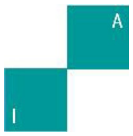
Inteligencia Artificial 25(69) (2022), 1-12
doi: 10.4114/intartif.vol25iss69pp01-12



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

A Gene Expression Clustering Method to Extraction of Cell-to-Cell Biological Communication

Hui Wang [1,#], Yan Sha [1,#], Dan Wang [1,*], Hamed Nazari [2]

[1] School of Medical Information & Engineering, Xuzhou Medical University, Xuzhou 221004, China
[2] Department of Computer Engineering, Urmia University, Urmia, Iran
[#] Hui Wang and Yan Sha are contribution equal to first author
[*] Corresponding Author: Dan Wang
Email: wanghui07401@163.com, uestcsy2009@126.com, ndskywd@126.com, st_h.nazari@urmia.ac.ir

Abstract Graph-based clustering identification is a practical method to detect the communication between nodes in complex networks that has obtained considerable comments. Since identifying different communities in large-scale data is a challenging task, by understanding the communication between the behaviours of the elements in a community (a cluster), the general characteristics of clusters can be predicted. Graph-based clustering methods have played an important role in clustering gene expression data because of their ability to show the relations between the data. In order to be able to identify genes that lead to the development of diseases, the communication between the cells must be established. The communication between different cells can be indicated by the expression of different genes within them. In this study, the problem of cell-to-cell communication is expressed as a graph and the communication are extracted by recognizing the communities. The FANTOM5 dataset is used to simulate and calculate the similarity between cells. After pre-processing and normalizing the data, to convert this data into graphs, the expression of genes in different cells was examined and by considering a threshold and Wilcoxon test, the communication between them were identified through using clustering. The results of the comparisons showed that the proposed method Silhouette coefficient of 0.814 with a threshold of 0.2 for cells and 0.789 with a threshold of 0.1 performs better than the state-of-the-art methods.

Keywords: Gene Expression, Normalization, Wilcoxon test, Cell-To-Cell Communication, Louvain Clustering.

1 Introduction

In the real world, networks are used to show the communication between different types of complex systems, such as social networks, biological networks, Internet networks, etc. [1]. One of the salient features of networks that has become a contravention subject of research is the structure of society [2]. A community is a subset of nodes that are very similar to each other. The structure of society, such as the expression of genes, proteins, and promoters, is used to diagnose various diseases, analyse social networks, and so on. Various algorithms have been designed to identify these communities [3] [4]. These algorithms can be generally categorized as random methods [5], spectral clustering and partitioning [6], modulation [7], spectral matrix and matrix factorization [8].

One of the cases in which community identification can be used in medicine and bioinformatics is the analysis of gene expression data [9]. Gene expression data are used to diagnose diseases based on different conditions. With the increasing development of technology, a lot of data on gene expression have been obtained, but no useful information has been extracted from them. In order to extract information from gene expression data, it is possible

ISSN: 1137-3601 (print), 1988-3064 (on-line)

to identify clusters and communities based on graph-based clustering, so that each community (cluster) has the most internal similarity and is the most different from other communities [10].

The continuation of this paper is as follows: The concept of gene expression and previous research is presented in Section II. Section III deals with the technical terms of the research (graph analysis, normalization, Wilcoxon test, and silhouette index). Extraction of cell-to-cell communication based on gene expression clustering is discussed in Section VI. Section V reports the results and experiments related to the proposed clustering algorithm. This section also presents the results of clustering and inter-cellular and interstitial relation on the FANTOM5 dataset. Finally, conclusions and further studies are discussed in Section VI.

2 RESEARCH BACKGROUND

Numerous computational studies in the field of mathematics and graphs have been performed to analyse cell behaviour, with extensive contributions to biological knowledge [11], community identification [12], and clustering [13] [14]. Identifying more communities is used to investigate the structural features of complex graphs [15]. For example, various studies have shown that community discovery can help identify the authors of the articles with similar themes on the subject of social graphs [16].

In recent years, several methods have been proposed to identify communities. One of the basic algorithms for this purpose is the minimum shear algorithm [17]. In a minimum shear algorithm, a graph is divided into a number of predefined segments so that the number of edges between communities is minimal. Graph addition algorithms, such as, spectral clustering algorithms [18] are another type of community recognition algorithms. Also, optimization-based algorithms, including modularity optimization, external optimization, spectral optimization to solve community discovery problem have been proposed [16] [19].

Hou et al. (2020) introduced an approach for predicting cell-to-cell communication networks through Network Analysis Toolkit for Multicellular Interactions (NATMI) [17]. This approach uses single-cell expression data to predict and visualize cell-to-cell communication networks. Lizio et al. (2015) investigated the relationships between FANTOM5 elements at the promoter level [18]. The authors created this dataset in a useful format for multiple purposes through a set of database systems, where they are complementary in terms of hosted data or context. Mojarad et al. (2021) modelled the behaviour of inherited diseases on the FANTOM5 dataset [19]. In this paper, the identification of intercellular and interstitial connections for different diseases has been performed using an innovative similarity criterion of the characteristics of the topological structure of the graph and a set of extensive clustering.

Paolicelli et al. (2019) examined the biogenesis and function of extracellular vesicles associated with microglia and focused on their possible role in the pathology of Alzheimer's disease [20]. Extracellular vesicles are important mediators of cell-to-cell communication. Also, microglia-associated extracellular vesicles play an important role in neurodegeneration. In fact, microglia act as a source and receptor for extracellular vesicles in the brain. AlMusawi et al. (2021) analyzed the current in vitro and in vivo mono-cellular and multi-cellular cultures models of colorectal cancer [21]. This was done with the aim of understanding cell-to-cell biological communication and signaling in the microenvironment of this disease on the FANTOM5 dataset. In addition, the process of separating different types of molecules based on single-cell multi-omics approaches has been investigated. Almet et al. (2021) analyzed the prospect of cell-to-cell communication through single-cell transcription [22]. Here, methods that use non-spatial single-cell and spatial data are reviewed. In addition, the authors introduce various approaches to advance current cell-to-cell communication inference.

