BI-COPAM ENSEMBLE CLUSTERING APPLICATION TO FIVE ESCHERICHIA COLI BACTERIAL DATASETS

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, UB8 3PH, Greater London, United Kingdom
[2]National Health Service Blood and Transplant, The University of Oxford, John Radcliffe Hospital, Oxford, OX3 9UB, United Kingdom
[3]Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland
email: {basel.abujamous, rui.fa}@brunel.ac.uk, david.roberts@ndcls.ox.ac.uk, asoke.nandi@brunel.ac.uk

ABSTRACT

Bi-CoPaM ensemble clustering has the ability to mine a set of microarray datasets collectively to identify the subsets of genes consistently co-expressed in all of them. It also has the capability of considering the entire gene set without pre-filtering as it implicitly filters out less interesting genes. While it showed success in revealing new insights into the biology of yeast, it has never been applied to bacteria. In this study, we apply Bi-CoPaM to five bacterial datasets, identifying two clusters of genes as the most consistently co-expressed. Strikingly, their average profiles are consistently negatively correlated in most of the datasets. Thus, we hypothesise that they are regulated by a common biological machinery, and that their genes with unknown biological processes may be participating in the same processes in which most of their genes known to participate. Additionally, our results demonstrate the applicability of Bi-CoPaM to a wide range of species.

Index Terms— Bi-CoPaM, microarray data analysis, gene clustering, *Escherichia coli* bacteria

1. INTRODUCTION

Recently, there has been an increasing trend in developing computational methods which collectively analyse multiple high-throughput biological datasets to obtain consensus results and conclusions [1, 2]. Gene expression analysis, including gene clustering, is not an exception [3, 4].

We have recently developed an unconventional paradigm of gene clustering through the proposal of the *binarisation of consensus partition matrices (Bi-CoPaM)* method [5]. In a tunable manner, this method 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. Regarding the resulting clusters, they can therefore be complementary, wide and overlapping, or tight and focused with many genes left without being assigned to any of them [5].

Further enhancements took place with regard to the approach in which the method is applied. Rather than filtering the thousands of genes within the genome of a species prior to clustering, the Bi-CoPaM can be fed the entire genome while exploiting its capability of tightening clusters in order to implicitly filter the less interesting genes within the process of clustering [6].

In a separate study, the Bi-CoPaM was applied to yeast datasets, which revealed important findings about the poorly understood gene CMR1 [7]. Although it was stated that the Bi-CoPaM can be applied to other species in order to participate in gene discovery studies at a wider scope, that has not been realised yet.

In this study, we apply the Bi-CoPaM to a set of five *Escherichia coli (E. coli)* bacterial gene expression microarray datasets generated under different biological conditions. Bacteria differ greatly from yeast in that the former belongs to the less developed *prokaryotic* branch of species while the later belongs to the more developed *eukaryotic* branch of species. To illustrate the huge gap between the two species, note that the eukaryotic branch also includes animals and plants. Thus bacteria and yeast are quite distant in their biological nature. We aim at verifying the biological validity of the Bi-CoPaM method when applied to bacterial datasets while trying to plot a number of testable biological hypotheses.

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. Asoke K. Nandi would like to thank TEKES for their award of the Finland Distinguished Professorship.

2. BI-COPAM

Given a number of datasets and individual clustering methods (e.g. k-means [8], hierarchical clustering [9], etc.), the binarisation of consensus partition matrices (Bi-CoPaM) method is applied through the following four main steps [5]:

- Partition generation: each individual clustering method is applied separately to each of the given datasets to generate a pool of partitions. The number of clusters should be fixed in all of those partitions.
- Relabelling: due to the fact that clustering is unsupervised, there are no labels readily associated with the generated clusters. Thus, the clusters of each of the partitions are rearranged such that the i^{th} cluster of all of the rearranged partitions are matched with each other. A min-min approach is considered for this step [6].
- Fuzzy consensus partition matrix (CoPaM) generation: The relabelled partition matrices are averaged in an element-by-element manner to produce a single fuzzy consensus partition matrix (CoPaM), in which the fuzzy membership $u \in [0, 1]$ of any of the genes in any of the clusters represents the fraction of partitions which has assigned that particular gene to this particular cluster.
- Binarisation: The fuzzy CoPaM is binarised to produce a binary consensus partition matrix.

