Multi-View Gene Clustering using Gene Ontology and Expression-based Similarities

Swagarika Jaharlal Giri
Department of Computer Science and Engineering
Indian Institute of Technology Patna
swagarika95@gmail.com

Sriparna Saha
Department of Computer Science and Engineering
Indian Institute of Technology Patna
sriparna.saha@gmail.com

***Abstract*—Gene is a fundamental, physical and functional unit of hereditary. A proper grouping of genes is needed for a better understanding of the natural patterns amongst them. Most of the existing approaches utilize gene expression values for the clustering of genes. They mostly ignore the semantic similarity between genes obtained from the global database like Gene Ontology(GO). We present a multi-view multi-objective clustering approach where Euclidean distances between the gene expression values and GO-based multi-factored gene-gene semantic similarity are considered as two complementary views and a consensus partitioning is obtained that satisfies both the aspects or views. Two real-life gene expression datasets are used to demonstrate the effectiveness of the proposed multi-view clustering technique. We have compared our approach with various standard single-view, multi-view and multi-objective optimization based clustering approaches. The results show that better co-expressed and biologically significant gene partitions are obtained by our proposed approach. Also, to validate the significance of the obtained clusters statistically and biologically, various biological, statistical and visual significance tests were conducted, and the corresponding results are reported.**

***Index Terms*—Gene Ontology (GO), Molecular Function (MF), Cellular Component (CC), Biological Function (BF), Lowest Common Ancestor (LCA), Information Content (IC)**

I. Introduction

Gene analysis is a popular means of evaluation that offers exceptional results in the domains of Genetic Analysis and Viral diagnosis. Also, the post-genomic era has experienced a tremendous increase in genetic analysis, as such studies can provide us with explanations and even help in the prediction of inherited human disorders [23]. To assist the process of gene analysis, technologies like DNA microarray were used that can measure the expression levels of thousands of genes simultaneously [7]. With the increase in the number of genes, analyzing each and every gene at an individual level is a challenging task. Hence, to cope up with this challenge, a probable alternative is to find interesting patterns and relatedness between the genes, within the dataset. This process is assisted by various bi-clustering [2], [17] and clustering techniques [3], [5], [27]. The efficiency of such clustering and bi-clustering techniques immensely depends on similarity/distance measure between the genes, hence finding the best similarity/distance measure that can capture the relatedness between the genes has always been an open challenge.

Existing literature [2] [3] [27] suggests, co-regulated genes can be identified by performing clustering on gene expression datasets. But the similarity in expression patterns does not always mean functional relatedness between the genetic materials. It can also be because of the noise that may lead to misidentified functional and biological relationships.

To overcome this limitation, Gene Ontology (GO) [9] based biological knowledge was incorporated along with gene expression values during the clustering process to obtain better partitioning results that are biologically significant [12], [21], [22]. GO serves as a knowledge source to establish the global functional relatedness between the genes. In [21], authors have proposed a multi-objective optimization algorithm (MOC-GaPK) that uses GO-based Wang semantic similarity [30] along with the expression values to cluster genes. The results suggest how GO-based similarity helps in getting better clustering results on the expression datasets.

It was observed that clustering on such a heterogeneous dataset has some limitations. Sometimes a single set of features is not enough to capture all the important aspects of a particular dataset; in such cases, multi-view learning is a popular choice. Multi-view clustering is the method that is used to group the samples based on their similarities while considering multiple aspects simultaneously [31]. Many methods and approaches were proposed in the literature to solve the problem of multi-view clustering [8], [25]. One of the popular approaches is formulating multi-view clustering into a multi-objective optimization problem. In [25], Saha et al. have formulated the problem of multi-view clustering in a multi-objective optimization framework and obtained better clustering results on many benchmark datasets.

A. Motivation

Clustering of genes provides important insights regarding the functionalities of genes, that is the reason that many clustering techniques [28] [27] [10] [3] [18] have been applied on microarray dataset to identify the co-regulated or co-expressed genes. Microarray datasets do not guarantee global similarity as they hide the vital information required to understand the biological processes that take place in a particular organism [18] and focus only on the expression levels.

Another popular dataset that is used to establish the relatedness between the genes, is through establishing the semantic similarity between them. Literature like [14] [16] [30] [1]

978-1-7281-6929-3/20/$31.00 ©2020 IEEE

show how semantic similarity can be used to establish the functional relatedness between the genes. Also in [1], Acharya et al. proposed a new semantic similarity that considers various topological and information-theoretical properties of GO terms simultaneously. The results show that GO-based multi-factored semantic similarity establishes a better functional relationship between genes than other single factored semantic similarities. Gene expression data and GO-based multi-factored semantic similarity matrix are conditionally independent, sufficient and complementary sources of information for grouping genes into different clusters. But both these datasets were never considered simultaneously for clustering the available set of genes. In [10], authors showed that if multiple views ensure *consensus partitioning* and *complementary information* then better approximation of the clustering can be obtained. Hence in this work, we have come up with a multi-view clustering framework to solve the problem of gene clustering where gene expression values and multi-factored semantic similarity are considered as two complementary views. Our proposed framework of gene clustering not only considers the global biological knowledge extracted from GO but also considers expression patterns of the genes. As it is difficult for any multi-view clustering technique to capture the intrinsic properties of different views by just considering a single objective function, hence inspired by [25], we have formulated the problem of multi-view clustering in multi-objective optimization framework where we aim to optimize simultaneously the goodnesses of partitionings obtained using individual views and also increase the agreement amongst partitionings obtained using two different views. The main contributions of the current work are summarized below:

- Generation of GO-based view is a novel contribution to the current scenario.
- Two complementary information, GO-based and gene-expression based features, are utilized simultaneously to cluster the available set of genes.
- A unique technique for calculating the multi-factored semantic similarity between genes using GOATOOL's API is proposed in the current paper.
- Another interesting contribution is the use of multi-objective based multi-view clustering technique for gene clustering which can automatically determine the number of clusters from a given dataset.

II. Methodology

In this section, we have discussed our proposed method and briefly discussed the formation of multi-factored semantic similarity using GOATOOLS's API.

A. Datasets

Two data sets used in the current paper are described below. **Yeast Dataset:** Yeast (Saccharomyces cerevisiae)[1] expression matrix based on Tavazoie et al. [26] has 2884 rows and 17 columns where each row indicates expression values (samples) for the corresponding gene over 17 different time conditions. **B-CLL Dataset:** B-cell chronic lymphocytic leukemia progression dataset [2] that analyzes primary lymphocyte in B-CLL patients. Its expression matrix has 12624 genes with 21 samples each.

B. Framework for Multi-Factored Semantic Similarity

The entire framework for multi-factored semantic similarity computation is divided into three modules.

- Module 1: Creation of gene-GO term annotation matrix by using the GO database.
- Module 2: Extraction of GO-based features like depth, level, Information Content(IC), immediate ancestors, all ancestors for each GO term and calculation of multi-factored semantic similarity between GO terms.
- Module 3: Calculation of multi-factored semantic similarity between genes from the semantic similarity of its GO terms.

A detailed explanation is provided in the later sections.

1) Module 1: Creation of Gene-GO Term Annotation Matrix: Annotation information is available at GO-based websites like [3] and [4]. We have downloaded the complete GO tree in *.obo* format which is the JSON file for simple data recovery. We conducted some pre-processing measures on the B-CLL dataset so that genes with significant expression values are considered. A hypothesis test is performed on the entire dataset, and the p-value corresponding to each expression pattern is calculated. Smaller the p-value (typically$\leq$ 0.05), stronger is the evidence against the null hypothesis. Hence we have sorted them in increasing order of p-values (least first) and considered the top 5000 genes for our further processing. The statistics of these datasets are as follows: In our case, we have considered 1842 Yeast genes and 2891 B-CLL genes that had their annotations in all three ontologies (BP, MF, CC). These genes are considered for further processing. As the statistics suggest, the three ontologies have an exclusive set of GO terms. Let a, b, c be the number of significant GO terms in CC, MF, BP, respectively. Hence $(a + b + c)$ is

TABLE I: GOTermMapper annotation data for 2884 yeast genes and 5000 B-CLL genes, here BP, CC, MF refer, respectively, to biological process, cellular component and molecular function.

		Yeast			B-CLL		
		BP	MF	CC	BP	MF	CC
Mapped gene		2264	1978	2466	3564	2989	3634
Unique GO terms		100	43	23	70	40	34
Unmapped genes	identified ambiguous	1	1	1	8	8	8
	unannotated	224	224	250	246	246	246
	not annotated in slim	16	77	20	35	589	35
	no root annotation	292	593	168	29	31	21

[1] http://arep.med.harvard.edu/biclustering/

[2] http://www.ncbi.nlm.nih.gov/geo,GDS1388

[3] https://go.princeton.edu/cgi-bin/GOTermMapper

[4] http://www.geneontology.org/

the total number of GO terms and the resultant matrix is of dimension $n \times (a+b+c)$ where n is the number of genes under consideration. In the yeast dataset, the gene-GO term annotation matrix is of size 1842×166, and for the B-CLL dataset, the gene-GO term annotation matrix has a size of 2891×144.

The Mathematical formulation is as follows:
Let $\exists n$ genes and a, b, c be the number of significant GO terms present in BP, MF, CC, respectively.
The size of gene-GO term matrix $|M| = n \times (a+b+c)$
GO_k represents k^{th} significant GO term where $k \in [1, a+b+c]$.

Let M[n][a+b+c] be the binary annotation matrix of size $n \times (a+b+c)$. The matrix is generated as described in [1] as follows:

$$M[i][GO_k] = \begin{cases} 1, & \text{if } G_i \text{ is annotated with } GO_k \\ 0, & \text{otherwise} \end{cases}$$

where $i \in [1, n]$ and $k \in [1, a+b+c]$.

2) Module 2: Calculation of Multi-Factored Semantic Similarity between the GO terms: The multi-factored semantic similarity between the GO terms is the combination of 3 individual semantic similarity measures, that are Lin's, Shen's and Normalized $struct_{depth}$. In this subsection, we will briefly explain the calculation of multi-factored semantic similarity measures between the GO terms using GOATOOL's API.