Other algorithms that are widely used to identify communities are hierarchical clustering algorithms [23] that use the criterion of similarity between pairs of nodes for clustering. In these algorithms, nodes with the highest similarity criteria are placed in a community. Newman has proposed a hierarchical division-based algorithm for identifying communities [24]. In this algorithm, the edge that has a high partition value is deleted. Newman has articulated the problem of community discovery as an optimization problem by considering a criterion called Modularity.

In this study, we use a graph-based clustering-based on modulation method to cluster gene expression data. In order to reduce the scattering of gene expression data and their compatibility with basic clustering algorithms, it is necessary to normalize and calculate the semantic communication between samples. The purpose of clustering gene expression data is to extract communication between different cells in this data. Due to the unique characteristics of gene expression data and its difference from the general data set, the proposed method by mapping this data to graphs in addition to reducing the computational volume, the possibility of applying basic graph-based clustering

algorithms is suggested. Therefore, this case can be expressed as one of the main differences between the proposed method and other similar methods.

3 TECHNICAL TERMS OF RESEARCH

In this section, the technical terms of the research include graph mining, normalization, and Wilcoxon test. Graph mining is one of the new methods for extracting data represented by the graphs. Normalization is one of the essential steps for analysing gene expression data and Wilcoxon test is one of the statistical tests to evaluate the semantic communication of the two samples depending on the ranking scale.

3.1 Graph mining

Graph G is a binary set of $G = (V, E)$ in which V is a set of nodes and E contains the edges of the graph. $n = |V|$ the number of nodes (graph order) and $m = |E|$ indicates the number of edges (graph size). In a weighted graph, the weight function W is defined as $W \rightarrow R$, which assigns a weight to each edge of the graph. In general, the density of a graph is the ratio of the number of edges in the graph to the level of the maximum possible edges. Eq. (1) defines the density of a graph [25].

$$\partial(G) = \frac{m}{n} for\ n \in \{0.1\}, we\ set\ \partial(G) = 0 \quad (1)$$

There are different criteria for graph analysis according to its size, which can be referred to as centrality of proximity, continuity index, power, specificity of centrality, accessibility, coherence and the degree of node [26].

3.2 Normalization of gene expression data

Normalization is an important step with a considerable effect on the analysis of the data among the gens. In terms of gene expression, RNA-seq is a type of technology that uses the "next generation sequencing" technique to obtain a general overview of the amount of RNA genome in a specific time period [27]. RNA-seq normalization methods provide comparisons of gene expression differences between samples. Therefore, in order to determine the most appropriate methods for normalizing gene expression data, a comparison is made between some common methods in this field.

- **In-sample and between-sample normalization methods:** These methods show the modification of the expression level in each gene related to other genes in the same sample. The most common methods in this field are RPKM [28] and FPKM [29].
- **In-sample normalization methods:** The changes in the count of reading a gene between samples are due to differences in the depth of the sequence, so in-sample normalization uses raw readings. The simplest method of normalization in this field is TC.
- **Global normalization methods:** Since the diversity among the genes of a sample and the variations of each gene throughout the samples must be modified, two methods, Med-pgQ2 and UQ-pgQ2 were introduced [27]. In Med-pgQ2, the new normalized count $Y_{gj}^{Med-pgQ2}$ for each gene and every 100 readings is defined as Eq. (2).

$$Y_{gj}^{Med-pgQ2} = \frac{Y_{gj}^{Med}}{Q2_g^{Med}} \times 100 \quad (2)$$

Where, Y_{gj}^{Med} is the expression value for the g gene in sample j and also $Q2_g^{Med}$, is the middle gen of the g after normalization of each sample.

In UQ-pgQ2, it is assumed that Y_{gj}^{UQ} are the expression values for the g gene in sample j and are normalized by UQ (75%); It is also supposed that $Q2_g^{UQ}$ is the middle g gene in the samples which is normalized after UQ. Thus, the new normalized count $Y_{gj}^{UQ-pgQ2}$ for each gene and per 100 readings is defined as Eq. (3).

$$Y_{gj}^{UQ-pgQ2} = \frac{Y_{gj}^{UQ}}{Q2_g^{UQ}} \times 100 \quad (3)$$

- **Scalable normalization methods:** These methods are used to calculate the cover scales as well as to normalize the gene expression values of very large cells. The most common methods in this field are RLE

and TMM [30]. In RLE, first, the average for each sample is calculated, for each j gene, namely $med(y_{*j})$. Where, y_{*j} is the value of the j-th column of the matrix $[y_{ij}]$. Then the deviation value is calculated from the mean $y_{ij} - med(y_{*j})$, so that the y_{ij} expressed logarithm for each j-th gene is from the ith sample.

In order to evaluate the performance of RNA-seq normalization methods, the analysis and evaluation of the differences in the expressed genes are examined using AUC values. Table 1 shows AUC for different methods according to the z-test of two one-way samples.

Table 1. Performance of IEC-MLP with and without features selection

Methods	$\boldsymbol{Z}$-statistics	$\boldsymbol{P}$-value
RLE	0.72	0.232
UQ-pgQ2	0.76	0.225
FPKM	2.00	0.022
TMM	2.01	0.022
DESeq	2.58	0.004
FQ	2.69	0.004
TC	2.75	0.003

In this table, for each method, a list of the results of p values on the FANTOM5 dataset are presented [31]. The results show that AUC value in RLE method is superior to other methods.