Six binarisation techniques were proposed in [5] and extended in [10], but we consider one of them in this study, namely the *difference threshold binarisation (DTB)* technique. In DTB, a gene is assigned to the cluster in which it has its maximum fuzzy membership value if, and only if, its membership value in the closest competitive cluster is less than that of the maximum by at least the value of the tuning parameter δ. The gene is otherwise left without being assigned to any of the clusters.

The value of δ ranges between zero and unity, inclusively. When δ is zero, each gene is assigned to the cluster in which it has its maximum fuzzy membership value. Therefore, every gene will be assigned to a single cluster, and the generated clusters will be complementary. Slight overlaps may occur though, when a gene belongs to more than one top cluster exactly with the same fuzzy membership value. At the other extreme, when δ is unity, a gene is assigned to a cluster if, and only if, all of the partitions have assigned to it consensually. In this case, the tightest clusters are obtained, and many of them might be totally empty. Taken together, increasing the value of δ tightens the clusters and leaves more genes unassigned.

3. DATASETS & EXPERIMENTAL SETUP

Table 1 lists the five *E. coli* bacterial datasets used in this study. The first column shows the unique identifiers with which we will hereinafter refer to the datasets. The second column shows the unique National Centre for Biotechnology Information (NCBI) accession number with which one can access those datasets in the NCBI online data depository. The third to the fifth columns show the number of samples (N), a brief description, and a reference for each of the datasets, respectively.

ID	Acc. No.	N	Description	Ref.
A	GSE9923	10	Indole signalling at low temperatures	[11]
B	GSE10159	9	Treatment with with cefsulodin and mecillinam	[12]
C	GSE20374	3	Response to cofactor perturbations	[13]
D	GSE34275	6	Growth in presence and absence of glycerol	[14]
E	GSE37026	4	Treatment with colicin	[15]

Table 1. *E. coli* bacterial datasets

All of the five datasets were generated by using the same microarray chip, namely the Affymetrix E. coli Genome 2.0 Array, and include the expression profiles for 3,956 genes constituting the genome of *E. coli*.

We have applied the Bi-CoPaM method to those five datasets while considering a K value (number of clusters) of three. The adopted individual clustering methods are k-means with the deterministic Kaufmann's initialisation [8, 16], hierarchical clustering with Ward's linkage [9, 17], and self-organising maps (SOMs) with a bubble neighbourhood [18]. The δ values considered for the DTB binarisation technique range from zero to unity with a step size of 0.1. The objective is to identify the subsets of genes which are consistently co-expressed in all of those different datasets.

4. RESULTS

The numbers of genes included in each of the three clusters, respectively labelled as C1, C2, and C3, at all of the considered δ values are listed in Table 2. It can be clearly seen in this Table that while increasing the δ value, the number of genes in the clusters decreases. Before giving the clusters their labels and listing them in Table 2, they had been ordered based on the number of genes they preserve at the tightest level when δ is equal to unity.

δ	C1		C2		C3	
	Genes	**MSE**	**Genes**	**MSE**	**Genes**	**MSE**
0.0	2076	0.70	1735	0.72	460	0.74
0.1	1520	0.66	1209	0.68	193	0.70
0.2	1208	0.63	864	0.64	97	0.66
0.3	885	0.59	599	0.59	33	0.60
0.4	565	0.52	377	0.55	11	0.49
0.5	378	0.48	234	0.49	2	0.04
0.6	283	0.45	149	0.44	1	0.00
0.7	120	0.32	57	0.35	1	0.00
0.8	61	0.27	20	0.32	0	-
0.9	21	0.25	3	0.19	0	-
1.0	21	0.25	3	0.19	0	-

Table 2. Number of genes and the mean squared error (MSE) average values for each of the three clusters generated by the Bi-CoPaM method at all of the considered δ values.