Calculation of Lin's Similarity using GOATOOLS's API: For calculating Lin's similarity between the GO terms, the different structural and information-theoretical properties that we need to know are **1.** Lowest Common Ancestor (LCA) between the two GO terms, t_i and t_j, **2.** The Information Content(IC) of the LCA and **3.** The individual IC of the terms, t_i and t_j. LCA is the most informative common ancestor of the GO term pair. GOATOOLS'S API provides us with an interface to find the structural, topological and information-theoretical properties of GO terms like depth (longest path length between the GO term and the root), level (shortest path length between the GO term and the root), all immediate ancestors, etc. The LCA for the GO term pairs can be extracted by recursively traversing the ancestors' list and finding all common parents for the pair. The parent having maximum IC is regarded as the LCA, and its information is used for calculation of the following semantic similarity.

Calculation of Shen's Similarity using GOATOOLS's API: This is a hybrid semantic similarity measure that considers both the IC of ancestor terms as well as shortest path length connecting the LCA with the individual term. For calculating the Shen's similarity between the GO terms, we have arranged the GO terms and their ancestors in the form of a weighted directed graph where each node represents a GO term. For calculating the Shen's similarity, we need to find the value of $\sum_{t_1 \epsilon path_i} \frac{1}{IC[t_1]}$, here $path_i$ is the shortest path connecting the GO term, t_i with it's LCA, and t_1 is the set of ancestor GO terms that appear in the shortest path. The topological and information-theoretical information were already extracted as mentioned above. Utilizing these, we develop the weighted graph framework and use the Dijkstra algorithm for finding the shortest path between the GO term and its LCA.

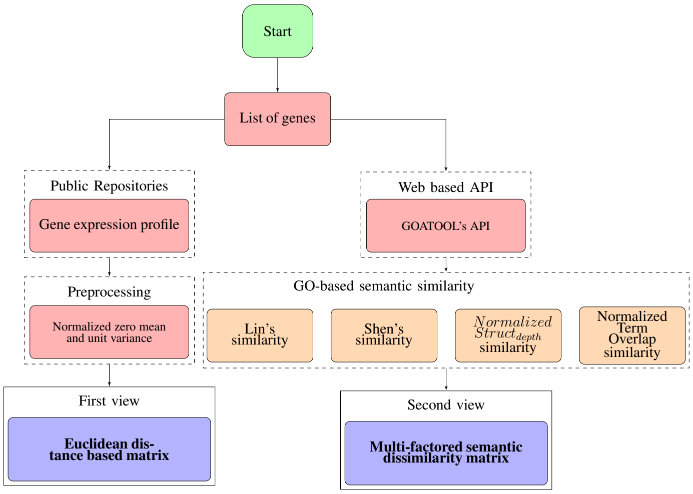
Start
List of genes
Public Repositories
Gene expression profile
Web based API
GOATOOL's API
Preprocessing
Normalized zero mean and unit variance
GO-based semantic similarity
Lin's similarity
Shen's similarity
Normalized Struct_depth similarity
Normalized Term Overlap similarity
First view
Euclidean distance based matrix
Second view
Multi-factored semantic dissimilarity matrix

Fig. 1: Flowchart of view formation

Calculating Normalized $struct_{depth}$ Similarity using GOATOOLS's API: Normalized $struct_{depth}$ is based on simple GO tree property and requires the depth information of LCA and the entire GO tree. Using GOATOOLS's API, the depth and level information corresponding to each GO term can be extracted easily. Using the above three semantic similarities, the Multi-Factored Semantic Similarity between the GO terms, t_i and t_j, can be calculated using the following equation

$$\textit{Multi-sim}(t_i, t_j) = \frac{arctan[Y]}{\pi/2} \tag{1}$$

Where,

$$Y = [sim_{Lin}(t_i, t_j) + sim_{Shen}(t_i, t_j) + sim_{norm-struct_{depth}}(t_i, t_j)]$$

$sim_{Lin}(t_i, t_j)$ is Lin's similarity between the GO terms.
$sim_{Shen}(t_i, t_j)$ is Shen's similarity between the GO terms.
$sim_{norm-struct_{depth}}(t_i, t_j)$ is Normalized $struct_{depth}$ between the GO terms.

3) Module 3: Calculating Multi-Factored Semantic Similarity between Genes: Multi-factored semantic similarity between gene products, g_i and g_j, can be calculated according to the following Equation

$$\textit{Multi-SIM}(g_i, g_j) = \frac{\frac{1}{m \times n} \sum_{t_k \epsilon A_i, t_p \epsilon A_j} \textit{Multi-sim}(t_k, t_p) + sim_{NTO}(g_i, g_j)}{2} \tag{2}$$

Here, $sim_{NTO}(g_i, g_j)$ is the normalized term overlap score between genes, g_i and g_j, and can be easily calculated utilizing the gene-GO term binary matrix.