3.3 Wilcoxon test

The Wilcoxon test [32] is a non-parametric statistical test that is used to assess the similarity of the two samples dependent on the ranking scale. Non-parametric sign tests, McNemar's and Wilcoxon tests are used to compare and detect pairwise semantic relations between the samples.

The sign test and Wilcoxon require high-level variables. The Wilcoxon test has one advantage over the sign test, and that is it shows a significant difference between the two samples. McNemar's test is not applicable for the variables that have two levels and for the variables that have more than two levels. Therefore, since the purpose of this study is to investigate the semantic differences of multivariate, Wilcoxon test is used. One of the requirements for performing this test is the relevance of the data as well as having an order scale for the values of the pairs so that the difference in the pairs can be calculated.

To calculate the Wilcoxon test, we assume that an n-sized sample is available with paired data; therefore $2n$ data is available. For pairs $X_{1,i}$ and $X_{2,i}$ as $i = 1,2, \dots, n$, zero and one hypotheses are suggested. Null hypothesis; the difference between pairs with a symmetrical distribution is considered around zero and one; and expresses the difference between pairs with a symmetrical distribution around zero. Here for all pair sizes, the value is calculated as $|X_{2,i} - X_{1,i}|$ and the sign difference is also recorded. In the next step, we delete all the zero differences and call the dimensions of the new sample n_r. The data is then sorted into a new sample from the smallest to the largest absolute value and the data is ranked as (the smallest rank one). This rank is indicated by the variable R_i. Finally, the Wilcoxon test is defined according to Eq. (4).

$$W = \sum_{i=1}^{n_r} [sign(X_{2,i} - X_{1,i}) \times R_i] \tag{4}$$

Where, the null hypothesis is rejected, if $|W| > W_{critical,n_r}$.

3.4 Silhouette index

Due to the lack of a target class in female expression data, there is a need for internal validation indicators to measure the accuracy of clustering results. In this research, the internal silhouette index [33] is used for this purpose. This criterion calculates the evaluation of clusters using the internal values of each cluster and their appearance.

The silhouette index is based on the calculation of cluster validity based on the difference in the distances between and within the cluster and offers a combination of in-cluster and inter-cluster similarities. The average of this index can take a value in the range of $[-1, +1]$. If the mean of the index is close to +1, then the clustering model is considered satisfactory. Negative values close to zero indicate the inadequacy of the model and poor performance of the clustering algorithm in creating clusters. This index is calculated for a data sample such as x_i in three steps.

Step 1: Calculate the average distance of data x_i from all other data in its cluster (a_i).

Step 2: Calculate the average distance x_i data from all other data in another cluster. The lowest value obtained from the $k-1$ cluster specifies the average distance calculated for selection (b_i).

Step 3: Calculate the silhouette coefficient with Eq. (5).

$$S_i = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{5}$$

Where, a_i and b_i represent the mean distance between observation i and other observations in a similar cluster, respectively, and the mean observation distance i to all observations in other clusters. In order to check the appropriateness of a clustering method, the average S_i is calculated for all data.

4 EXTRACTION OF CELL-TO-CELL COMMUNICATION

Gene expression clustering techniques allow thousands or even more genes to be placed into smaller bunches. One of the features of gene expression clustering is the definition of measuring the similarity (for example, distance) between gene expression characteristics [34]. In this study, Wilcoxon method is used to obtain inter-cellular communication, which is based on gene expression data. A proper clustering method plays a key role in obtaining communication between cells.

In the previous work, the detection of inter-cellular communication in different diseases was performed according to the characteristics of the topological structure of the graph and an improved cumulative clustering method. The previous method had two steps; In the first stage, several clustering models were combined to detect the initial communication between cells in order to produce better results than individual algorithms, and in the second stage, the similarity between cells in each cluster was calculated using a similarity criterion based on the topological structure of the graph. The present study uses the efficiency of a graph-based algorithm to detect inter-cellular communication extracted by Wilcoxon test. The flowchart of the proposed method is shown in Figure 1.

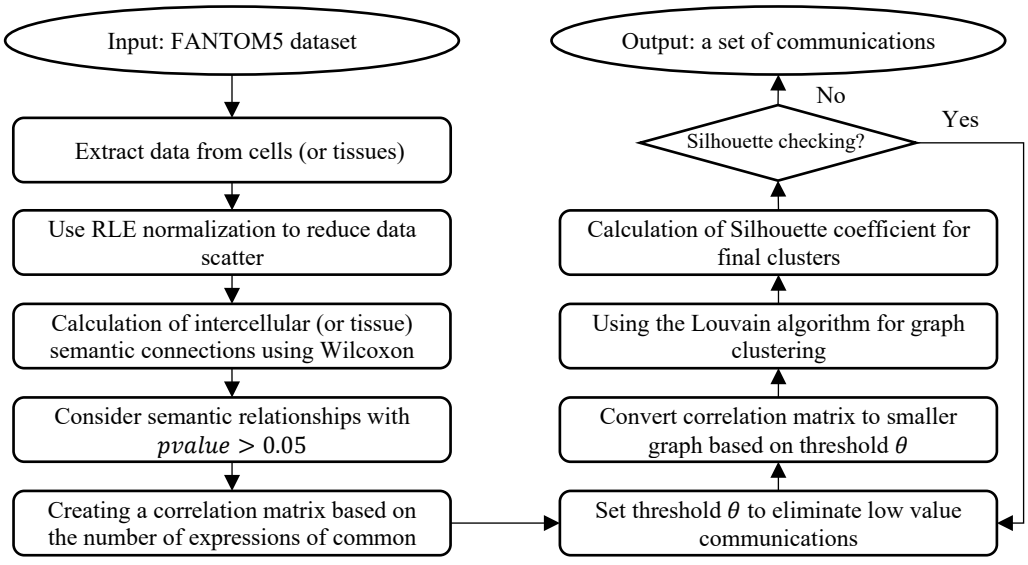


Figure 1. Flowchart of the proposed method

Given the distribution of values in the FANTOM5 dataset (information about this dataset is available at http://fantom.gsc.riken.jp/5), we must first normalize the data. The purpose of normalization is to prevent data scatter by placing data values in a specified range. There are various methods and models for normalization, in [35]

