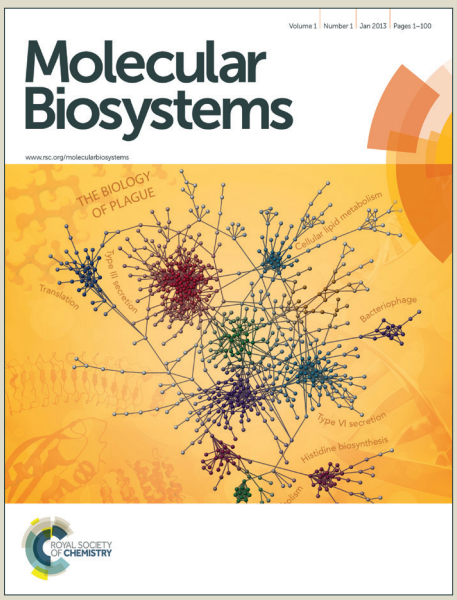
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

ARTICLE TYPE www. rsc.org/xxxxxx | XXXXXXXX

City Block Distance and Rough-Fuzzy Clustering for Identification of Co-Expressed microRNAs†

Sushmita Paul and Pradipta Maji

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

The microRNAs or miRNAs are short, endogenous RNAs having ability to regulate mRNA expression at the post-transcriptional level. Various studies have revealed that miRNAs tend to cluster on chromosomes. The members of a cluster that are at close proximity on chromosome are highly likely to be processed as co-transcribed units. Therefore, a large proportion of miRNAs are co-expressed. Expression profiling of miRNAs generates a huge volume of data. Complicated networks of miRNA-mRNA interaction increase the challenges of comprehending and interpreting the resulting mass of data. In this regard, this paper presents a clustering algorithm in order to extract meaningful information from miRNA expression data. It judiciously integrates the merits of rough sets, fuzzy sets, *c*-means algorithm, and normalized range-normalized city block distance to discover co-expressed miRNA clusters. While the membership functions of fuzzy sets enable efficient handling of overlapping partitions in noisy environment, the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in cluster definition. The city block distance is used to compute the membership functions of fuzzy sets and to find initial partition of a data set, and thereby helps to handle minute differences between two miRNA expression profiles. The effectiveness of the proposed approach, along with a comparison with other related methods, is demonstrated on several miRNA expression data sets using different cluster validity indices. Moreover, the gene ontology is used to analyze the functional consistency and biological significance of generated miRNA clusters.

1 Introduction

MicroRNAs or miRNAs are a class of short approximately 22-nucleotide non-coding RNAs found in many plants and animals. They often act post-transcriptionally to inhibit mRNA expression. Hence, miRNAs are related to diverse cellular processes and regarded as important components of the mRNA regulatory network. Recent genome wide surveys on non-coding RNAs have revealed that a substantial fraction of miRNAs is likely to form clusters. However, the evolutionary and biological function implications of clustered miRNAs are still elusive.

The genes of miRNAs are often organized in clusters in the genome. It has been reported that at a very conservative maximum inter-miRNA distance of 1kb, over 30% of all miRNAs are organized into clusters[1]. Expression analyses showed strong positive correlations among the closely located miRNAs, indicating that they may be controlled by common regulatory element(s). In fact, experimental evidence demonstrated that clustered miRNA loci form an operon-like gene structure and that they are transcribed from common promoter. Existence of co-expressed miRNAs is also demonstrated using expression profiling analysis in[2]. Several miRNA clusters have been experimentally shown by RT-PCR or Northern blotting[3,4]. These findings suggest that members of a miRNA cluster, which are at a close proximity on a chromosome, are highly likely to be processed as co-transcribed units. Expression data of miRNAs can be used to detect clusters of miRNAs as it is suggested that co-expressed miRNAs are co-transcribed, so they should have similar expression pattern.

