorithm. Also, we have analyzed the *Saccharomyces cerevisiae* cell cycle gene expression data with the proposed method. To confirm the superiority of the proposed method, we have compared predicted clusters to biological knowledge and found a useful congruence, which suggests that this approach may be of high value in the analysis of gene expression data.

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc Natl Acad Sci*, USA, vol. 96, no.12, pp. 6745-6750, 1999.

[2] A. P. Gasch and M. B. Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, vol. 3, no. 11, research 0059.1-0059.22. 2002.

[3] F. Höppner, F. Klawonn, R. Kruse and T. Runkler, *Fuzzy Cluster Analysis*, Wiley, 2000.

[4] J. C. Bezdeck, Cluster validity with fuzzy sets. *J. Cybernit.*, vol. 3, pp. 58-72, 1974.

[5] X. L. Xie and G. Beni, A validity measure for fuzzy clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841-846, 1991.

[6] Y. Fukuyama and M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method. *Proceedings of 5th Fuzzy Systems Symposium*, pp. 247-250, 1989.

[7] D. W. Kim, K. H. Lee, D. and H. Lee, Fuzzy cluster validation index based on inter-cluster proximity. *Pattern Recognition Letters*, vol. 24, pp. 2561-2574, 2003.
[8] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.
[9] K. Y. Yeung, et al., "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, no. 4, pp. 309-318, 2001.
[10] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data," *SIGPRO*, vol. 21, no. 82, pp. 1-9. 2002.
[11] A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biology*, vol. 3, no. 11, research 0059.1-0059.22, 2002.
[12] D. Dembele and P. Kastner,"Fuzzy c-means method for clustering microarray data," *Bioinformatics*, vol. 19, no. 8, pp. 973-980, 2003.
[13] N. R. Pal and J. C. Bezdeck, "On cluster validity for the fuzzy c-means model," *IEEE Trans. Fuzzy Systems*, vol. 3, no. 3, pp. 370-379, 1995.
[14] M. R. Rezaee, B. P. F. Lelieveldt and J. H. C. Reiber, "A new cluster validity index for the fuzzy c-means," *Pattern Recognition Letters*, vol. 19, pp. 237-246, 1998.
[15] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673-679, 2001.
[16] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65-73, 1998.