IDENTIFICATION OF GENES CONSISTENTLY CO-EXPRESSED IN MULTIPLE MICROARRAY DATASETS BY A GENOME-WIDE BI-COPAM APPROACH

Basel Abu-Jamous[1], Rui Fa[1], David J. Roberts[2], Asoke K. Nandi[1,3]

[1]Department of Electronic and Computer Engineering, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK;
[2]National Health Service Blood and Transplant, The University of Oxford, Oxford, UK;
[3]Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland
{Basel.AbuJamous, Rui.Fa}@brunel.ac.uk, david.roberts@ndcls.ox.ac.uk, asoke.nandi@brunel.ac.uk

ABSTRACT

Many methods have been proposed to identify informative subsets of genes in microarray studies in order to focus the research. For instance, the recently proposed binarization of consensus partition matrices (Bi-CoPaM) method has, amongst its various features, the ability to generate tight clusters of genes while leaving many genes unassigned from all clusters. We propose exploiting this particular feature by applying the Bi-CoPaM over genome-wide microarray data from multiple datasets to generate more clusters than required. Then, these clusters are tightened so that most of their genes are left unassigned from all clusters, and most of the clusters are left totally empty. The tightened clusters, which are still not empty, include those genes that are consistently co-expressed in multiple datasets when examined by various clustering methods. An example of this is demonstrated in this paper for cyclic and acyclic genes as well as for genes that are highly expressed and that are not. Thus, the results of our proposed approach cannot be reproduced by other methods of genes' periodicity identification or by other methods of clustering.

Index Terms— Bi-CoPaM, co-expressed genes, genome-wide clustering, microarray data analysis

1. INTRODUCTION

Advances in microarray technology have enabled the measurement of expressions of a huge number of genes simultaneously. Most microarray experiments consider measuring the expression values of the entire genome (all of the genes) of a specific organism over multiple time-points, biological developmental stages, different types of tissues, or different conditions. The resulting data structure is a gene-sample or a gene-time-point matrix whose rows represent genes and whose columns represent samples or time-points [1].

Although these microarray datasets include thousands of genes which represent the entire genome, most of the genes in the dataset are expected to be irrelevant in a specific case study [1]. So, many studies resort to filtering the genes in order to keep a small subset of informative genes only. This small subset can be exposed to further analysis, such as clustering. Gene clustering mines for co-expressed genes, i.e. genes with similar expression profiles. Although co-expression does not necessitate that these genes have the same biological function, it indicates that they may do; so, clustering lights up some promising regions of further biological experimentation [2].

The gene-filtering step has been carried out by different approaches, which in many cases depend on the nature of the dataset under consideration. For example, many datasets have been generated by measuring the expression of the budding yeast ~6000 genes over two yeast cell-cycles [3,4,5,6]; the subset of genes which show periodic expression profiles over these cell-cycles was chosen as the subset of further analysis by many studies [3,4,5,6,7]. Many other studies choose *differentially expressed* genes as the subset of genes for further concentration. The problem of identifying differentially expressed genes has been tackled in various approaches, but, in general, differentially expressed genes are those whose expression profiles show high expression and/or high fold-changes over different time-points or conditions [1,2].

The gene clustering step has been approached by many non-ensemble methods, e.g. k-means [8], hierarchical clustering (HC) [7,9], self-organizing maps (SOMs) [10,11], self-organizing oscillator networks (SOON) [12], and ensemble methods, e.g. relabeling and voting [13], co-association matrix [14], hypergraph methods [15], and the recently proposed binarization of consensus partition matrices (Bi-CoPaM) [16,17]. In contrast to most of the other methods, Bi-CoPaM provides the unique ability to generate tunable wide overlapping clusters with many multiply assigned genes as well as tunable tight clusters while leaving many genes unassigned from all clusters [16,17].

Bi-CoPaM clusters the expression profiles of the same set of genes from multiple microarray datasets and/or by using various clustering methods. The generated partitions are then combined into a single fuzzy consensus partition matrix (CoPaM) which is then binarized by one of the six Bi-CoPaM binarization techniques to provide the final binary partition [16,17].

In this study, we propose a novel way of using the Bi-CoPaM method by (i) applying it over the entire genome rather than a filtered subset, (ii) clustering the same set of genes from multiple microarray datasets, (iii) using a large number of clusters (K), and (iv) exploiting the tightness feature in order to obtain a small number of non-empty clusters which include a few genes that bare the characteristic of being consistently co-expressed in multiple datasets. This approach of using the Bi-CoPaM embeds the filtering step within the clustering step in a tunable way.