the RLE method is used to normalize the data similar to the FANTOM5 data set and good results are reported. In the present study, this method is used for normalization.

In the FANTOM5 dataset, the lines representing the expression numbers of genes and columns represent different cell samples extracted from multiple human samples, so that there may be several samples from one cell. Table 2 shows a schematic of the FANTOM5 dataset with 108 samples and 86428 genes for cells.

Table 2. The schematic plan of the FANTOM5 dataset for cells

	Cell 1					**…**		**Cell 108**			
	Sample 1	***Sample 2***	**...**	***Sample N***		**...**		***Sample 1***	***Sample 2***	**...**	***Sample N***
Gen 1	Gene expression	Gene expression	...	Gene expression		...		Gene expression	Gene expression	...	Gene expression
⋮	⋮	⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮
Gen 86428	Gene expression	Gene expression	...	Gene expression		...		Gene expression	Gene expression	...	Gene expression

After normalizing the data, the correlation matrix between the columns should be calculated to detect inter-cellular communication. At least 2 samples are required to obtain a correlation matrix from each cell, therefore, one-time observed samples in the data set are not considered. Wilcoxon method with p-value greater than 0.05 was used to obtain inter-cellular communication. The output of the correlation matrix shows the number of expressions of common genes between both cell samples. In fact, the Wilcoxon method and the p-value value calculate the semantic communication of each cell pair for all genes. Table 3 shows a diagram of the output of the correlation matrix for cells.

Table 3. Correlation matrix output for cells

	Cell 1	**Cell 2**	**...**	**Cell 108**
Cell 1	Gene expression	Gene expression	...	Gene expression
Cell 2	Gene expression	Gene expression	...	Gene expression
⋮	⋮	⋮	⋮	⋮
Cell 108	Gene expression	Gene expression	...	Gene expression

For example, if cell number 1 is shared with cell number 2 in 50 gene expressions, the matrix value for these two cells (corresponding number of genes expressed) is 50. In general, Table 3 form the graphs of cell-to-cell communication, so that the nodes represent the cells and the edges represent the weights between them.

Given that the graph created in the representation section is complete, it is clear that the weight of all edges does not affect the clustering of cells, and their presence may reduce the efficiency of the proposed method. For this purpose, before running the community detection algorithm, by applying a threshold to the edges of the graph, we remove those edges that weigh less than the threshold θ.

In the next step, we use a graph-based clustering method to detect inter-cellular and interstitial communication. By doing this, we put cells in a community (or a cluster) and extract communication. In this way, using this communication, it can be checked that; first of all, in which communication genes are expressed in the same way, second, what is the expression of genes in several cells, and third, what is the expression of genes in a particular cell.

The goal of clustering here is to place similar cells in the same clusters. Most graph clustering methods have problems and disadvantages [36]. In most methods, the parameter k (number of clusters) must be specified by the user before the algorithm is executed. On the other hand, the distribution of data in each cluster is one of the important criteria in clustering that has not been considered in most of the previous methods. Considering the scattering rate of features in each cluster increases the performance of the clustering algorithm. In this study, to solve this problem and to solve the above problem, the Louvain graph-based clustering algorithm is used [4]. Louvain is a greedy algorithm that tries to maximize the measurement criterion in a graph. Unlike all other clustering methods, the input graph size limit does not depend on memory as well as processing time. For this reason, this algorithm can be easily applied and distributed for the graphs with hundreds of millions of nodes.

In benchmark networks with heterogeneous distributions of cluster sizes, the simultaneous elimination of both biases is not possible and multiresolution modularity is not capable to recover the planted community structure, not even when it is pronounced and easily detectable by other methods, for any value of the resolution parameter. This holds for other multiresolution techniques and it is likely to be a general problem of methods based on global optimization. However, in the studied data set, most communities are large in size and this modularity constraint has little effect on the proposed method.

Modularity function optimization is a widely used method for identifying communities. Modularity is defined in the above algorithm as the optimized value between $[-1, +1]$, which measures the density of links within communities relative to the communication between communities. Eq. (6) shows the modulus function.

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{6}$$

Where, A_{ij} is the sum of the weights of the edges between two nodes i and j, k_i and k_j are the sum of the weights of the edges connected to nodes i and j. Also, m is the sum of all the weights of the edges in the graph; c_i and c_j are the communication of nodes i and j. In addition, δ is a simple delta function according to Eq. (7).

$$\delta_{ij} = \begin{cases} 0 & if\ i \neq j \\ 1 & if\ i = j \end{cases} \tag{7}$$

Complete optimization of the modularity function with the Louvain clustering algorithm is done in two steps:

1. With local optimization, the algorithm looks for small groups. For node $\boldsymbol{i}$, the allocation benefit to cluster $\boldsymbol{C}$ is calculated by Eq. (8).

$$AQ = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{8}$$

Where, Σ_{in} is the sum of the weights in cluster C, Σ_{tot} is the sum of the weights of the edges connected to the nodes of cluster C, k_i is the sum of the edges of node i and $k_{i,in}$ represents the sum of the weights of one end of the node which is i on one side and is the C cluster on the other side. Also, m is the sum of the weights of all the edges of the graph.