A miRNA expression data set can be represented by an expression table, where each row corresponds to one particular miRNA, each column to a sample or time point, and each entry of the matrix is the measured expression level of a particular miRNA in a sample or time point, respectively. The complex networks of miRNA-mRNA interaction greatly increase the challenges of comprehending and interpreting the resulting mass of data[5]. A first step towards addressing this challenge is the use of clustering techniques, which is essential in the pattern recognition process to reveal natural structures and identify interesting patterns in the underlying data[6].

Cluster analysis is a technique for finding natural groups present in the miRNA set. It divides a given miRNA set into a set of clusters in such a way that two miRNAs from the same cluster are as similar as possible and the miRNAs from different clusters are as dissimilar as possible[7]. To understand the

Biomedical Imaging and Bioinformatics Lab, and Machine Intelligence Unit, Indian Statistical Institute, Kolkata, 700 108, India. E-mail: {sushmita_t,pmaji}@isical.ac.in

† This work is partially supported by the Indian National Science Academy, New Delhi (grant no. SP/YSP/68/2012).

role of miRNAs in different cellular processes and diseases, and the mechanism of repression of mRNA translation, clustering techniques have proven to be helpful. The co-expressed miRNAs, that is, miRNAs with similar expression patterns and co-transcribed, can be clustered together having similar cellular functions. This approach may further understanding of the functions of many miRNAs for which information has not been previously available[8].

In this background, several authors used hierarchical clustering algorithms[5,9,10] and self organizing maps[11] to group miRNAs having similar function. Other clustering techniques such as *k*-means algorithm[12], graph theoretical approaches[13–16], model based clustering[17–20], and density based approach[21], which have been widely applied to find co-expressed gene clusters, can also be used to group co-expressed miRNAs from microarray data.

However, one of the main problems in expression data analysis is uncertainty. Some of the sources of this uncertainty include imprecision in computations and vagueness in class definition. In this background, the possibility concept introduced by fuzzy sets[22] and rough sets[23] provides a mathematical framework to capture uncertainties associated with human cognition process[6,24]. Also, the empirical study has demonstrated that miRNA expression data are often highly connected, and the clusters may be highly overlapping with each other or even embedded one in another. Moreover, expression data often contains a huge amount of noise due to the complex procedures of microarray experiments. Therefore, fuzzy *c*-means[25] and different rough-fuzzy clustering algorithms such as rough-fuzzy *c*-means[26] can be used to effectively handle these situations and to find co-expressed miRNA clusters.

In general, the quality of generated clusters is always relative to a certain distance measure. Different distance measures may lead to different clustering results. However, every distance measure tries to compute the dissimilarity among miRNAs present in different clusters. Several similarity or dissimilarity measures such as Euclidean distance, Jaccard index, Pearson correlation coefficient, and city block distance (CBD) are used in various clustering algorithms. The performance of a clustering algorithm highly depends on the distance measure used. One of the important properties of the CBD, not shared by Euclidean distance, is dimensional additivity, that is, the total distance is a sum of the distances per dimension. Moreover, the time required to calculate the CBD is less than the time required to calculate the Euclidean distance.

In this regard, the paper presents a rough-fuzzy clustering algorithm, integrating the concepts of lower and upper approximations of rough sets, probabilistic and possibilistic memberships of fuzzy sets, *c*-means algorithm, and normalized range-normalized city block distance (NRNCBD), to discover groups of co-expressed miRNAs from huge miRNA expression data. While the integration of both membership functions of fuzzy sets enables efficient handling of overlapping partitions in noisy environment, the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in cluster definition. Moreover, the use of the NRNCBD helps to handle minute differences between two miRNA expression profiles. Each cluster is represented by a set of three parameters, namely, a cluster prototype or centroid, a possibilistic lower approximation, and a probabilistic boundary. The cluster prototype depends on the weighting average of the possibilistic lower approximation and probabilistic boundary. The NRNCBD is used to calculate both possibilistic and probabilistic membership functions as well as to find initial partition of a data set. The effectiveness of the NRNCBD over Pearson distance and Euclidean distance is presented in this paper. The performance of the proposed miRNA clustering algorithm, along with a comparison with other related methods, is demonstrated on four miRNA expression data sets using standard cluster validity indices. Biological validation of the clustering solutions is also done using gene ontology based analysis.

