M-N SCATTER PLOTS TECHNIQUE FOR EVALUATING VARYING-SIZE CLUSTERS AND SETTING THE PARAMETERS OF BI-COPAM AND UNCLES METHODS

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

The recently proposed UNCLES method has the ability to unify clustering results from multiple datasets under different types of external specifications. It can also tunably tighten the results such that many objects are unassigned from all of the clusters to obtain few tight clusters. Despite the success of this method, setting its parameters, such as the number of clusters (K) and the tuning parameters δ and (δ^+, δ^-), has never been automated. As its clusters vary in size, they cannot be validated by the existing validation indices. In this study we present a technique of validation based on our proposed M-N scatter plots. This technique has the ability to provide better fitness values for the clusters which include more objects while preserving their tightness. This well suits the nature of the results of UNCLES. We have applied this technique to a set of bacterial microarray datasets as well as a set of English vowels datasets. Our results demonstrate the success of the M-N plots in selecting the best few clusters out of a pool of clusters generated under varying K, δ, and (δ^+, δ^-) values. Our results also show that the best few clusters can be originated from different partitions, which shows the power of our technique in evaluating individual clusters rather than whole partitions. Finally, despite proposing this technique within the context of the UNCLES framework, it is readily applicable to other clustering results, especially when the parameters are not confidently predefined.

Index Terms— M-N plots, UNCLES, Bi-CoPaM, clustering validation, gene expression

1. INTRODUCTION

Unsupervised clustering methods have been widely used in a variety of applications, including the identification of the subsets of co-expressed genes, i.e. the genes whose genetic expression profiles are highly correlated. Generic clustering methods, such as k-means [1], self-organising maps (SOMs) [2], and hierarchical clustering (HC) [3], have been commonly used by the bioinformatics community to achieve the aforementioned objective [1,2,3]. Other clustering methods have been proposed within the specific context of gene clustering to tackle some aspects that emerge within this context, such as the recently proposed method, *the binarisation of consensus partition matrices (Bi-CoPaM)* [4,5,6], and its more recent generalisation, *the unification of clustering results from multiple datasets using external specifications (UNCLES)* [7]. Despite being proposed within the context of gene clustering, those methods can be applied to any other application with analogous aspects.

Rather than identifying the subsets of co-expressed genes in a single gene expression dataset (e.g. microarray dataset), the Bi-CoPaM method was proposed to allow for the identification of the subsets of genes *consistently* co-expressed over multiple datasets [4]. From a biological point of view, the expression profiles for a subset of genes might be found correlated for reasons other than that they are contributing to the same biological process [8], even though, consistent co-expression (correlation) of the same subset of genes over multiple datasets provides a stronger hypothesis that those genes contribute to the same biological process [8]. Moreover, the Bi-CoPaM method, in contrast to other clustering methods, allows any single gene to have any of the three eventualities, to be exclusively assigned to a single cluster, to be simultaneously assigned to multiple clusters, or not to be assigned to any of the clusters [4]. This tackles the biological fact that any gene can participate in one biological process, multiple biological processes, or to be irrelevant to the processes under investigation [4]. This leads to the generation of clusters with varying levels of tightness from being wide and overlapping, to being complementary, to being tight and focused with many genes left without assignment [4].

In ICASSP 2013 [6], Abu-Jamous and colleagues proposed the application of the Bi-CoPaM method to genome-wide scale datasets, i.e. datasets with all of the genes included without pre-filtering [6]. They demonstrated the usefulness of the tunable tightness feature which allows the Bi-CoPaM to filter out the genes which are not consistently co-expressed over the given datasets, and that the other commonly used pre-filtering techniques would pre-eliminate many genes that are preserved by the Bi-CoPaM [6]. In a more recent study [7], they proposed the UNCLES method as a generalisation of the Bi-CoPaM. UNCLES unifies the clustering results from multiple datasets using different types of external specifications [7]. The first type is equivalent to Bi-CoPaM, which unifies the results to identify the subsets of genes consistently co-expressed in *all* of the provided datasets. The second type unifies those results in order to identify the subsets of genes consistently co-expressed in one subset of datasets while being poorly consistently co-expressed in another subset of datasets [7]. This issue has been mentioned in the literature without being fully tackled by an unsupervised method [8,9].