C. Framework for Microarray Distance Matrix

In this subsection, we will discuss the generation of distance matrix from the microarray dataset. Gene expression vector across each gene is standardized to have zero mean and unit variance so that the Euclidean distance and Pearson Correlation of any two gene vectors are essentially equivalent. This is done to normalize the gene vectors and remove the sources of variations that affect the measured expression levels. Euclidean distance between every two gene vectors is calculated and stored in a matrix form. For the ease of computation, we have converted Multi-Factored Similarity between genes into Multi-Factored Dissimilarity between genes using the following formulation,

$$Multi - DISSIM(g_i, g_j) = 1 - Multi - SIM(g_i, g_j) \quad (3)$$

as dissimilarity is comparable with distance. In one case (view 1) the objective is to minimize the dissimilarity, and another case (view 2) is the minimization of distance to get better clustering results.

The next section will cover how multi-view clustering is applied to the above discussed two views to obtain a consensus partitioning.

D. Multi-view Clustering Technique

This subsection aims to obtain consensus partitioning on the above gene list after considering both the views or aspects, simultaneously. Some cluster validity measures calculated on individual views are optimized simultaneously to obtain the final partitioning from the data set. Also, all the alternative partitionings of the particular dataset need to be captured simultaneously. Hence to meet the above-mentioned requirements, the problem is formulated as a multi-objective multi-view clustering problem.

Multi-view clustering problem can be formulated in a multi-objective optimization framework where the following objectives are required to be optimized simultaneously:

1. Better clustering results on the first view, **2.** Better clustering results on the second view, **3.** Better Agreement Index between the partitionings obtained using two different views. The underlying optimization strategy used for solving the above mentioned multi-objective optimization problem is AMOSA (archived multi-objective simulated annealing) [4]. Any other multi-objective optimization algorithm could have been used but as it has already been shown in the literature that AMOSA outperforms several other existing multi-objective evolutionary algorithms [2] hence that has made it a valid choice. Here we have applied a multi-objective multi-view based clustering technique to combine the two different views while generating the gene-clusters. Our proposed algorithm is inspired by the approach proposed in Ref. [25]. The general flow follows that of Ref. [25]. But some modifications are incorporated in the current framework to better handle the newly generated views.

The key steps of this multi-view based multi-objective clustering technique are elaborated below:

1) Solution Representation: In this work, a center-based solution representation is used. In such representation, only cluster centers are encoded in the solution. For each solution, three structures are maintained simultaneously, one for each view and the last one for the consensus partitioning. The number of centroids/medoids in each structure are the same. Each center corresponds to the index of the gene that is the representative point of the cluster.

2) Archive Initialization: The first step in the proposed methodology is to initialize the archive with some random solutions. Here each solution contains a different set of centroids. As the proposed clustering technique is automatic in nature, *the number of centroids* and *the cluster centers* are chosen randomly and the number of centroids present in each solution varies over a range, $K = 2$ to $K = \sqrt{N}$.

3) Membership Calculation: Assigning the genes to their respective clusters is one of the important steps in our proposed approach. We have used K-medoid clustering [20] technique for calculating the membership values of the sample gene points. It is a minimum-distance based assignment of the sample points to their nearest centroids.

4) Updation of Cluster Centers: After assigning the genes to their respective clusters, it is important to update the centroid to the most central element in the cluster. Multiple *Calculate Membership* and *Update Center* operations are applied on the clusters to identify better cluster centers for each solution.

5) Objective Functions: Many cluster validity indices could have been used but considering the crisp and fuzzy nature of our obtained clusters, PBM index [19] is used to validate the clusters formed in the above steps. We have also used an agreement index that measures the agreement between the partitionings obtained using both the views. The *cluster validity index* on clusters obtained from *gene expression* dataset (first view), *cluster validity index* on clusters obtained from *GO-based* dataset (second view) and *Agreement index* (AI) computed on partitionings obtained using *both the views* are the three objective functions that are to be optimized simultaneously. AI quantifies the similarity between partitionings obtained using both the views. The search capability of any MOO based technique can be utilized to optimize these three objective functions simultaneously. The aim is to identify some good partitionings using different views, which are also consensus (similar) partitionings across different views.

6) Mutation: Due to the large search space, there is a chance that the AMOSA algorithm may get stuck at local optima. To overcome this drawback and efficiently explore the search space using AMOSA, various mutation operations were performed on the individual solutions.

Normal Mutation: This type of mutation operation keeps the number of existing clusters the same but makes some changes in the cluster centers. For an update, a random value is drawn using the Laplacian distribution and the current cluster center is replaced with a new cluster center. The Laplacian distribution is used so that the probability of generating a value similar to the old value would be high.