2. METHODS

2.1. Bi-CoPaM

The Bi-CoPaM method consists of four main steps [16,17]:

The project (Ref. NIHR-RP-PG-0310-1004-AN) is supported by National Institute for Health Research (NIHR), UK. A.K. Nandi would like to thank TEKES for their award of the Finland Distinguished Professorship.

1. Generation of many partitions for the same set of genes by applying various clustering methods over the expression profiles of these genes from multiple microarray datasets.
2. Relabeling the generated partitions such that each cluster from one partition is matched with its corresponding cluster from every other partition.
3. Fuzzy consensus partition matrix (CoPaM) generation by element-by-element averaging of the relabeled partitions.
4. Binarization of the CoPaM by one or more of the six tunable binarization techniques proposed in [16].

The six binarization techniques scrutinize the CoPaM in different ways to produce binary partitions with different features. Our concentration in this study is mainly on the *difference threshold binarization (DTB)* technique.

Conventional binarization assigns each gene to the cluster in which it has its maximum fuzzy membership; this generates complementary clusters in which each gene is exclusively assigned to one and only one cluster. The DTB technique imposes a stricter policy; it assigns this gene to that maximum-membership cluster only if the closest cluster competing on this gene has a fuzzy membership value which is lower than the maximum by at least the value of the parameter (δ). Otherwise, the gene is unassigned from all of the clusters accordingly. Tighter clusters with more unassigned genes are obtained when δ is increased until it reaches one. When its value is one, only genes that have been consensually assigned to the same clusters by all of the single partitions are preserved; all of the other genes are left unassigned.

2.2. Mean Squared Error (MSE) Metric

The mean squared error (MSE) metric has been used by many studies to evaluate the quality of the generated clusters so that comparisons between different methods can be performed [18,19]. We adopt the MSE metric for evaluating the generated clusters.

Because the total number of genes assigned to the clusters by Bi-CoPaM at any specific tightness level is variable, we use a normalized MSE measure to be *per gene*. The $MSE_{cluster}$ metric which quantifies the total MSE for the k^{th} cluster is defined as:

$$MSE_{cluster(k)} = \frac{1}{N \cdot M_k} \sum_{x_i \in C_k} \|x_i - z_k\|^2, \qquad (1)$$

where N is the number of dimensions (time-points) in the dataset, M_k is the number of genes in the k^{th} cluster, C_k is the set of genetic expression profiles $\{x_i\}$ for the genes in the k^{th} cluster, and z_k is the mean expression profile for the genes in the k^{th} cluster.

If multiple datasets were used for clustering, genes profiles and the clusters centroids will vary from one dataset to another for the same partition. In this case, the MSE metric can be calculated multiple times for each dataset and then averaged over them.

3. DATASETS

We consider six yeast cell-cycle microarray datasets in our study. Each of these datasets measures the genetic expression of the entire yeast genome (~6000 genes) over about two complete yeast cell-cycles. The details of the datasets are listed in Table 1.

The first column shows the unique name which is used hereinafter to refer to each of these datasets. The names of the first four datasets, as commonly used in the literature, were derived from the synchronization methods used in producing them. The last two datasets were generated by Orlando and colleagues [6]. Orlando and colleagues generated four datasets, two of which are used in our study; they are those which they labeled as "wild-type replicate 1" and "wild-type replicate 2". Accordingly, we refer to these two datasets respectively as orl-wt1 and orl-wt2. See Table 1.

Table 1. Budding yeast cell-cycle microarray datasets

Name	Year	Genes	Time points	Spacing (min)	Missing values allowed	Ref.
Cdc28	1998	6223	17	10	1 / 17	[3]
Alpha	1998	6178	18	7	1 / 18	[4]
Alpha-30	2006	6266	25	5	1 / 25	[5]
Alpha-38	2006	6266	25	5	1 / 25	[5]
Orl-wt1	2008	5667	15	16	0/15	[6]
Orl-wt2	2008	5667	15	16	0/15	[6]

The second to the seventh columns of Table 1 respectively show the year in which the dataset was published, total number of genes, number of time-points, time spacing between each two consecutive time-points, number of allowed missing values in any single gene not to be filtered out, and reference.

There are 4910 genes that are common to the six datasets and do not have more missing values than allowed. These genes were the ones considered in our Bi-CoPaM clustering experiments.