Some important aspects in those methods were stated to be unsolved yet [4]. The first aspect is setting the number of clusters (K) and the tuning parameter δ used to tighten or widen the clusters; these used to be either predetermined or set semi-manually [4,5,6]. The second aspect is an appropriate clustering validation technique which suits the tunable nature of these methods' results. Although

This article summarises independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Grant Reference Number RP-PG-0310-1004). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. A.K. Nandi would like to thank TEKES for their award of the Finland Distinguished Professorship.

those aspects were identified and stated while proposing the Bi-CoPaM [4], they were not solved even after proposing its generalisation, UNCLES [7]. In fact new parameters were introduced with UNCLES, namely the pair (δ^+, δ^-), which also need to be properly set [7].

In this study, we propose a novel clustering evaluation technique, *M-N plots*, which suits the varying tightness nature of both types of the UNCLES' method (the first type is equivalent to the Bi-CoPaM), and can be used to resolve the issues of identifying the most appropriate number of clusters (K) as well as the other tuning parameters. We also demonstrate that this method can be applied to applications other than gene clustering.

2. METHODS

2.1. Bi-CoPaM

Given a set of datasets, e.g. gene expression datasets generated under various biological conditions and contexts, and a set of clustering methods (e.g. k-means, SOMS, etc.), the Bi-CoPaM method is applied through the following four main steps [4,6]:

1) Each of the clustering methods is applied to each of the datasets to provide a pool of individual partitions.
2) The partitions are relabelled such that each cluster from any partition is mapped to its corresponding cluster from all of the other partitions.
3) Relabelled partitions are combined to produce a single fuzzy consensus partition matrix (CoPaM).
4) The fuzzy CoPaM is binarised by one of six binarisation techniques to produce a final binary CoPaM.

Here we concentrate on one binarisation technique, which is the difference threshold binarisation (DTB) [4,6]. DTB assigns any object (e.g. gene) to the cluster in which it has its maximum fuzzy membership only if its closest competing cluster is at least far by the value of the parameter δ; this object is not assigned to any of the clusters otherwise. When δ is zero, each object is assigned to the cluster in which it has its higher membership unconditionally. While δ is increased, more objects are unassigned from all of the clusters, and therefore tighter and more focused clusters are produced. The tightest clusters are produced when δ reaches unity, the case at which an object is assigned to a cluster only if its membership in that cluster is one with zero membership in all of the other clusters.

2.2. UNCLES

The UNCLES method [7] has two types, A and B, where type A is equivalent to the Bi-CoPaM method. In terms of gene expression analysis, type B tackles the question of identifying the subsets of genes consistently co-expressed in a subset of datasets, S^+, while being poorly co-expressed in another subset of datasets, S^-. UNCLES is applied through the following steps:

1) The Bi-CoPaM is applied to each of the two subsets of datasets, S^+ and S^-, independently and respectively with the DTB δ values δ^+ and δ^-, i.e. the parameter pair (δ^+, δ^-).
2) The genes which are preserved in the results of the Bi-CoPaM applied to S^- at δ^-, i.e. the genes which are consistently co-expressed in S^-, are excluded from the clusters generated by the Bi-CoPaM applied to the S^+ datasets at δ^+.

The resulting clusters include the subsets genes deemed as consistently co-expressed in S^+ at δ^+ while being deemed as consistently poorly co-expressed in S^- at δ^-. Therefore, δ^+ controls how strongly consistently co-expressed the genes need to be in S^+ in order to be included, and δ^- controls how strongly co-expressed the genes need to be in S^- in order to be excluded.

2.3. Proposed M-N Scatter Plots

The mean squared error (MSE) metric has been used in many studies to evaluate the quality of the generated clusters [6,7,10,11]. The $MSE_{cluster}$ metric, which quantifies the average MSE for the k^{th} cluster, is defined as:

$$MSE_{cluster(k)} = \frac{1}{D \cdot N_k} \sum_{x_i \in C_k} \|x_i - z_k\|^2, \quad (1)$$

where D is the number of dimensions (time-points) in the dataset, N_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.

MSE is biased towards smaller clusters which include fewer genes, where the trivial case of a cluster which includes a single gene is provided the best MSE value of zero [6]. This renders the MSE metric, as it is, inappropriate for evaluating the results of both types of UNCLES because of their tunable tightness / size nature. Therefore, we propose evaluating the clusters based on both the number of genes included in them (N) and an MSE-based metric (M). The objective is to minimise the dissimilarity of the genes within the cluster, i.e. to minimise the MSE-based metric value, while maximising the number of objects included.

The MSE-based metric (M) for UNCLES type A is the average of all MSE values calculated based on each of the given datasets. For type B, it is defined as the signed difference between the average of MSE values based on the S^+ subset of datasets, and the average of MSE values based on the S^- subset of datasets.