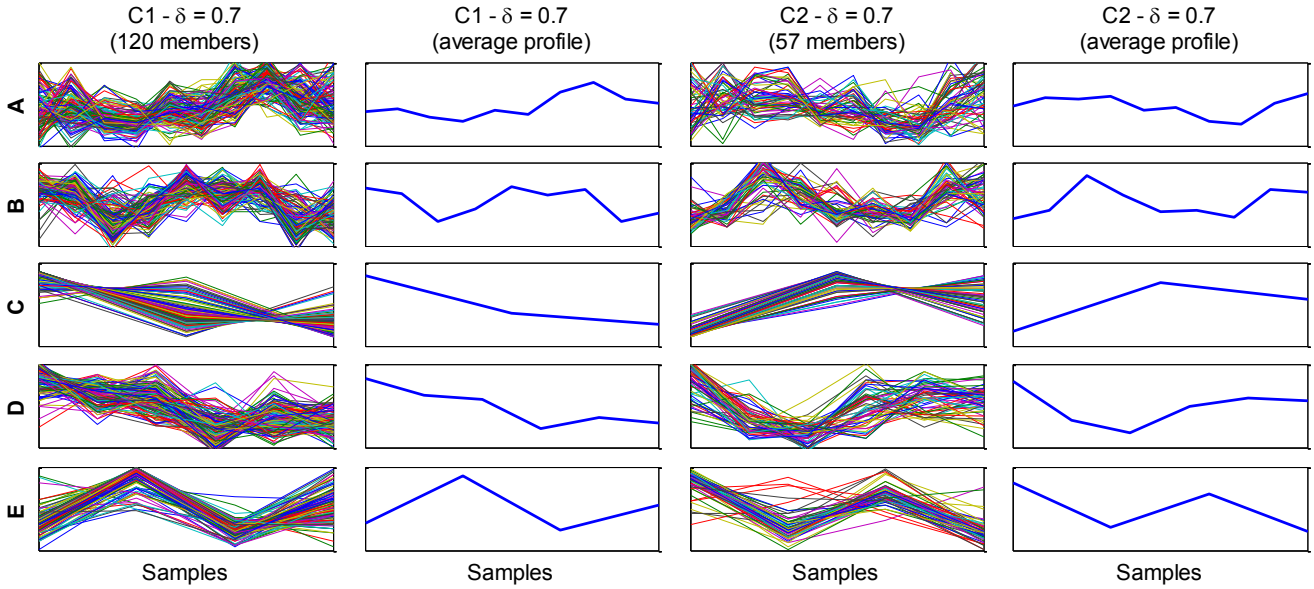

Figure 1. Individual profiles (first and third columns) and average profiles (second and fourth columns) of the genes included in the clusters C1 and C2 at δ = 0.7 in all of the five datasets. The five rows in this grid of sub-plots represent the five datasets A to E.

The third cluster loses all, or most, of its genes at a relatively low δ value. For example, at δ = 0.5, it includes only a couple of genes, which are totally lost when δ reaches the value of 0.8. On the other hand, C1 preserves a considerable amount of genes even when δ reaches unity.

4.1. Mean squared error (MSE) analysis

To examine the quality of the clusters at different δ values, we calculated average mean square error (MSE) values for each of the clusters at all of the δ values. The MSE metric measures the tightness of a cluster by giving it lower values when its genes' profiles are better correlated. The mathematical formulation of this metric is shown in equation (1):

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

where k is the number of the cluster being examined, N is the number of samples or time-points in the dataset, M_k is the number of genes in the cluster, C_k is the set of vectors $\{x_i\}$ representing the gene expression profiles for the genes included in the cluster (k), and z_k is a vector representing the average expression profile for the genes in that cluster.

The MSE values are listed alongside the numbers of genes in Table 2. As seen in this Table, the MSE values for the clusters at higher δ values are smaller, the observation which indicates that they are tighter clusters and correlates with what is expected. Moreover, the three clusters differ largely in terms of their MSE values at similarly sized cases. For instance, the clusters C1, C2, and C3, at the respective δ values of 0.7, 0.6, and 0.2, include the comparative numbers of genes of 120, 149, and 97 genes, respectively, while having the respective distant average MSE values of 0.32, 0.44, and 0.66. This illustrates that the cluster C1 has higher quality than the other two clusters because it preserves larger numbers of genes while maintaining relatively smaller values of MSE. The cluster C2 is not far from C1, but the cluster C3 is significantly distance from both C1 and C2. With this analysis, we can filter out the cluster C3 from our further analysis, and we choose the clusters C1 and C2 at δ = 0.7 for further analysis.

4.2. Expression Profile Analysis

Figure 1 shows the profiles of the genes included in the clusters C1 and C2 at δ = 0.7 in all of the five datasets A to E. The first and the third columns of the Figure show the profiles of the individual genes within those clusters while the second and the fourth columns show their average expression profiles.