4. EXPERIMENTAL SETUP

The considered 4910 genes were clustered into 25 clusters by k-means with Kauffman initialization (KA) [8], self-organizing maps (SOMs) with bubble neighborhood and five-by-five grid, and hierarchical clustering (HC) with Ward linkage. This was applied over their profiles from all of the six datasets. Generated partitions were relabeled by a min-min approach and then combined into a single CoPaM matrix. Because alpha-30 and alpha-38 were generated by technical replication, each of them was given half weight in generating the CoPaM. Similarly, the Orl-wt1 and Orl-wt2 datasets are biological replicates and were given half weight in combination. The CoPaM was binarized by the DTB technique with δ values ranging from zero to one and then analyzed by MSE.

Prior to clustering, the one-color datasets cdc28, orl-wt1 and orl-wt2 were normalized by quantile normalization [20] then by making each gene's expression profile zero-mean and unity standard deviation. The log-ratios of the two-color datasets alpha, alpha-30 and alpha-38 were zero-centered by subtracting genes' log-ratios' mean values [21].

5. RESULTS

The numbers of genes in the tightest 15 clusters at all of the δ values are shown in Table 2. Clusters were ordered based on their tightness, such that those clusters that preserve at least seven genes up to higher δ values are considered tighter. When many clusters preserve at least seven genes up to the same δ value, they are ordered based on the number of genes they include at that level.

The interpretation of the inclusion of a set of genes within one cluster at the $\delta = 1$ case is that these genes were assigned to the same cluster by all of the partitions. In other words, they are co-expressed, i.e. they show the same expression profile, in all of the considered datasets and when clustered by all of the adopted clustering methods. On the other hand, the genes included at lower levels of δ were assigned to the same cluster *significantly more* than to any other cluster. The definition of '*significantly more*' is tunable and controlled by δ. The fact that these genes were not assigned to the same cluster by *some* partitions can be because of the differences between the clustering methods or because these genes show less co-expression in some of the considered datasets.

Table 2. Numbers of genes included in each of the 15 tightest clusters at all of the considered δ values

Tightness	δ	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
Complementary	0.0	123	489	270	568	686	541	530	395	360	290	226	259	172	141	179
	0.1	59	269	143	235	241	203	174	107	86	67	52	53	48	24	38
	0.2	44	196	98	179	120	96	69	35	26	23	23	21	15	9	9
	0.3	35	135	57	77	42	29	19	4	4	3	6	3	5	2	0
	0.4	31	99	42	41	24	12	8	2	1	1	2	1	0	0	0
	0.5	26	70	28	21	10	5	2	0	0	0	2	1	0	0	0
	0.6	24	41	19	2	0	0	0	0	0	0	0	0	0	0	0
	0.7	22	16	6	0	0	0	0	0	0	0	0	0	0	0	0
	0.8	19	4	4	0	0	0	0	0	0	0	0	0	0	0	0
	0.9	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Tightest	1.0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0

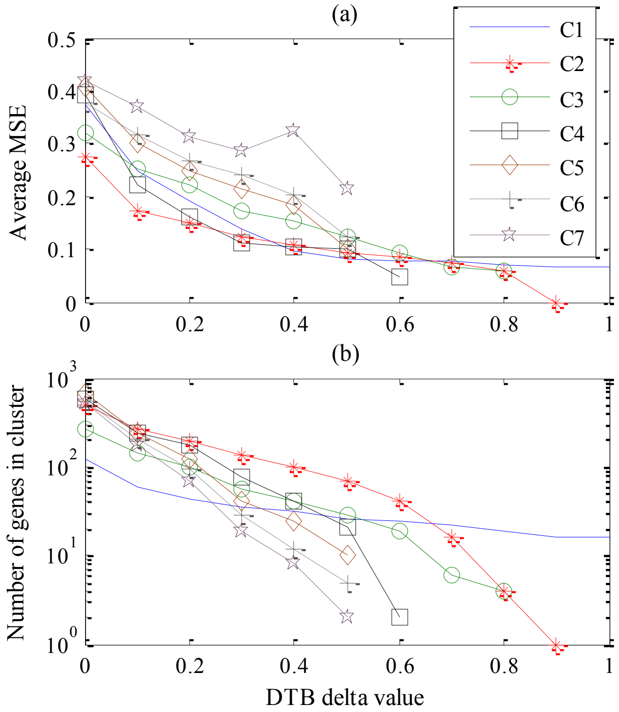

Figure 1. (a) Average MSE values and (b) number of genes included for the tightest seven clusters.