The M-N scatter plot is a plot on which the clusters are scattered, whose horizontal axis is the MSE-based metric (M), and whose vertical axis is the logarithm of the number of genes included in the cluster (N). When the axes limits are set to the limits of scattered clusters and normalised, the top left corner is considered as the best point at which N is maximised and M is minimised.

The clusters which are scattered on an M-N plot can be from all of the UNCLES experiments which consider various values of K, δ, and/or (δ^+, δ^-). The best of all of those clusters is that which is closest in Euclidean distance to the top left corner of the M-N plot. After selecting this best cluster, all of the other clusters which share at least one object with it are removed from the plot; that is because they are considered as other versions of the same cluster but with degraded quality. The process of selecting the best cluster and removing the ones similar to it is repeated many times in order to select the best distinct clusters from that pool of all available clusters. Once there is a big leap in the distances from the top left corner between two consecutively selected clusters, the process terminates.

3. DATA, EXPERIMENTS, AND RESULTS

We have conducted two experiments to test our proposed technique. The first experiment involves five Escherichia coli (E. coli) bacteria microarray datasets while the second involves a non-genetic English vowels data tensor split into five datasets.

3.1. Real Bacterial Microarray Datasets

E. coli bacteria, which is a widely used model organism for microbiological studies, is commonly found in the lower intestine of warm-blooded organisms. Five microarray datasets have been considered in this experiments and they are listed in Table 1. The first column shows the labels we shall use hereinafter to refer to those datasets, such that the letters 'P' and 'N' respectively indicate *positive* and *negative* datasets with reference to their treatment under UNCLES type B. The second to the fifth columns respectively show

Table 1. Escherichia coli bacteria microarray datasets

ID	GEO acc.	D	Description	Ref.
P1	GSE9923	10	Perturbations including temperature	[12]
P2	GSE10159	18	Peptidoglycan stress	[13]
N1	GSE20374	9	Varying cofactors NADH and ATP	[14]
N2	GSE34275	12	Varying glycerol	[15]
N3	GSE34631	6	Varying glycol and glycerol	[16]

the GEO accession identifier, the number of dimensions (samples) (D), a short description, and the reference for each of the five datasets. The K-12 strain (genetic variant) was used in all of the five datasets and therefore we have considered the entire K-12 strain's genome, consisting of 4,345 genes, in our analysis.

UNCLES type A was applied to the five datasets and UNCLES type B was applied to them while considering the datasets P1 and P2 as the positive subset of datasets (S^+) and N1, N2, and N3 as the negative subset of datasets (S^-). The number of clusters (K) was varied for both types to include all values from two to twenty in addition to 25, 30, and 50. The parameters δ for type A, and δ^+ and δ^- for type B were varied from zero to unity with steps of 0.1.

3.1.1. Results

UNCLES types A and B respectively have generated 3,454 and 37,994 individual clusters while varying the values of K, δ and (δ^+, δ^-), 1,167 and 23,260 of them were non-empty, respectively. The clusters from type A are scattered in the top-left sub-plot in Figure 1 with three different symbols. The blue solid circle represents the closest cluster to the top-left corner, and therefore it is the cluster selected by the proposed technique as the best cluster. The red stars represent all of the other clusters that share at least one gene with that selected cluster, and the black squares are the rest of the clusters. This first selected cluster is referred to with the label 'A1'. We have found 571 clusters that share genes with A1. The clusters represented by the blue solid circle and all of the red stars are removed from this M-N plot to produce the M-N plot 'A2'. The same procedure has been applied to this reduced M-N plot version to select the second best cluster that is distinct from A1. The first three iterations of this process, which select A1, A2, and A3, are shown in Figure 1. This Figure also shows three similar iterations applied to the results of type B to select B1, B2, and B3.

Table 2 shows some details about the top four selected clusters for both types A and B. The Table has two almost identical subsets of columns for the two types A and B. The first column in each subset of columns shows the identifiers with which the clusters have been labelled. The second to the sixth columns respectively show the numbers of genes, MSE-based metric values, the K values under which the clusters have been generated, δ or (δ^+, δ^-) values, and the distances (d) from the top-left corners of the normalised M-N plots. It can be seen in this Table that the top clusters are selected from various K values and under various δ or (δ^+, δ^-) values.