The most interesting observation in this Figure is that, in almost all of the datasets, the average profiles of the two clusters are reciprocal, i.e. highly negatively correlated. To evaluate this quantitatively, Pearson's correlation was calculated between the average profiles of the clusters C1 and C2, and as control, the same was calculated for the average profiles of the cluster pairs C1 – C3 and C2 – C3. The results are shown in Figure 2.

It is clearly shown in Figure 2 that there is a strong negative correlation (ρ < -0.7) between the clusters C2 and C3 in four out of five datasets, namely by excluding the dataset (D). On the other hand, there is no similar pattern between any of the other two pairs of clusters. This indicates that those two clusters might be oppositely co-regulated as well as being negatively correlated. In other words, they may be controlled by a common biological machinery, which while activating the subset of genes in C1, represses the subset of genes in C2, and vice versa.

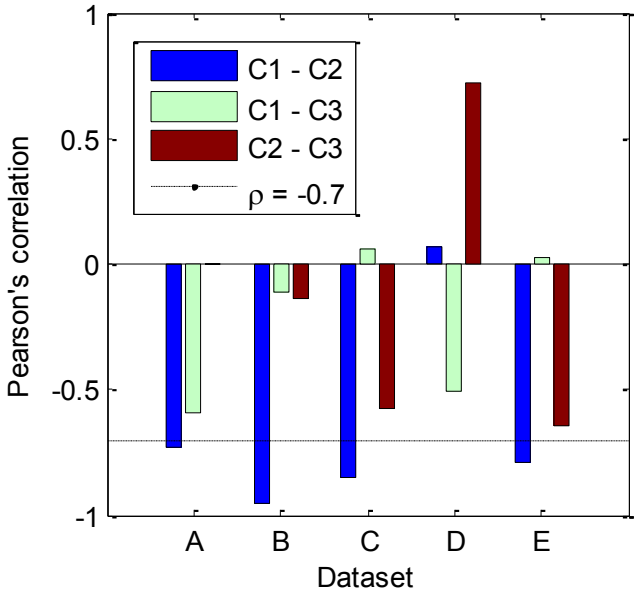

Figure 2. Pearson's correlation values between average profiles of pairs of the clusters. The clusters C1 and C2 in this case were considered at $\delta = 0.7$, while C3 was considered at $\delta = 0.2$.

4.3. Biological Analysis

We have carried out the commonly used *gene ontology (GO) term* analysis to investigate the biological relevance of the clusters C1 and C2. GO terms are terms which represent biological processes, molecular functions, or cellular components. While new discoveries are unveiled by researchers regarding genes, these discoveries are encoded by the Gene Ontology Consortium in the form of associating the considered genes with their corresponding processes, functions, and components. GO term enrichment analysis mines a subset of genes for those GO terms with which significant numbers of content genes are associated.

We have mined the clusters C1 and C2 at $\delta = 0.7$ for the enriched biological processes GO terms. C1 is highly enriched with processes related to protein synthesis as well as the cell-cycle, such as "translation" (p-value 6.7×10^{-4}), "tRNA processing" (p-value 3.4×10^{-5}), "DNA repair" (p-value 1.2×10^{-3}), and "methylation" (p-value 2.0×10^{-3}). Note that protein synthesis processes, such as ribosome biogenesis, have also been found as the most consistently co-expressed subset of genes in yeast [6, 19]. In contrast, C2 is highly enriched with processes related to transport and carbohydrate metabolism, such as "transport" (p-value 1.6×10^{-3}), "carbohydrate transport" (p-value 3.2×10^{-8}), "maltose transport" (p-value 2.8×10^{-5}), "carbohydrate metabolic process" (p-value 1.2×10^{-1}, and p-value 1.1×10^{-4} at $\delta = 0.6$), and "L-ascorbic acid catabolic process" (p-value 1.2×10^{-3}).

Another interesting aspect is those genes with unknown biological processes. At $\delta = 0.7$, C1 includes 28 genes with unknown processes out of 120 while C2 includes 9 such genes out of 57. In light of the aforementioned biological processes known to be highly enriched in these clusters, the unknown genes can be hypothesised to have similar processes as they are consistently co-expressed (correlated) with them across five different datasets generated under different conditions. This plots such hypotheses about those unknown genes which guide future work in gene discovery.