The rest of this paper is organized as follows: Section 2 reports the miRNA expression data sets used, while Section 3 presents the basic concepts of city block distance, method for selection of initial cluster prototypes, and the proposed rough-fuzzy clustering algorithm. Implementation details, experimental results, discussions, and a comparison among different clustering algorithms are presented in Section 4. Finally, concluding remarks are given in Section 5.

2 Data Sets Used

In this work, publicly available four miRNA expression data sets are used to compare the performance of different clustering methods. This section gives a brief description of the following four miRNA expression data sets, which are downloaded from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/).

1. **GSE16473:** It is the analysis to evaluate the role of miRNAs in skeletal muscle regeneration[27]. Hence, global miRNA expression is measured during muscle cell growth and differentiation. This data set contains 231 miRNAs and 7 time points.

2. **GSE17155:** It is the analysis to test the hypothesis that there is a specific miRNA expression signature which characterizes male breast cancers. The miRNA microarray analysis was performed in a series of male breast cancers and compared them to cases of male gynecomastia and female breast cancers[28]. This data set contains 774 miRNAs and 38 time points.

3. **GSE29495:** The miRNA profiling of kidney tissue from C57BL/6 mice that received a 30 minute ischemic injury compared with control kidney tissue from mice that received sham operation only has been conducted. The number of miRNAs and time points are 574 and 17, respectively.

4. **GSE35074:** It is the analysis to identify miRNAs participating in SNAI1-orchestrated regulatory pathways, a time-resolved microarray data of SNAI1-induced EMT is analyzed, obtained during conditional expression of SNAI1 in a Tet-Off MCF7-SNAI1 breast carcinoma cell model[29]. It contains 837 miRNAs and 21 time points.

3 Proposed Clustering Method

This section describes the proposed miRNA clustering algorithm. It is developed by integrating judiciously rough sets, fuzzy sets, *c*-means algorithm, and the CBD.

3.1 City Block Distance

The CBD, also known as the Manhattan distance or taxi distance, is closely related to the Euclidean distance. Whereas the Euclidean distance corresponds to the length of the shortest path between two points, the CBD is the sum of distances along each dimension. The distance between two objects x_i and x_j is defined as follows:

$$\mathrm{CBD}(x_i,x_j)=\sum_{k=1}^{m}|x_{ik}-x_{jk}| \tag{1}$$

where m is the number of features of the objects x_i and x_j. As for the Euclidean distance, the expression data are subtracted directly from each other, and therefore should be made sure that they are properly normalized. There are many variants of the CBD. The normalized CBD (NCBD) is defined as follows:

$$\mathrm{NCBD}(x_i,x_j)=\frac{1}{m}\sum_{k=1}^{m}|x_{ik}-x_{jk}| \tag{2}$$

while the range-normalized CBD (RNCBD) is defined as follows:

$$\mathrm{RNCBD}(x_i,x_j)=\sum_{k=1}^{m}\left[\frac{|x_{ik}-x_{jk}|}{k_{max}-k_{min}}\right] \tag{3}$$

where k_{max} and k_{min} denote the maximum and minimum values along the kth feature, respectively. On the other hand, the normalized RNCBD (NRNCBD) is defined as follows:

$$\mathscr{N}(x_i,x_j)=\frac{1}{m}\times\mathrm{RNCBD}(x_i,x_j). \tag{4}$$

From the above discussions, following properties can be derived:

1. $0\leq\mathscr{N}(x_i,x_j)\leq 1$.
2. $\mathscr{N}(x_i,x_j)=\mathscr{N}(x_j,x_i)$.
3. $\mathscr{N}(x_i,x_i)=0$.
4. $\mathscr{N}(x_i,x_j)\leq\mathscr{N}(x_i,x_k)+\mathscr{N}(x_k,x_j)$.