Figure 2 (a) and (b) show the distances from the M-N plots top-left corners of the top eight clusters selected for UNCLES types A and B respectively. This Figure can aid the researcher's decision on how many top clusters should be selected before termination. For example, there is an obvious leap in the distances after the second cluster in each of the two types, and therefore, one might consider those first clusters and discard the remaining ones. One might consider the third and the fourth clusters as well with a lower level of confidence. Indeed this is an application specific decision.

Table 2. Top four E. coli clusters generated by UNCLES types A and B and selected by M-N plots

ID	N*	M	K	δ	d	ID	N*	M	K	(δ^+, δ^-)	d
A1	291	0.53	2	0.8	0.70	B1	258	-0.27	6	(0.4,0.8)	0.45
A2	27	0.32	4	0.7	0.71	B2	309	-0.22	5	(0.1,0.8)	0.47
A3	55	0.46	6	0.6	0.75	B3	98	-0.34	9	(0.4,0.8)	0.51
A4	43	0.45	9	0.4	0.76	B4	135	-0.26	7	(0.6,0.7)	0.52

* 'N' in the M-N plots is the base-10 logarithm of the number of genes while in this Table it is the absolute number of genes.

3.1.2. Biological Reasoning

The Gene Ontology (GO) initiative associates genes with the terms of the biological processes in which they participate. This is actively updated based new findings in biology and related sciences. We have performed GO term analysis over the top clusters generated by both UNCLES types A and B and selected by the M-N plots technique.

The cluster A1 was found to be enriched with many terms related to translation, which is the process that produces proteins in the cells. Some of those terms are 'translation' (p-value 1.7×10^{-4}), 'tRNA modification' (p.v. 1.4×10^{-4}), 'tRNA methylation' (p.v. 3.9×10^{-4}), and 'polyamine transport' (p.v. 1.4×10^{-4}). Translation, which is at the core of the *central dogma of molecular biology*, is a global property of cells under various conditions and across species from bacteria and fungi to plants and animals [17]. Thus, discovering A1 in our results meets this biological fact because UNCLES type A identifies the subsets of genes consistently co-expressed in *all* of the considered datasets generated under various conditions.

The top cluster in UNCLES type B 'B1' has been found highly enriched with many processes needed under stress and perturbation conditions such as 'DNA repair' (p.v. 2.4×10^{-4}) and 'response to heat' (p.v. 1.7×10^{-3}). Recall from Table 1 that the *positive* subset of datasets (P1 and P2), in contrast to the *negative* subset of datasets (N1 to N3), is related to different types of perturbations and stresses. Again, the top cluster discovered by our UNCLES type B method well meets this biological reasoning.

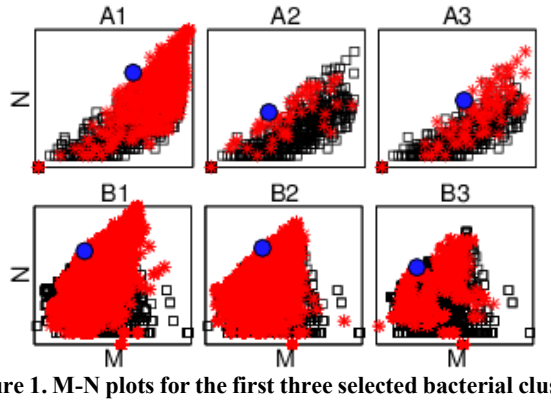

Figure 1. M-N plots for the first three selected bacterial clusters by types A and B of UNCLES. M is the MSE-related metric and N is the logarithm of the number of genes.

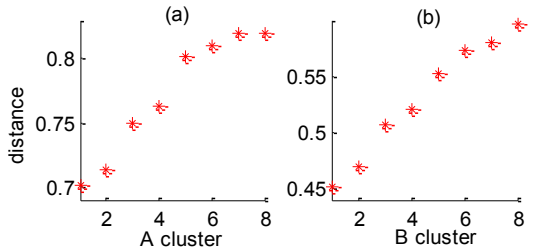

Figure 2. Distances from the top-left corner of the M-N plots for the top eight clusters in UNCLES (a) type A and (b) type B.

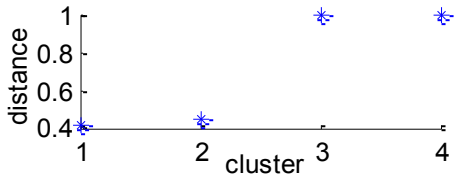

Figure 3. M-N plots' distances for the top four clusters of English vowels.