2. It then continues clustering by merging small groups that have the ability to form larger groups. The steps are repeated until there is no change in the clusters and the modulus is maximized.

Due to the integration of small clusters, the Louvain algorithm automatically detects the number of final clusters as the modulus value increases. Therefore, the number of final clusters is equal to the number of clusters with the maximum modulus value. The graph forms cell-to-cell or tissue-to-tissue relationships so that the nodes represent the cells and the edges represent the weight (number of genes equivalent to each row of the data set) between the nodes. Fig. 2 shows the pseudo code of the proposed method for extracting inter-cellular communication.

Cell-to-Cell Communication Extraction Algorithm

- **Input:** FANTOM5 dataset
- **Output:** Cells clustering

1: Apply filters to dataset (Remove cells with only one sample and Lines without $GeneID$).
2: Use Relative Log Expression (RLE) to normalize FANTOM5 dataset.
3: **for** i = 1 to $nCells$ **do**
4: **for** j = 1 to $nCells$ **do**
5: **if** $i \neq j$ **then**
6: $GraphCommunication\ (i, j)$ = Use the Wilcoxon test to calculate Communication between i and j cells, with p-value > 0.05.
7: **else**
8: $GraphCommunication(i, j) = 0$
9: **end if**
10: **end for**
11: **end for**
12: Apply Louvain aalgorithm ($Graph\ Communication$) clustering method to detect Communication with the highest Gene Expression.

Figure 2. Pseudo-code of the proposed algorithm

In line 1, the FANTOM5 data set filtering process is performed and cells with only one sample as well as expression of genes without $GeneID$ are deleted. In line 2, the FANTOM5 dataset is normalized using the RLE method. In line 3, variable i is repeated for the number of cells. Line 4, the variable j is repeated in the same way as the number of cells. This is to identify the communication of each cell pair. Line 5 shows whether the two cells i and j are similar or not. If they are not the same in line 6, the semantic communication between them is calculated using the Wilcoxon statistical technique. Here, communication with a p-value > 0.05 are considered. Lines 7 and 8 show that two similar cells have 0 communication. Lines 9, 10 and 11 are the end of the repeating loops. In line 12, the Louvain clustering algorithm is applied based on p-value > 0.05 and clusters the cells.

5 RESULTS AND DISCUSSION

In this section, the performance of the proposed algorithm for detecting communication between cells on a real FANTOM5 dataset is evaluated. In the FANTOM5 dataset, there are 1829 samples with 201802 promoters (gene expression). The features in the FANTOM5 dataset are promoters that actually contain information about the genes that lead to its production. The purpose of extracting this communication is to identify cells that have the same gene expression in one or more diseases.

In this dataset, samples (columns) include the names of cells from different patients. Promoters (rows) represent the gene expression number that is specified using Entre $GeneID$. Some of the Entre $GeneID$ values in the original dataset are valueless and are marked with NA, so these rows are deleted. Since our aim is to obtain an inter-cellular communication, only the columns that are relevant to the cell are considered. Here 702 columns related to cells were identified. Specifically, several samples may be taken from one cell. Table 4 summarizes the FANTOM5 dataset information.

Table 4. Summary of FANTOM5 dataset

FANTOM5 dataset	#Samples	#Promoters	#Cells
Original dataset	1829	201802	498
Dataset after filtering	125	86427	108

Due to the lack of a target class, we need internal validation indicators to measure the accuracy of clustering results. The purpose of cluster validation is to find clusters that best fit the data in question. The data belonging to a cluster should be as close to each other as possible (density criterion). A common criterion for determining data density is data variance. Also, the clusters themselves must be sufficiently separated from each other (separation criterion). In this paper, the silhouette index has been used for this purpose.

Table 5. Comparison of clustering results with different thresholds

Methods	Threshold ($\boldsymbol{\theta}$)	0.05	0.1	0.15	0.2	0.25	0.3	0.35
Binary clustering [36]	No. of clusters	16	19	19	12	10	12	8
	Silhouette index	0.643	0.758	0.601	0.754	0.699	0.435	0.381
	Run time (s)	671	627	601	587	577	560	547
Ensemble clustering [37]	No. of clusters	8	8	10	9	11	6	4
	Silhouette index	0.565	0.561	0.589	0.698	0.678	-0.701	-0.789
	Run time (s)	2067	1895	1806	1772	1715	1682	1675
Similarity criterion based on the graph topological [19]	No. of clusters	11	9	9	13	13	14	15
	Silhouette index	0.520	0.619	0.604	0.723	0.762	0.700	0.737
	Run time (s)	1865	2152	1932	1806	1816	2006	1765
Proposed method	No. of clusters	8	8	8	9	9	5	2
	Silhouette index	0.592	0.592	0.618	0.814	0.748	-0.702	-0.956
	Run time (s)	513	506	489	468	450	441	436

The cell communication occurs when the number of promoters in a number of cells is significantly expressed. Analysis of the results shows that the values of gene expression with different thresholds produce different values

in the silhouette coefficient. For this purpose, the results are examined with different thresholds. Also, for better analysis, the results are compared with previous research (binary and cumulative clustering) on the criteria of the number of clusters, silhouette and execution time. The results of applying the proposed method for extracting inter-cellular communications with the optimal threshold value $\theta = 0.2$ are shown in Table 5 for 108 cells in 9 clusters. Due to the use of Louvain for clustering, the number of clusters is determined automatically. So according to the threshold setting we can also have the number of clusters.