Each gene is definitely assigned to at least one cluster at DTB $\delta = 0$. So, considering the clusters' contents at this level of tightness for further analysis means considering all of the genes whether they were relevant or not. When tightness increases, many clusters shrink quickly, and many genes are left unassigned. These quickly lost clusters and genes are considered relatively irrelevant because they do not show consistent co-expression at different conditions. The contents of the clusters which survive up to high levels of tightness are more informative and worth being considered for further analysis.

5.1. MSE Analysis

The MSE values for each of tightest seven clusters were calculated at all of the DTB δ values. Each of these MSE values was calculated based on the six datasets and then averaged and plotted in Figure 1 (a). Figure 1 (b) shows the numbers of genes included in each of these seven clusters at all of the DTB δ values. Missing points in both plots represent empty clusters.

It can be seen in this Figure that the MSE values generally support the Bi-CoPaM results in that the clusters that lose genes quicker tend to show higher (worse) MSE values. Moreover, after the fifth cluster, clusters start to show significantly higher values of MSE with significantly lower numbers of genes included. Thus, we consider the first five clusters C1 to C5 as significant and discard the rest of the clusters for our current study.

The importance of considering both plots (a) and (b) in Figure 1 in tandem is because the differences in clusters' sizes at different δ values make them incomparable by merely using the MSE. For example, C2 at $\delta = 0.9$ has the perfect MSE value of zero, but this is because it includes only a single gene. Another example is C4 whose MSE value shows a small increase from $\delta = 0.5$ to $\delta = 0.3$ while the number of genes included in it increases significantly from 21 to 77. In other words, the best tightness level of any cluster can be chosen such that it includes as many genes as possible while maintaining a reasonable value of MSE.

Accordingly, we intuitively choose the best tightness level for each of the five clusters C1 to C5. We call the chosen cases as the *cores* of these clusters, see Table 3. Note that the total number of genes included in all of these five clusters' cores is 187 genes out of possible 4910 genes originally included in the study.

The expression profiles for all of the genes included in each of these five clusters' cores from the two datasets cdc28 and orl-wt1 are plotted in Figure 2. Although the expression profiles in the other four datasets are not identical to these two, these plotted profiles are representative and serve well in demonstrating that these core genes are consistently co-expressed in multiple datasets.

Table 3. Cores of clusters C1 to C5

	C1	C2	C3	C4	C5
δ value	0.4	0.6	0.5	0.3	0.5
Number of genes	31	41	28	77	10
Average MSE	0.097	0.086	0.122	0.113	0.101

5.2. Comparison to the Literature – Periodicity

Many studies which considered one or more of these six yeast cell-cycle datasets in their analysis have started by identifying those genes that are cell-cycle regulated, i.e. that show periodic expression profiles over cell-cycles. The studies in [3], [4], [5] and [6] have respectively identified 384, 800, 1000 and 1271 genes as periodic, and they performed their further analysis over these subsets of genes. Some studies considered analyzing these subsets of genes or an intersection between some of them [6,7,17].

The numbers of genes included in each of the five clusters cores C1 to C5 and considered periodic by each of the aforementioned four studies are listed in Table 4. The numbers in

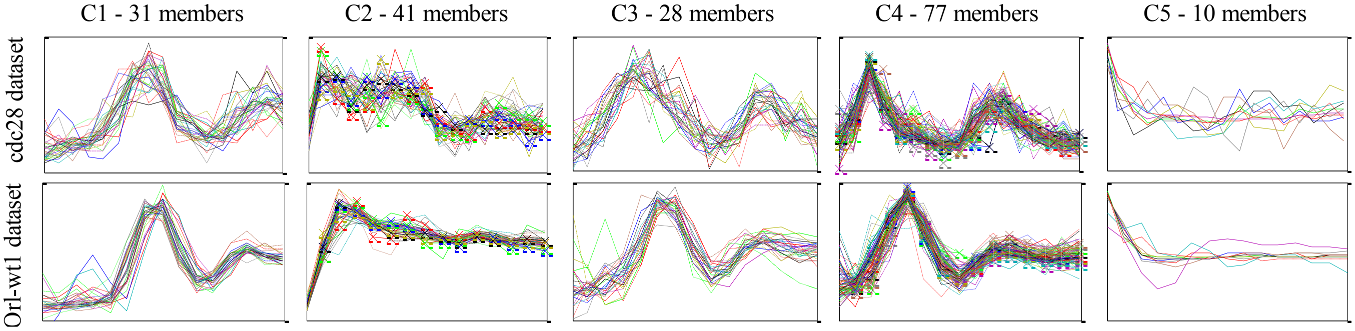

Figure 2. Normalized expression profiles for the genes in the cores of the clusters C1 to C5 from the datasets cdc28 and orl-wt1.