Insert Mutation: The purpose of this mutation operator is to increase the number of clusters present in a solution. A random gene is selected from the gene set and that is inserted in the current solution. The randomly selected gene will act as the new center. This mutation will increase the number of centroids. The mutation step is then followed by *Update*

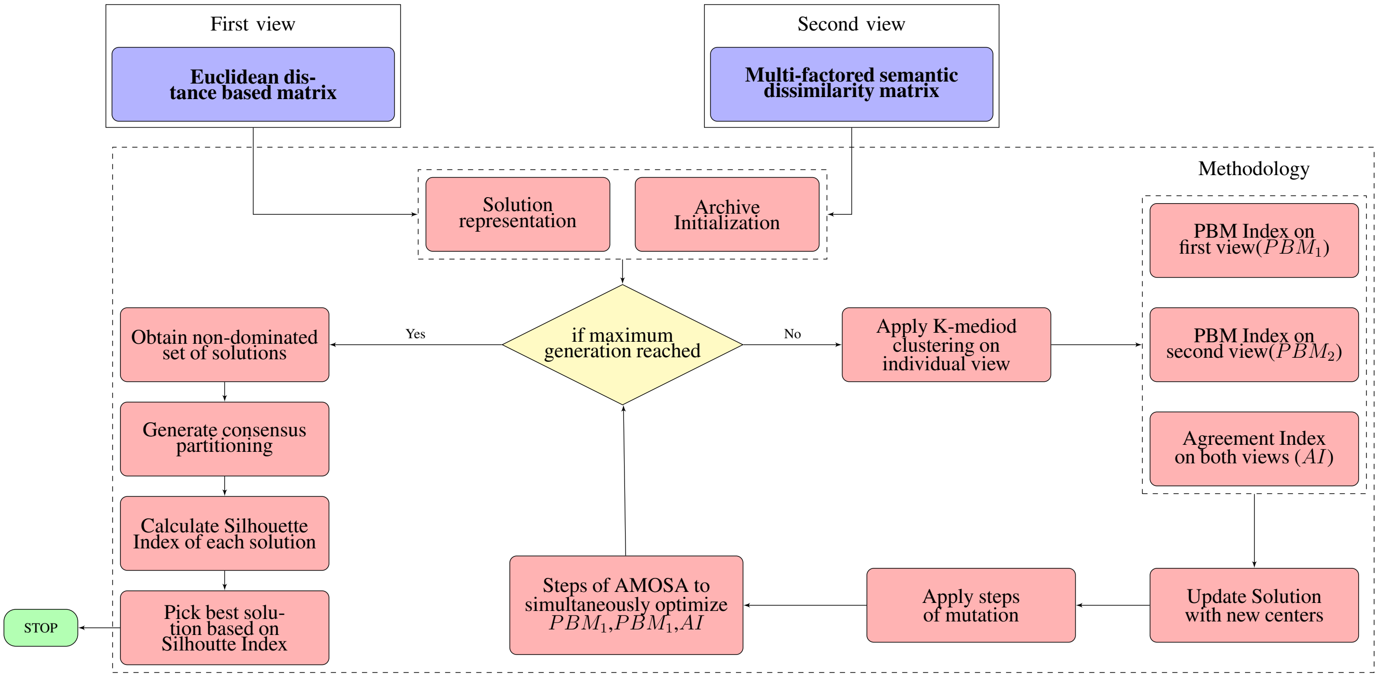
First view
Euclidean distance based matrix
Second view
Multi-factored semantic dissimilarity matrix
Methodology
Solution representation
Archive Initialization
PBM Index on first view(PBM_1)
Yes
if maximum generation reached
No
Obtain non-dominated set of solutions
Apply K-mediod clustering on individual view
PBM Index on second view(PBM_2)
Generate consensus partitioning
Agreement Index on both views (AI)
Calculate Silhouette Index of each solution
STOP
Pick best solution based on Silhoutte Index
Steps of AMOSA to simultaneously optimize PBM_1,PBM_1,AI
Apply steps of mutation
Update Solution with new centers

Fig. 2: Flowchart of our proposed methodology

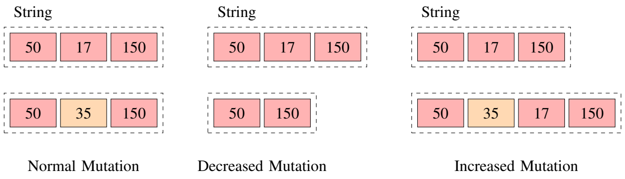

Fig. 3: Examples of different types of mutation operations

Membership operation as membership has to be re-computed for the new cluster center.

Delete Mutation: This type of mutation is used to decrease the number of cluster centers encoded in a solution. A random cluster center is selected and then deleted from the solution. As the center no longer exists, all the sample points assigned to that cluster center are required to be reassigned.

7) Consensus Partitioning: The pivotal task is to identify the one-one correspondence between different partitions obtained using both the views. We extracted the common/consensus sample points in both the partitionings. Further, these points are used to determine the new cluster centers. The most central point from those present in the consensus partitioning is chosen as a new cluster center for that partitioning and the other points are then assigned to these centers using a minimum distance criterion to get a final consensus partitioning.

E. Selection of Single Solution

On termination, a set of non-dominated solutions is obtained on the final Pareto optimal front and those are stored in the archive. For a single solution, membership matrices corresponding to different views are combined to obtain a single consensus partitioning. The procedure to obtain consensus partitioning is already explained in Section II-D7. All the solutions present in the archive are having equal importance, any single solution can be selected by the user based on his/her requirement.