According to the test results, the threshold of 0.2 has an optimal value with a silhouette coefficient of 0.814 for cells. This is while in binding and cumulative clustering methods, the best silhouette coefficient is 0.698 and 0.758, respectively. The results of graph-based clustering (proposed method) with the optimal threshold parameter are shown in Table 6. The symbol C_i represents the i-th cell, whose names are given in Appendix 1.

According to the obtained results, the highest amount of inter-cellular communication is related to hes3.gfp.embryonic.stem.cells (C_{82}) and cd14.cd16..monocytes.2 (C_{29}) cells with 64580 gene expression. In the second and third ranks, respectively, cd14.monocytes (C_{17}) with ciliary.epithelial.cells (C_{49}) and basophils (C_{15}) with cd14.monocytes (C_{17}) have the highest communication of gene expression. This number of gene expressions indicates the similar behavior of these two cells in exposure to different diseases. This information can help extract patterns of behaviour from a particular virus. In general, most of the communication are expressed in the ABLIM1 gene, followed by the TACC2 and KIAA1217 genes.

Table 6. Clustering results with optimal threshold

Clusters	Samples (Cells)
Cluster 1	C_{10}
Cluster 2	C_{11}
Cluster 3	C_{46}
Cluster 4	C_{45}
Cluster 5	C_{30}, C_{31}, C_{62}
Cluster 6	C_{3}, C_{9}, C_{34}-C_{37}, C_{40}, C_{44}, C_{80}, C_{84}, C_{85}, C_{90}, C_{98}
Cluster 7	C_{4}, C_{12}-C_{15}, C_{17}, C_{19}-C_{28}, $_{33}$, C_{38}, C_{39}, C_{41}, C_{42}, C_{49}, C_{52}, C_{58}, C_{78}, C_{81}, C_{91}, C_{96}, C_{102}, C_{103}
Cluster 8	C_{1}, C_{6}, C_{47}, C_{63}-C_{66}, C_{68}, C_{72}, C_{74}-C_{76}, C_{79}, C_{94}, C_{99}, C_{100}, C_{107}, C_{108}
Cluster 9	C_{2}, C_{5}, C_{7}, C_{8}, C_{16}, C_{18}, C_{29}, C_{32}, C_{43}, C_{48}, C_{50}, C_{51}, C_{53}-C_{57}, C_{59}-C_{61}, C_{67}, C_{69}-C_{71}, C_{73}, C_{77}, C_{82}, C_{83}, C_{86}-C_{89}, C_{92}, C_{93}, C_{95}, C_{97}, C_{101}, C_{104}-C_{106}

6 Conclusion

Relations between cells will help identify different diseases and their causes. In fact, cell communication indicates hereditary communication among patients. These communication help identify common areas of the body that are affected by various diseases. In this study, the detection of inter-cellular communication in different diseases with the combination of RLE normalization, Wilcoxon method and Louvain graph-based clustering algorithm are presented. Evaluation of the performance of the proposed clustering algorithm with silhouette index has proved its high accuracy. The Wilcoxon method with a p-value greater than 0.05 is used to obtain inter-cellular communication. The output of the correlation matrix shows the number of expressions of common genes between both cell samples. This matrix is a weighted matrix that is expressed as a complete graph. The Louvain algorithm uses a greedy method of modulation and extracts the final inter-cellular communication through clustering. The proposed method is tested on the FANTOM5 dataset.

The results show that on average, the proposed clustering algorithm based on the silhouette index well detects the communication between cells. One of the advantages of the proposed method is reducing the volume of gene expression data by mapping them in graphs. This approach uses graph-based clustering algorithms based on gene expression data. Instead of calculating the similarity between objects by a threshold, several different thresholds were used, and finally a threshold of 0.2 for cells were obtained as the best results with respect to the silhouette index. In addition, cell-to-cell communication in 9 clusters was reported to be optimal for 108 cells.

APPENDIX 1: Full cell names (C_i) in Table 6.