this Table and the profiles in Figure 2 lead to the same conclusion that the clusters C1, C3 and C4 are cell-cycle regulated (cyclic) while C2 and C5 are not. All of the five clusters from the Bi-CoPaM method, including the ones detected by previous cell-cycle studies and the ones that were not, follow the same criterion: these sets of genes are co-expressed with each other in six different datasets when examined by various clustering methods.

Table 4. Periodic genes in the cores of the clusters C1 to C5

No.	Ref.	No. of genes	C1	C2	C3	C4	C5
1	[3]	384	27	0	12	51	0
2	[4]	800	31	0	21	63	1
3	[5]	1000	31	4	28	64	3
4	[6]	1271	31	0	25	65	1

5.3. Expression Levels Analysis

Based on the cdc28 dataset, the quartiles of the 4910 genes peak values were calculated and listed in the first row of Table 5. The last row of the Table shows the numbers of genes from the 187 Bi-CoPaM core clusters' genes whose peak expression levels are between each two consecutive quartiles. It can be seen in this Table that significant numbers of these core genes are within each of these four quartile-intervals. This indicates that expression-level-based gene-filtering prior to Bi-CoPaM clustering cannot be used to focus such analysis.

Table 5. Core genes distribution over expression levels

Quartile →	**min**		**Q1**		**Q2**		**Q3**		**max**
Quartile expression value	9		205		393		773		14107
Core genes in the interval		22		50		67		48	

6. DISCUSSION AND CONCLUSIONS

We have proposed a novel approach of using the binarization of consensus partition matrices (Bi-CoPaM) method over genome-wide microarray datasets in order to identify a small subset of genes organized within clusters to allow for more focused biological analysis. Abu-Jamous and colleagues have proposed the Bi-CoPaM method as an ensemble clustering method which allows for generating wide overlapping clusters and/or tight clusters with many genes being unassigned from all clusters [16]. They also proposed applying this method over the expression profiles of the similar set of genes from multiple datasets [16,17]. That approach was useful to mine for groups of genes that are co-expressed (have the same expression profile) in multiple datasets consistently. In their first study, they tested Bi-CoPaM over a synthetic dataset of cyclic genes' profiles which belong to five clusters [16], and in their second study over 500 cyclic yeast genes from five different yeast cell-cycle datasets [17].

Our approach in this paper mines the entire set of genes within the available microarrays, i.e. the entire genome, for genes that are consistently co-expressed in multiple datasets. This is different from those previous studies in that – (i) no a priori filtering of genes must be carried out before applying the Bi-CoPaM, i.e. all informative and noisy genes are included in the study initially, and (ii) it is able to identify genes that are consistently co-expressed in multiple datasets whether they were cyclic or not (e.g. core clusters C2 and C5 in our analysis), and whether their expression levels are differentially high or not.

Our approach of applying Bi-CoPaM not only serves as a way of finding useful clusters; it also serves as a way to identify focused informative genes subsets from the entire set of available genes. Many studies aimed at identifying such informative subsets of genes in different ways, e.g. Cho [3], Spellman [4], Pramila [5], as well as Orlando [6] and their colleagues, have identified different subsets of yeast genes by considering their periodicity in the yeast cell-cycle. Other research instances, as reviewed by Roberts [1], as well as by some others [2], have considered high or differential-expression between different time-points or conditions as the criterion by which the subsets of interesting genes are identified. Although this is expected to minimize the number of irrelevant genes in the initial study, many genes that are neither periodic nor highly- or differentially-expressed might still be identified by our approach as consistently co-expressed with the same sets of genes in multiple datasets (e.g. C2 and C5). Thus, none of these methods can find the same subsets of genes which we have found by applying the Bi-CoPaM over genome-wide datasets. Moreover, filtering the complete set of genes by any of the aforementioned methods prior to applying the Bi-CoPaM can result in loss of information which might still be useful for such Bi-CoPaM analysis. Anyway, the filtering step is implicitly embedded within this Bi-CoPaM-based approach.