III. Results and Discussion

This subsection focuses on the necessary parameter settings required to get the clustering results using the AMOSA based framework. We have also discussed the results obtained by our proposed approach on a standard dataset with the existing clustering approaches. Unlike SOO where we get a single clustering solution, in our proposed framework, we obtain a set of solutions that form the Pareto-optimal front. Every solution has a variable number of cluster centers encoded in it. All the solutions on the final archive are evaluated in terms of an internal cluster validity index, namely *Silhouette Index* [24]. This is an internal cluster validity index, which can be used to measure the goodness of the partitioning. Also, to establish the statistical and biological significance of the results we have also conducted the various biological and statistical significance tests on the obtained clusters.

A. Parameter Setting

Parameter setting plays a very important role to exploit the search capability of AMOSA. The detailed discussions of all the parameters, their functionalities and their interdependence are already explained in original work of AMOSA [4]. Table:II suggests the parameter settings used in our proposed approach.

TABLE II: Parameter settings for underlying MOO-algorithm, AMOSA

Parameters	Value	Discussion
Dimension	3	As there are three structures of indices of centroids; one for each view and one combined
Min_Len	2	Minimum number of centroids possible
Max_Len	60	Maximum number of centroids possible
SL	40	The maximum size of the Archive at the initial stage before applying clustering.
HL	30	The maximum size of the Archive on termination. It holds the set of non-dominated solutions.
Mutate Normal	$0.75 > and \geq 0.15$	Encoded cluster center is perturbed by small amount.
Mutate Delete	$0.15 > and \geq 0.1$	Decreases solution length by deleting random cluster centers
Mutate Insert	< 0.1	Increases solution length by adding random cluster centers
alpha	0.75	cooling rate
T_min	0.001	Initial Temperature
T_max	100	Final Temperature
iter	30	Number of iterations at each temperature
K_{min}	2	Minimum value of number of clusters
K_{max}	$\sqrt{N}$	Maximum value of number of clusters
Max_generation	20	-

B. Discussion on Compared Methods

The experimental results obtained by the proposed approach are compared with several traditional, single-view and multi-objective optimization based clustering techniques. The proposed approach is compared with traditional techniques like K-means [15], K-medoid [29] and hierarchical [11] clustering algorithms. These traditional clustering techniques are computationally less complex but very sensitive to noise. Also, to use such clustering techniques, the number of clusters is to be known beforehand. Recently, some techniques were introduced which integrate the biological knowledge along with the gene expression profile for clustering the genes. In [22], authors have applied C-means clustering algorithm on integrated dissimilarity measure obtained by considering both expression dataset and GO-based semantic distance. The drawback of such an integrated dataset is that it cannot capture the intrinsic and extrinsic properties of the individual datasets. Recently, Parraga et al. [21] have proposed gene clustering algorithm (MOC-GaPK) which is also a multi-objective optimization algorithm. This approach also incorporates biological knowledge along with expression values to cluster the genes. In order to illustrate the effectiveness of the proposed technique, we have compared our proposed approach with all the above mentioned approaches. Our proposed approach, in terms of finding the number of clusters, is automatic in nature whereas the compared methods are required to be supplied with the number of clusters beforehand. Hence K is varied in the range [2,$\sqrt{N}$] where N is the number of genes. Silhouette index [13], s(C), is an efficient cluster validity index that quantifies the clustering in terms of tightness and separation of the clusters; it was used to determine the best partitioning. The K value corresponding to the best Silhouette index, s(C), is finally selected. Our proposed approach is a MOO-based clustering technique; hence, we get a set of non-dominated solutions where each solution is a different partitioning result with a different number of clusters. So Silhouette index was again used to select the best solution amongst them. All comparative approaches along with the proposed approach are executed 20 times on each dataset and then the results are reported.

C. Discussion on Results

In this section we have compared the performance of the proposed approach with several state-of-the-art techniques in terms of four different aspects.

1) Number of Clusters: It is very difficult to identify the appropriate number of clusters from such biological datasets. Also because of the ever-changing nature of the genome, the number of clusters keeps on changing. Our proposed approach is automatic in comparison to other comparative approaches where the number of clusters is to be known beforehand. For the other comparative methods, the K value was varied in the range [2,$\sqrt{N}$] and the cluster partition having the highest Silhouette index, s(C), was reported. Our proposed approach provides a set of non-dominated solutions with a variable number of clusters and optimal partitioning can be automatically identified.

2) Integration of GO-based Biological Knowledge with Gene Expression Data: The results prove that better clustering partitions that share more common biological properties are obtained by incorporating biological knowledge along with expression data during the clustering process. In [22] authors have combined the dissimilarity scores obtained from the two knowledge sources. Comparative results show that the combined dissimilarity score fails to capture the individuality of the independent knowledge source hence better results were obtained by our proposed approach that allows the individual knowledge source to retain it's intrinsic and extrinsic properties, and a consensus partitioning is obtained after considering both the views.

3) Underlying Multi-objective Optimization Algorithm: The underlying MOO algorithm used in our approach is AMOSA [4] whereas that used in MOC-GaPBK is NSGA-II. Both of these MOO algorithms are extensively used for clustering. Comparative results are reported in Table:III. In our approach, we have used multi-factored semantic distance between two genes as a biological knowledge source whereas Parraga et al. [21] have used G-SemSim library that gives the Wang semantic measure between the genes. The literature [1] has already shown the superiority of multi-factored semantic measure over Wang semantic measure to capture the biological relationships between the genes. Hence we have observed better results than the above-mentioned method.