5. CONCLUSIONS

The Bi-CoPaM method has the ability to process multiple genome-wide datasets, i.e. multiple datasets with the entire set of genes without filtering, collectively and comprehensively. Filtering of less interesting genes is done implicitly within the course of Bi-CoPaM's application through two main means, tightening clusters by increasing the value of the tuning parameter δ, and filtering out clusters with lower quality from the consequent steps of analysis. By the application of Bi-CoPaM to five genome-wide *E. coli* bacterial datasets generated under different biological conditions, we have identified two major clusters for being consistently co-expressed (correlated) across different conditions. The first cluster includes genes which participate in protein synthesis and cell-cycle while the second cluster includes genes which participate in transport and carbohydrate metabolism. One hypothesis suggests that the few genes included in those two clusters with unknown biological processes may be participating in the same processes in which the other genes in those clusters participate in. Another important hypothesis is based on the striking observation that those two clusters are consistently negatively correlated with other across the different datasets, and therefore may be regulated by a common biological regulatory machinery which while activating one of them represses the other.

REFERENCES

[1] B. Palsson and K. Zengler, "The challenges of integrating multi-omic data sets," *Nature Chemical Biology,* vol. 6, pp. 787-789, 2010.

[2] P. Cahan, F. Rovegno, D. Mooney, J. C. Newman, G. S. Laurent and T. A. McCaffrey, "Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization," *Gene,* vol. 401, no. 1-2, p. 12–18, 2007.

[3] R. Nilsson, I. J. Schultz, E. L. Pierce, K. A. Soltis, A. Naranuntarat, D. M. Ward, J. M. Baughman, P. N. Paradkar, P. D. Kingsley, V. C. Culotta, J. Kaplan, J. Palis, B. H. Paw and V. K. Mootha, "Discovery of Genes Essential for Heme Biosynthesis through Large-Scale Gene Expression Analysis," *Cell Metabolism,* vol. 10, pp. 119-130, 2009.

[4] S. Monti, P. Tamayo, J. Mesirov and T. Golub, "Consensus clustering -- A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning,* vol. 52, p. 91–118, 2003.

[5] B. Abu-Jamous, R. Fa, D. J. Roberts and A. K. Nandi, "Paradigm of Tunable Clustering using Binarization of

Consensus Partition Matrices (Bi-CoPaM) for Gene Discovery," *PLOS ONE,* vol. 8, no. 2, 2013.

[6] B. Abu-Jamous, R. Fa, D. J. Roberts and A. K. Nandi, "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.

[7] B. Abu-Jamous, R. Fa, D. J. Roberts and A. K. Nandi, "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.

[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, P. T. Spellman, P. O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," in *Proc Natl Acad Sci (PNAS)*, 1998.

[10] B. Abu-Jamous, R. Fa, D. J. Roberts and A. K. Nandi, "Hybrid binarisation technique for the Bi-CoPaM method," in *Proceedings of the 2013 Constantinides International Workshop on Signal Processing (CIWSP-2013)*, London, UK, 2013.

[11] J. Lee, X.-S. Zhang, M. Hegde, W. E. Bentley, A. Jayaraman and T. K. Wood, "Indole cell signaling occurs primarily at low temperatures in Escherichia coli," *The ISME Journal,* vol. 2, p. 1007–1023, 2008.

[12] 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, p. 2065–2074, 2008.

[13] A. K. Holm, L. M. Blank, M. Oldiges, A. Schmid, C. Solem, P. R. Jensen and G. N. Vemuri, "Metabolic and transcriptional response to cofactor perturbations in Escherichia coli," *The Journal of Biological Chemistry,* vol. 285, no. 23, p. 17498–17506, 2010.

[14] K. Arunasri, M. Adil, K. V. Charan, C. Suvro, S. H. Reddy and S. Shivaji, "Effect of simulated microgravity on E. coli K12 MG1655 growth and gene expression," *PLOS ONE,* vol. 8, no. 3, 2013.

[15] S. Kamenšek and D. Žgur-Bertok, "Global transcriptional responses to the bacteriocin colicin M in Escherichia coli," *BMC Microbiology,* vol. 13, 2013.

[16] L. Kaufman and P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, John Wiley & Sons, Inc., 2005.

[17] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association,* vol. 58, no. 301, pp. 236-244, 1963.

[18] X. Xiao, E. R. Dow, R. Eberhart, Z. B. Miled and R. J. Oppelt, "Gene clustering using self-organizing maps and particle swarm optimization," in *IEEE-IPDPS*, Indianapolis, 2003.

[19] 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, p. 293–306, 2006.