C_1: x293slam.rinderpest.infection, C_2: arpe.19.emt.induced.with.tgf.beta.and.tnf.alpha, C_3: adipocyte...breast, C_4: adipocyte...omental, C_5: adipocyte...subcutaneous, C_6: adipocyte.differentiation, C_7: alveolar.epithelial.cells, C_8: amniotic.epithelial.cells, C_9: anulus.pulposus.cell, C_{10}: aortic.smooth.muscle.cell.response.to.fgf2, C_{11}: aortic.smooth.muscle.cell.response.to.il1b, C_{12}: astrocyte...cerebellum, C_{13}: astrocyte...cerebral.cortex, C_{14}: b.lymphoblastoid.cell.line..gm12878.encode, C_{15}: basophils, C_{16}: bronchial.epithelial.cell, C_{17}: cd14..monocytes, C_{18}: cd14..monocyte.derived.endothelial.progenitor.cells, C_{19}: cd14..monocytes...mock.treated, C_{20}: cd14..monocytes...treated.with.b.glucan, C_{21}: cd14..monocytes...treated.with.bcg, C_{22}: cd14..monocytes...treated.with.candida, C_{23}: cd14..monocytes...treated.with.cryptococcus, C_{24}: cd14..monocytes...treated.with.group.a.streptococci, C_{25}: cd14..monocytes...treated.with.ifn...n.hexane, C_{26}: cd14..monocytes...treated.with.salmonella, C_{27}: cd14..monocytes...treated.with.trehalose.dimycolate..tdm., C_{28}: cd14..monocytes...treated.with.lipopolysaccharide, C_{29}: cd14.cd16..monocytes, C_{30}: cd14.cd16..monocytes.1, C_{31}: cd14.cd16..monocytes.2, C_{32}: cd19..b.cells..pluriselect., C_{33}: cd19..b.cells, C_{34}: cd34.cells.differentiated.to.erythrocyte.lineage, C_{35}: cd34..progenitors, C_{36}: cd34..stem.cells...adult.bone.marrow.derived, C_{37}: cd4..t.cells, C_{38}: cd4.cd25.cd45ra..naive.regulatory.t.cells.expanded, C_{39}: cd4.cd25.cd45ra..naive.regulatory.t.cells, C_{40}: cd4.cd25.cd45ra..memory.regulatory.t.cells.expanded, C_{41}: cd4.cd25.cd45ra..memory.conventional.t.cells.expanded, C_{42}: cd4.cd25.cd45ra..memory.conventional.t.cells, C_{43}: cd8..t.cells..Pluriselect, C_{44}: cd8..t.cells, C_{45}: cobl.a.rinderpest.infection, C_{46}: cobl.a.rinderpest..c..infection, C_{47}: cardiac.myocyte, C_{48}: chondrocyte...de.diff, C_{49}: ciliary.epithelial.cells, C_{50}: corneal.epithelial.cells, C_{51}: dendritic.cells...monocyte.immature.derived, C_{52}: dendritic.cells...plasmacytoid, C_{53}: endothelial.cells...aortic, C_{54}: endothelial.cells...artery, C_{55}: endothelial.cells...lymphatic, C_{56}: endothelial.cells...microvascular, C_{57}: endothelial.cells...thoracic, C_{58}: endothelial.cells...umbilical.vein, C_{59}: endothelial.cells...vein, C_{60}: eosinophils, C_{61}: esophageal.epithelial.cells, C_{62}: fibroblast...aortic.adventitial.donor2...cytoplasmic.fraction. C63: fibroblast...aortic.adventitial, C_{64}: fibroblast...cardiac, C_{65}: fibroblast...choroid.plexus, C_{66}: fibroblast...conjunctival, C_{67}: fibroblast...dermal, C_{68}: fibroblast...gingival, C_{69}: fibroblast...lung, C_{70}: fibroblast...lymphatic, C_{71}: fibroblast...mammary, C_{72}: fibroblast...periodontal.ligament, C_{73}: fibroblast...villous.mesenchymal, C_{74}: fibroblast...skin.dystrophia.myotonica, C_{75}: fibroblast...skin.normal, C_{76}: fibroblast...skin.spinal.muscular.atrophy, C_{77}: fibroblast...skin, C_{78}: gingival.epithelial.cells, C_{79}: h1.embryonic.stem.cells.differentiation.to.cd34..Hsc, C_{80}: h9.embryoid.body.cells, C_{81}: h9.embryonic.stem.cells, C_{82}: hes3.gfp.embryonic.stem.cells, C_{83}: hair.follicle.dermal.papilla.cells, C_{84}: hair.follicle.outer.root.sheath.cells, C_{85}: hep.2.cells.mock.treated, C_{86}: hep.2.cells.treated.with.streptococci.strain.5448, C_{87}: hep.2.cells.treated.with.streptococci.strain.jrs4, C_{88}: hepatic.sinusoidal.endothelial.cells, C_{89}: hepatic.stellate.cells..lipocyte., C_{90}: hepatocyte, C_{91}: k562.erythroblastic.leukemia.response.to.hemin, C_{92}: keratinocyte...epidermal, C_{93}: keratocytes, C_{94}: lens.epithelial.cells, C_{95}: lymphatic.endothelial.cells.response.to.vegfc, C_{96}: mcf7.breast.cancer.cell.line.response.to.egf1, C_{97}: mcf7.breast.cancer.cell.line.response.to.hrg, C_{98}: macrophage...monocyte.derived, C_{99}: mast.cell, C_{100}: melanocyte...dark, C_{101}: melanocyte...light, C_{102}: melanocyte, C_{103}: meningeal.cells, C_{104}: mesenchymal.stem.cells...adipose, C_{105}: mesenchymal.stem.cells...bone.marrow, C_{106}: mesenchymal.stem.cells...umbilical, C_{107}: mesothelial.cells, C_{108}: monocyte.derived.macrophages.response.to.lps.

Acknowledgement

Fund Project: Xuzhou Science and Technology Innovation Project. Project Name: Effect and mechanism of GDNF-mediated fusion gene BCL2L2-PABpN1 on anti-apoptosis in glioma development, Project No.: KC19061, establishment time: July, 2019.

REFERENCES

[1] C. Pizzuti and S. E. Rombo, "Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods," Bioinformatics, 30(10):1343–1352, 2014.

[2] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Mixing local and global information for community detection in large networks," Journal of Computer and System Sciences, 80(1):72–87, 2014.

[3] J. Xie and B. K. Szymanski, "Towards linear time overlapping community detection in social networks," Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, pages 25–36. Springer, 2012.

[4] N. Ozaki, H. Tezuka, and M. Inaba, "A Simple Acceleration Method for the Louvain Algorithm," Journal of Computer and Electrical Engineering, 8(3):207–218, 2016.

[5] K. Macropol, T. Can, and A. K. Singh, "RRW: Repeated random walks on genome-scale protein networks for local cluster discovery," BMC Bioinformatics, 10(1):1–10, 2009.

[6] T. Qin and K. Rohe, "Regularized spectral clustering under the degree-corrected stochastic blockmodel," International Conference on Neural Information Processing Systems, 3:3120–3128, 2013.

[7] M. S. Aslanpour, S. E. Dashti, M. Ghobaei-Arani, and A. A. Rahmanian, "Resource provisioning for cloud applications: a 3-D, provident and flexible approach" The Journal of Supercomputing, 74(12):6470–6501, 2018.