4) Single-factored Semantic Measure vs Multi-factored Semantic Measure: We have also compared the effect of multi-factored semantic measures between the genes with other single factored semantic measures on our proposed framework. The results in Table:IV show that multi-factored semantic measure helps in obtaining better partitions that are biologically similar and share similar biological properties.

D. Statistical Significant Test

To validate the improvements attained by our proposed method statistically, we have performed a statistical significant test named Welch's t-test [6] at 5%(0.05) significant level.

TABLE III: The maximum *Silhoutte Score* values obtained by different single-view clustering techniques as compared to our proposed multi-view clustering technique, here $View_{GE}$ is the distance matrix obtained from Gene Expression values and $View_{Multi-fact}$ is the Multi-factored semantic dissimilarity matrix

Algorithm	View Name	Silhoutte Index on Yeast Dataset	Silhoutte Index on B-CLL dataset
K-Mean	$View_{GE}$	0.45	0.65
K-Medoid	$View_{GE}$	0.46	0.68
K-Medoid	$View_{Multi-fact}$	0.30	0.46
Hierarchial	$View_{GE}$	0.35	0.70
Hierarchial	$View_{Multi-fact}$	0.25	0.38
Approach proposed in [22]	Normalized CBD and GO-based semantic measure	0.44	0.64
MOC-GaPBK [21]	Pearson correlation and GO-based Wang semantic measure	0.56	0.82
Proposed Algorithm using single view	$View_{GE}$	0.52	0.75
Proposed Algorithm using single view	$View_{Multi-fact}$	0.35	0.58
Proposed Algorithm using multiple views	$View_{Multi-fact}$ and $View_{GE}$	**0.61**	**0.85**

TABLE IV: The maximum *Silhouette Score* values obtained by proposed method using various semantic measures. In each case, gene expression values were used as the complementary views.

GO-Based Semantic Similarity	Silhoutte Index on Yeast dataset	Silhoutte Index on B-CLL Dataset
Shen's	0.49	0.76
Normalized $struct_{depth}$	0.46	0.58
Normalized Term Overlap	0.30	0.75
Lin's	0.45	0.68
GOATOOLS	0.51	0.64
Multi-factored	**0.61**	**0.85**

It indicates that the performance improvement that we have obtained is statistically significant and not obtained by chance. For that, we have calculated the p-value obtained by Welch's test for comparison of two different groups. Each considered algorithm was executed 20 consecutive times and their performance metrics (Silhouette Score) were computed for both the benchmark datasets.

The list of Silhouette scores produced by our algorithm and another compared algorithm is supplied to Welch's t-test and the p-value was analyzed. In each case, a p-value (typically$<$ 0.05) was observed. The p-value (typically$<$ 0.05) claims that the performance improvements attained by our algorithm are statistically significant.

E. Cluster Profile Plots

To validate the partitions obtained by our proposed approach, we have used various data visualization techniques. These methods enable us to visualize the actual coherence between the genes within the same cluster. The cluster profile plots of Fig:4 show how the genes placed in the same cluster are coherently similar for the B-CLL dataset. For other obtained clusters, similar profile plots are obtained.

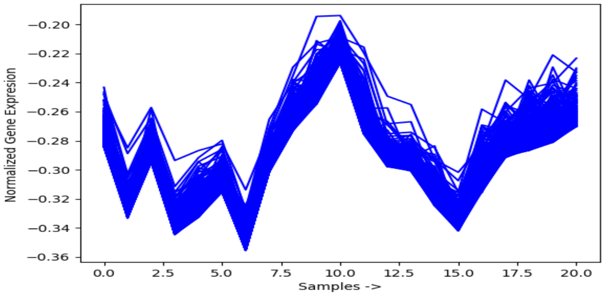

Fig. 4: Cluster profile plot of one cluster (having 103 genes and 21 samples) obtained from our proposed methodology on B-CLL dataset

TABLE V: Results for biological significance test: first two clusters obtained from proposed multi-view based clustering algorithm

Cluster	GO term	Cluster%	Genome%
Cluster1 103 genes	GO:0051179 localization	0.32	0.22
	GO:0050896 response to stimulus	0.33	0.18
	GO:0009987 Cellular Process	0.85	0.69
Cluster2 267 genes	GO:0008152 metabolic activity	0.78	0.52
	GO:0009056 catabolic activity	0.24	0.12

F. Biological Significance Test

To verify whether the partitions obtained after application of our proposed approach are biological significant or biologically enriched, we have performed a biological significance test with the help of GOTERMMAPPER. The results of the first two clusters out of the total four clusters obtained for the Yeast dataset are shown in Table:V. In this subsection, we have summarized significant GO terms shared by genes of corresponding clusters. For each GO term, the percentage of genes sharing that GO term within the same cluster and those in the entire genome was reported in the Table:V. The results suggest stronger biological relationships between the genes in the same cluster obtained by our proposed approach than the entire genome. The higher percentage of shared GO terms between the genes suggests that these genes are more involved in similar biological processes compared to the remaining genes of the genome.