3.1.3. Comparison with Other Validation Indices

We have compared our M-N scatter plots approach with two clustering validation indices, namely the MSE metric [10] and the silhouette [18]. The silhouette metric is an object-based validation metric which quantifies the fitness of any object (e.g. gene) in the cluster to which it is assigned. We have applied this metric to the partitions generated by both types of UNCLES and then considered that the fitness of any cluster is the average of the fitness values of its members. We have not considered the validation indices which only provide fitness values for whole partitions rather than individual clusters because such comparison would not be applicable in the context of our study.

The top 150 UNCLES type A clusters, as sorted by the MSE and the silhouette metrics, include no more than four and 34 genes respectively. This is out of 1,167 non-empty clusters. As for type B, the top 1,000 clusters, out of 23,260 non-empty ones, include no more than 56 and 33 genes when sorted by the MSE and the silhouette metrics respectively. This clearly shows the tendency of these two metrics towards praising very small clusters.

Moreover, the top clusters selected by M-N plots have been found not to top the sorted lists based on the MSE and the silhouette metrics. For example, A1 and A2 were overtaken by 42% and 21% of the non-empty clusters respectively based on the MSE metric. A2 was overtaken by 13% of them based on silhouette which silhouette failed to evaluate A1 because its partition happened to have only one non-empty cluster. Similarly, B1 and B2 were overtaken by 22% and 38% based on MSE, and by 45% and 55% based on the silhouette, respectively. This indicates that such metrics cannot be used to assess the results of the methods like UNCLES which generate tunable clusters ranging in sizes from empty to extremely large.

3.2. English Vowels Real Datasets

Harshman and colleagues measured the positions of thirteen points across the tongue of five native English speakers while uttering each of ten different vowels [19,20]. The ten vowels are the ones encapsulated between the letters *h* and *d* in the ten words "heed, hid, hayed, head, had, hod, hawed, hoed, hood, and who'd." The data from any of the speakers represent one single dataset and therefore there are five datasets; the ten words are the objects to be clustered while the thirteen distance values are their features.

UNCLES type A was applied to these five datasets to identify the subsets of words (vowels) at which the tongue's position is consistently similar across multiple speakers. The number of clusters (K) was varied from two to ten and the δ values were varied from zero to unity with steps of 0.1. Therefore, the results contained 594 individual clusters, 335 of which found to be not empty. Successive iterations of the M-N plots technique have been applied to this pool of clusters to select the top ones of them. No clusters were left in that pool after the fourth iteration and therefore the process terminated.

Table 3. English vowels clusters' members

C1	C2	C3	C4
hod, hawed, hoed, hood	heed, hid, hayed, head	who'd	had

Figure 3 shows the distances from the top-left corners of the M-N plots for those clusters which are labelled C1 to C4 respectively. The first two clusters are significantly better than the last two. Table 3 shows the membership of the ten words (vowels) in each of the four clusters. These results can be intuitively validated by noticing that the first two clusters clearly include two distinct groups of vowels. Moreover, the M-N plots ranking for the clusters has elevated the clusters which include more vowels while preserving some tightness because this better meets the original question of which subsets of vowels are consistently correlated in terms of tongue's geometry across various speakers.

4. DISCUSSION AND CONCLUSIONS

We have presented a novel validation technique based on the M-N scatter plots, which can resolve the issue of setting the parameters K, δ, and (δ^+, δ^-) for the UNCLES method.

Despite the previously demonstrated success of the UNCLES framework across various studies, it has always been considered with a predefined number of clusters (K), and with semi-manually selected δ and (δ^+, δ^-) values, and its results have always been mainly validated by biological reasoning rather than numerical validation [4,5,6,7]. Furthermore, Abu-Jamous and colleagues have explicitly stated that the unconventional nature of the results of the Bi-CoPaM (UNCLES type A) requires designing an applicable validation technique as well as a technique to identify the optimum parameters' values [4]. As shown in this study, those previously unresolved issues have been successfully tackled by our M-N scatter plots.

The clusters generated by UNCLES are tunable such that the same cluster can have various versions from being extremely small to extremely large when produced under various δ and (δ^+, δ^-) values [4,6,7]. Varying the number of clusters (K) as well would generate similar, split, or combined versions of clusters [4,6,7]. It has been shown in this study that the best clusters do not tend to belong to the same partition, i.e. to the same clustering result generated under the same K, δ, and (δ^+, δ^-) values (Table 2). Therefore, clustering validation techniques which provide a single fitness value for each whole partition, such as DI [21], CH [22], and GI [23], are not applicable to this problem in hand. On the other hand, the MSE metric [10] provides fitness values for individual clusters, and the silhouette [18] index provides fitness values for each of the objects (genes) within the clusters. The average of the values from the later one can be used as fitness values for individual clusters. As can be seen in our results, such indices have the tendency to give better values for smaller clusters and are not applicable to the nature of UNCLES' results.