To conclude, our proposed approach is to apply the Bi-CoPaM method over genome-wide expression data from multiple datasets to generate many clusters, and then to tighten them such that only a few tight clusters of genes are obtained. The central feature which our approach looks for is that the genes included in any of these tight clusters are tightly and consistently co-expressed in multiple datasets which might have been generated in different labs, in different years, and under different conditions. Some other common methods of gene filtering, such as periodically or highly expressed genes identification, mine for different features other than what our approach mines for. Thus, our approach can find novel and important results that are orthogonal to what these other methods can find, yet cannot be found by any of them.

7. REFERENCES

[1] P. C. Roberts, "Gene expression microarray data analysis demystified," *Biotechnol Annu Rev*, vol. 14, pp. 29-61, 2008.

[2] P. C. Boutros and A. B. Okey, "Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data," *Briefings in Bioinformatics*, vol. 6, pp. 331-343, 2005.

[3] R. J. Cho *et al*., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65–73, 1998.

[4] P. T. Spellman *et al*., "Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273–3297, 1998.

[5] T. Pramila *et al*., "The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phasegap in the transcriptional circuitry of the cell cycle," *Genes and Development*, vol. 20, pp. 2266–2278, 2006.

[6] D. A. Orlando *et al*., "Global control of cell-cycle transcription by coupled CDK and network oscillators," *Nature*, vol. 453, pp. 944-947, 2008.

[7] E. J. Cooke *et al*., "Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements," *BMC Bioinformatics*, vol. 12, 2011.

[8] J. M. Pena, J. A. Lozano, and P. Larranaga, "An empirical comparison of four initialization methods for the K-Means algorithm," *Pattern Recognition Letters*, vol. 20, no. 10, pp. 1027-1040, 1999.

[9] M. B. Eisen *et al*., "Cluster analysis and display of genome-wide expression patterns," in *Proc. Natl. Acad. Sci.*, vol. 95, 1998, pp. 14863-14868.

[10] X. Xiao *et al*., "Gene clustering using self-organizing maps and particle swarm optimization," in *IEEE Parallel and Distributed Processing Symposium Proceedings*, Indianapolis, 2003, pp. 154-163.

[11] D. Dikicioglu *et al*., "How yeast re-programmes its transcriptional profile in response to different nutrient impulses," *BMC Systems Biology*, vol. 5, pp. 148:163, 2011.

[12] S. A. Salem, L. B. Jack, and A. K. Nandi, "Investigation of self-organizing oscillator networks for use in clustering microarray data," *IEEE Trans. Nanobioscience*, vol. 7, no. 1, pp. 65-79, 2008.

[13] H. G. Ayad and M. S. Kamel, "On voting-based consensus of cluster ensembles," *Pattern Recognition*, vol. 43, pp. 1943-1953, 2010.

[14] A. Fred and A. K. Jain, "Data clustering using evidence accumulation," in *Proceedings of the Sixteenth International Conference on Pattern Recognition (ICPR)*, vol. 4, 2002, pp. 276-280.

[15] Z. Yu, H. S. Wong, and H. Wang, "Graph-based consensus clustering for class discovery from gene expression data," *Bioinformatics*, vol. 23, no. 21, pp. 2888-2896, 2007.

[16] B. Abu-Jamous *et al*., "Paradigm of Tunable Clustering using Binarization of Consensus Partition Matrices (Bi-CoPaM) for Gene Discovery," *PLOS ONE*, vol. 8, no. 2, 2013a, doi: 10.1371/journal.pone.0056432.

[17] B. Abu-Jamous *et al*., "Yeast gene CMR1/YDL156W is consistently co-expressed with genes participating in DNA-metabolic processes in a variety of stringent clustering experiments," *Journal of the Royal Society Interface*, vol. 10, no. 81, 2013b, doi: 10.1098/rsif.2012.0990.

[18] Y. K. Lam and P. W. Tsang, "eXploratory K-Means: A new simple and efficient algorithm for gene clustering," *Applied Soft Computing*, vol. 12, pp. 1149–1157, 2012.

[19] Z. Zhu *et al*., "Memetic clustering based on particle swarm optimizer and k-means," in *2012 IEEE Congress on Evolutionary Computation (CEC)*, Brisbane, Australia, 2012.

[20] B. Bolstad *et al*., "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, pp. 185-193, 2003.

[21] J. Quackenbush, "Microarray data normalization and transformation," *Nature Genetics*, vol. 32, pp. 496–501, 2002.