IV. CONCLUSION AND FUTURE WORKS

In this research article, we have proposed a multi-view multi-objective clustering framework in order to address the problem of gene clustering. Two views: one based on Euclidean distance between gene expression values and other based on a recently proposed GO-based gene-gene similarity measure are utilized as two complementary views to identify clusters of functionally similar genes. The proposed multi-view clustering framework is multi-objective in nature. Three objective functions, cluster quality measures calculated on the partitionings obtained using individual views and an agreement index measuring the consensus between partitionings obtained using two views are simultaneously optimized using the search capability of AMOSA. Two well-known benchmark datasets like Yeast and B-cell chronic lymphocytic leukaemia progression dataset, are utilized for conducting the experiments. A thorough comparative study has been performed with respect to some well known single view based clustering algorithms. Obtained results establish the fact that considering multiple views provides better clustering solution compared to existing single view based gene clustering techniques. Qualities of obtained clusters are verified using visual plots like cluster profile plot. At the end, to establish the superiority of our proposed algorithm statistically and biologically, a statistical significance test and a biological significance test have been conducted.

In the future, we would like to apply our proposed approach on datasets whose true labels are known so that we could have a better approximation of the resultant clusters formed. Also, we would like to integrate the proposed framework with deep-learning to develop some cluster ensemble-based techniques to identify the gene labels better. The multi-objective based approach provides a set of solutions on the final Pareto optimal front. All these solutions can be combined with the use of deep-learning based techniques to further improve the accuracy of the obtained partitioning.

Acknowledgement

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

References

[1] S. Acharya, S. Saha, and P. Pradhan. Multi-factored gene-gene proximity measures exploiting biological knowledge extracted from gene ontology: application in gene clustering. *IEEE/ACM transactions on computational biology and bioinformatics*, 2018.

[2] S. Acharya, S. Saha, and P. Sahoo. Bi-clustering of microarray data using a symmetry-based multi-objective optimization framework. *Soft Computing*, pages 1–22, 2018.

[3] H. Azzawi, J. Hou, Y. Xiang, and R. Alanni. Lung cancer prediction from microarray data by gene expression programming. *IET systems biology*, 10(5):168–178, 2016.

[4] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb. A simulated annealing-based multiobjective optimization algorithm: Amosa. *IEEE transactions on evolutionary computation*, 12(3):269–283, 2008.

[5] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of computational biology*, 6(3-4):281–297, 1999.

[6] D. Best and J. Rayner. Welch's approximate solution for the behrens–fisher problem. *Technometrics*, 29(2):205–210, 1987.

[7] P. O. Brown and D. Botstein. Exploring the new world of the genome with dna microarrays. *Nature genetics*, 21(1s):33, 1999.

[8] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136, 2009.

[9] O. Gene et al. Gene ontology consortium: going forward. *Nucleic Acids Res*, 43(Database issue):D1049–56, 2015.

[10] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: a survey. *IEEE Transactions on knowledge and data engineering*, 16(11):1370–1386, 2004.

[11] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[12] R. Kustra and A. Zagdanski. Incorporating gene ontology in clustering gene expression data. In *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pages 555–563. IEEE, 2006.

[13] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162, 1994.

[14] D. Lin et al. An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304. Citeseer, 1998.

[15] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[16] M. Mistry and P. Pavlidis. Gene ontology term overlap as a measure of gene functional similarity. *BMC bioinformatics*, 9(1):327, 2008.

[17] A. Oghabian, S. Kilpinen, S. Hautaniemi, and E. Czeizler. Biclustering methods: biological relevance and application in gene expression analysis. *PloS one*, 9(3), 2014.

[18] J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh, M. Achas, and E. Adebiyi. Clustering algorithms: Their application to gene expression data. *Bioinformatics and Biology insights*, 10:BBI–S38316, 2016.

[19] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik. Validity index for crisp and fuzzy clusters. *Pattern recognition*, 37(3):487–501, 2004.

[20] H.-S. Park and C.-H. Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

[21] J. Parraga-Alava, M. Dorn, and M. Inostroza-Ponta. A multi-objective gene clustering algorithm guided by apriori biological knowledge with intensification and diversification strategies. *BioData mining*, 11(1):16, 2018.

[22] S. Paul. Integration of gene expression and ontology for clustering functionally similar genes. In *International Joint Conference on Rough Sets*, pages 587–598. Springer, 2017.

[23] E. Poliakov, D. N. Cooper, E. I. Stepchenkova, and I. B. Rogozin. Genetics in genomic era. *Genetics research international*, 2015, 2015.

[24] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[25] S. Saha, S. Mitra, and S. Kramer. Exploring multiobjective optimization for multiview clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(4):44, 2018.

[26] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature genetics*, 22(3):281, 1999.

[27] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19):2405–2412, 2006.

[28] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 2017.

[29] M. Van der Laan, K. Pollard, and J. Bryan. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584, 2003.

[30] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[31] Y. Yang and H. Wang. Multi-view clustering: A survey. *Big Data Mining and Analytics*, 1(2):83–107, 2018.