Although we have proposed the M-N scatter plots technique in the context of the UNCLES framework, it is applicable to the broader area of applications in which clustering produces varying sized clusters. Moreover, the idea of selecting the best few distinct clusters out of a pool of clusters stimulates bypassing the boarders of the partitions to dissolve all of the clusters generated under various sets of parameters into such a pool. This can be applied to other clustering methods which involve parameter setting and even to a pool of clusters generated by various methods. Additionally, we have demonstrated, for the first time, the applicability of the UNCLES method to non-biological applications which widens the horizon of its possible applications.

In conclusion, we have shown the unique ability of our proposed M-N scatter plots technique in resolving the UNCLES methods' issue of cluster validation and parameter setting, as well as the applicability of the UNCLES method to a wider range of applications. We have also elucidated the potential usefulness of the M-N plots technique in other than the context of UNCLES.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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, 2013, doi: 10.1371/journal.pone.0056432.

[5] 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, 2013, doi: 10.1098/rsif.2012.0990.

[6] B. Abu-Jamous *et al*., "Identification of genes consistently co-expressed in multiple microarrays by a genome-wide approach," in *The Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 1172-1176.

[7] B. Abu-Jamous *et al*., "Method for the identification of the subsets of genes specifically consistently co-expressed in a set of datasets," in *Proceedings of the 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP-2013)*, Southampton, UK, 2013.

[8] R. Nilsson *et al*., "Discovery of Genes Essential for Heme Biosynthesis through Large-Scale Gene Expression Analysis," *Cell Metabolism*, vol. 10, pp. 119-130, 2009.

[9] C. H. Wade, M. A. Umbarger, and M. A. McAlear, "The budding yeast rRNA and ribosome biosynthesis (RRB) regulon contains over 200 genes," *Yeast*, vol. 23, pp. 293–306, 2006.

[10] 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.

[11] 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.

[12] J. Lee *et al*., "Indole cell signaling occurs primarily at low temperatures in Escherichia coli," *The ISME Journal*, vol. 2, pp. 1007–1023, 2008, doi: 10.1038/ismej.2008.54.

[13] M. E. Laubacher and S. E. Ades, "The Rcs phosphorelay is a cell envelope stress response activated by peptidoglycan stress and contributes to intrinsic antibiotic resistance," *Journal of Bacteriology*, vol. 190, no. 6, pp. 2065–2074, 2008, doi: 10.1128/JB.01740-07.

[14] A. K. Holm *et al*., "Metabolic and transcriptional response to cofactor perturbations in Escherichia coli," *The Journal of Biological Chemistry*, vol. 285, no. 23, pp. 17498–17506, 2010, doi: 10.1074/jbc.M109.095570.

[15] K. Arunasri *et al*., "Effect of simulated microgravity on E. coli K12 MG1655 growth and gene expression," *PLOS ONE*, vol. 8, no. 3, p. e57860, 2013, doi: 10.1371/journal.pone.0057860.

[16] H. Nam *et al*., "Network context and selection in the evolution to enzyme specificity," *Science*, vol. 337, no. 6098, pp. 1101-1104, 2012, doi: 10.1126/science.1216861.

[17] V. Piras, M. Tomita, and K. Selvarajoo, "Is central dogma a global property of cellular information flow?," *Frontiers in Physiology*, vol. 3, no. 439, 2012, doi: 10.3389/fphys.2012.00439.

[18] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 187, doi: 10.1016/0377-0427(87)90125-7.

[19] R. Harshman, P. Ladefoged, and L. Goldstein, "Factor analysis of tongue shapes," *The Journal of the Acoustic Society of America*, vol. 62, no. 3, pp. 693-713, 1977.

[20] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, "Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 26, no. 2, pp. 295-327, 2005, doi: 10.1137/S089547980139786X.

[21] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Cybernetics and Systems*, vol. 3, no. 3, pp. 32–57, 1973, doi: 10.1080/01969727308546046.

[22] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics - Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[23] B. SY. Lam and H. Yan, "Assessment of microarray data clustering results based on a new geometrical index for cluster validity," *Soft Computing*, vol. 11, no. 4, pp. 341–348, 2007, doi: 10.1007/s00500-006-0087-1.