[8] M. Ghobaei-Arani, and A. Shahidinejad, "An efficient resource provisioning approach for analyzing cloud workloads: a metaheuristic-based clustering approach," The Journal of Supercomputing, 77(1):711–750, 2021.

[9] G. J. McLachlan, R. W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data.," Bioinformatics, 18(3):413–422, 2002.

[10] S. Forouzandeh, M. Rostami, and K. Berahmand, "Presentation a Trust Walker for rating prediction in recommender system with Biased Random Walk: Effects of H-index centrality, similarity in items and friends," Engineering Applications of Artificial Intelligence, 104:104325, 2021.

[11] M. Rostami, K. Berahmand, E. Nasiri, and S. Forouzandeh, "Review of swarm intelligence-based feature selection methods," Engineering Applications of Artificial Intelligence, 100:104210, 2021.

[12] J. Hofbauer and K. Sigmund, "Evolutionary game dynamics," Bulletin of the American Mathematical Society, 40(4):479–519, 2003.

[13] S. Talatian Azad, G. Ahmadi, and A. Rezaeipanah, "An intelligent ensemble classification method based on multi-layer perceptron neural network and evolutionary algorithms for breast cancer diagnosis," Journal of Experimental & Theoretical Artificial Intelligence, in press, 2021.

[14] A. Rezaeipanah, and G. Ahmadi, "Breast Cancer Diagnosis Using Multi-Stage Weights Adjustment in the MLP Neural Network," The Computer Journal, in press, 2020.

[15] D. M. Lane and A. Sándor, "Designing Better Graphs by Including Distributional Information and Integrating Words, Numbers, and Images," Psychological Methods, 14(3):239–257, 2009.

[16] D. Chen, Y., Kamath, G., Suh, C., & Tse, "Community recovery in graphs with locality," International Conference on Machine Learning, p. 689–698, 2016.

[17] R. Hou, E. Denisenko, H. T. Ong, J. A. Ramilowski, and A. R. Forrest, "Predicting cell-to-cell communication networks using NATMI," Nature communications, 11(1):1–11, 2020.

[18] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin,T. Kasukawa, S. Sahin, and H. Kawaji, "Gateways to the FANTOM5 promoter level mammalian expression atlas," Genome biology, 16(1)1–14, 2015

[19] M. Mojarad, F. Sarhangnia, A. Rezaeipanah, H. Parvin, and S. Nejatian, "Modeling Hereditary Disease Behavior Using an Innovative Similarity Criterion and Ensemble Clustering," Current Bioinformatics, 16(5):749–764, 2021.

[20] S. AlMusawi, M. Ahmed, and A. S. Nateri, "Understanding cell-cell communication and signaling in the colorectal cancer microenvironment," Clinical and Translational Medicine, 11(2):e308, 2021.

[21] A. A. Almet, Z. Cang, S. Jin, and Q. Nie, "The landscape of cell–cell communication through single-cell transcriptomics," Current opinion in systems biology, 26:12–23, 2021.

[22] R. C. Paolicelli, G. Bergamini, and L. Rajendran, "Cell-to-cell communication by extracellular vesicles: focus on microglia," Neuroscience, 405:148–157, 2019.

[23] S. Fortunato, "Community detection in graphs," Physics Reportsm, 486(3-5):75–174, 2010.

[24] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," Physical review E, 74(3):036104, 2006.

[25] S. U. Rehman, A. U. Khan, and S. Fong, "Graph mining: A survey of graph mining techniques," Seventh International Conference on Digital Information Management, Macau, Macao, pages 88–92. IEEE, 2012.

[26] W. Maharani and A. A. Gozali, "Collaborative Social Network Analysis and Content-based Approach to Improve the Marketing Strategy of SMEs in Indonesia," Procedia Computer Science, 59:373–381, 2015.

[27] K. H. Li, P., Piao, Y., Shon, H. S., & Ryu, "Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data," BMC Bioinformatics, 16:347, 2015.

[28] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," Nature methods, 5(7):621–628, 2008.

[29] C. Trapnell et al., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," Nature biotechnology, 28(5):511–515, 2010.

[30] L. C. Gandolfo and T. P. Speed, "RLE plots: Visualizing unwanted variation in high dimensional data," PLoS One, 13(2):e0191629, 2018.

[31] A. R. R. Forrest et al., “A promoter-level mammalian expression atlas,” Nature, 507(7493):462–470, 2014.

[32] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” In Breakthroughs in statistics, New York, NY, pages 196–202. Springer, 1945.

[33] S. Aranganayagi and K. Thangavel, “Clustering categorical data using Silhouette coefficient as a relocating measure,” in Proceedings - International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 2007, volume 2, pages 13–17. IEEE, 2008.

[34] R. Syah, S. Wulandari, A. Arbansyah, and A. Rezaeipanah, “Design of Ensemble Classifier Model Based on MLP Neural Network For Breast Cancer Diagnosis,” Inteligencia Artificial, 24(67):147–156, 2021.

[35] E. Côme and P. Latouche, “Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood,” Stat. Modelling, 15(6):564–589, 2015.

[36] S. J. Nanda, R. Raman, S. Vijay, and A. Bhardwaj, “A new density based clustering algorithm for Binary Data sets,” International Conference on High Performance Computing and Applications, pages 1–6. IEEE, 2014.

[37] P. Rathore, J. C. Bezdek, S. M. Erfani, S. Rajasegarar and M. Palaniswami, “Ensemble fuzzy clustering using cumulative aggregation on random projections,” IEEE Transactions on Fuzzy Systems, 26(3):1510–1524, 2017.