The first three axioms are trivial: the first presents the range of the NRNCBD and says that it is always positive. The second says that the NRNCBD from x_i to x_j is the same with that from x_j to x_i; in other words, the measure is symmetric. The third says that the distance is necessarily 0 when two objects are identical. The fourth axiom, called the triangle inequality, may also seem intuitively obvious but is the more difficult one to satisfy.

3.2 Selection of Initial Cluster Prototypes

A limitation of any *c*-means algorithm is that it can only achieve a local optimum solution that depends on the initial choice of the cluster prototypes. Consequently, computing resources may be wasted in that some initial centers get stuck in regions of the input space with a scarcity of data points and may therefore never have the chance to move to new locations where they are needed. To overcome this limitation, the proposed algorithm begins with the selection of c distinct miRNAs from the given miRNA expression data set using the NRNCBD, which enables the algorithm to converge to an optimum or near optimum solutions.

The algorithm starts by computing the NRNCBD between pairs of miRNAs of a given microarray data set. If the NRNCBD $\mathscr{N}(x_i,x_j)$ between two miRNAs x_i and x_j is less than a predefined threshold λ, then they are considered as similar to each other. After computing the NRNCBD, the total number of similar miRNAs for each miRNA x_i is computed. After that, the miRNAs are sorted according to their similarity values. If the miRNA x_i has higher similarity value than another miRNA x_j and they are similar to each other with respect to the threshold λ, then the miRNA x_i is considered as the potential candidate for the set of initial centers and the miRNA x_j is not included in this set. Finally, c initial centers are selected from the reduced set as potential initial centers. Hence, the initialization method helps to identify different dense regions present in the data set. The identified dense regions ultimately lead to discovering natural groups present in the data set. The whole approach is, therefore, data dependent. The main steps for selection of initial miRNAs are as follows:

1. For each miRNA x_i, calculate $\mathscr{N}(x_i,x_j)$ between itself and the miRNA x_j, $\forall_{j=1}^{n}$.

2. Calculate similarity score between two miRNAs x_i and x_j as follows:

$$S(x_i,x_j) = \begin{cases} 1 & \text{if } \mathscr{N}(x_i,x_j) \leq \lambda \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

3. For each miRNA x_i, calculate total number of similar miRNAs of x_i as

$$N(x_i) = \sum_{j=1}^{n} S(x_i,x_j). \quad (6)$$

4. Sort n miRNAs according to their values of $N(x_i)$ such that $N(x_1) > N(x_2) > \cdots > N(x_n)$.

5. If $N(x_i) > N(x_j)$ and $\mathscr{N}(x_i,x_j) \leq \lambda$, then x_j cannot be considered as an initial cluster center, resulting in a reduced set of miRNAs to be considered for c initial cluster centers v_i, $i = 1,2,\cdots,c$.

6. Stop.

3.3 Rough-Fuzzy Clustering

The proposed rough-fuzzy clustering algorithm adds the concepts of fuzzy memberships, both probabilistic and possibilistic, of fuzzy sets, lower and upper approximations of rough sets, and the NRNCBD into c-means algorithm. While the integration of both probabilistic and possibilistic memberships of fuzzy sets enables efficient handling of overlapping clusters in noisy environment, the rough sets deal with uncertainty, vagueness, and incompleteness in cluster definition.

Let $X = \{x_1,\cdots,x_j,\cdots,x_n\}$ be the set of n objects and $V = \{v_1,\cdots,v_i,\cdots,v_c\}$ be the set of c centroids, where $x_j \in \Re^m$ and $v_i \in \Re^m$. Each of the clusters β_i is represented by a cluster center v_i, a lower approximation $\underline{A}(\beta_i)$ and a boundary region $B(\beta_i) = \{\overline{A}(\beta_i) \setminus \underline{A}(\beta_i)\}$, where $\overline{A}(\beta_i)$ denotes the upper approximation of cluster β_i. The proposed clustering algorithm partitions X into c clusters by minimizing the following objective function:

$$J = \begin{cases} \omega\mathscr{A}_1 + (1-\omega)\mathscr{B}_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset,\ B(\beta_i) \neq \emptyset \\ \mathscr{A}_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset,\ B(\beta_i) = \emptyset \\ \mathscr{B}_1 & \text{if } \underline{A}(\beta_i) = \emptyset,\ B(\beta_i) \neq \emptyset \end{cases} \quad (7)$$

$$\text{where } \mathscr{A}_1 = \sum_{i=1}^{c} \sum_{x_j \in \underline{A}(\beta_i)} (v_{ij})^{\acute{m}_2} \mathscr{N}(v_i,x_j) + \sum_{i=1}^{c} \eta_i \sum_{x_j \in \underline{A}(\beta_i)} (1-v_{ij})^{\acute{m}_2}; \quad (8)$$

$$\text{and } \mathscr{B}_1 = \sum_{i=1}^{c} \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{\acute{m}_1} \mathscr{N}(v_i,x_j). \quad (9)$$

The parameters ω and $(1-\omega)$ correspond to the relative importance of lower and boundary regions, respectively. Hence, to have the clusters and the centroids a greater degree of freedom to move, $0 < (1-\omega) < \omega < 1$. The parameters $\acute{m}_1 \in [1,\infty)$ and $\acute{m}_2 \in [1,\infty)$ are the probabilistic and possibilistic fuzzifiers, respectively. Note that $\mu_{ij} \in [0,1]$ is the probabilistic membership function as that in fuzzy c-means[25] and $v_{ij} \in [0,1]$ represents the possibilistic membership function that has the same interpretation of typicality as in possibilistic c-means[30].

In the proposed rough-fuzzy clustering algorithm, each cluster is represented by a centroid, a possibilistic lower approximation, and a probabilistic boundary. The lower approximation influences the fuzziness of final partition. According to the definitions of lower approximation and boundary of rough sets[23], if an object $x_j \in \underline{A}(\beta_i)$, then $x_j \notin \underline{A}(\beta_k), \forall k \neq i$, and $x_j \notin B(\beta_i), \forall i$. That is, the object x_j is contained in β_i definitely. Hence, the memberships of the objects in lower approximation of a cluster should be independent of other centroids and clusters. Also, the objects in lower approximation should have different influence on the corresponding centroid and cluster. From the standpoint of "compatibility with the cluster prototype", the membership of an object in the lower approximation of a cluster should be determined solely by how far it is from the prototype of the cluster, and should not be coupled with its location with respect to other clusters. As the possibilistic membership v_{ij} depends only on the distance of object x_j from cluster β_i, it allows optimal membership solutions to lie in the entire unit hypercube rather than restricting them to the hyperplane given by equation (11). On the other hand, if $x_j \in B(\beta_i)$, then the object x_j possibly belongs to cluster β_i and potentially belongs to another cluster. Hence, the objects in boundary regions should have different influence on the centroids and clusters, and their memberships should depend on the positions of all cluster centroids. So, in the proposed clustering algorithm, the membership function of the object in lower approximation is given by equation (12), which is identical to possibilistic c-means, while that of boundary region is given by equation (10), which is same as fuzzy c-means. Solving equation (7) with respect to μ_{ij} and v_{ij}, we get

$$\mu_{ij} = \left[\sum_{k=1}^{c} \left(\frac{\mathscr{N}(v_i,x_j)}{\mathscr{N}(v_k,x_j)} \right)^{\frac{1}{\acute{m}_1 - 1}} \right]^{-1}; \quad (10)$$

$$\text{subject to } \sum_{i=1}^{c} \mu_{ij} = 1, \forall j, \text{ and } 0 < \sum_{j=1}^{n} \mu_{ij} < n, \forall i, \quad (11)$$

$$v_{ij} = \left[1 + \left\{ \frac{\mathscr{N}(v_i,x_j)}{\eta_i} \right\}^{\frac{1}{(\acute{m}_2 - 1)}} \right]^{-1}; \quad (12)$$

$$\text{subject to } 0 < \sum_{j=1}^{n} v_{ij} \leq n, \forall i;\ \max_{i}\{v_{ij}\} > 0, \forall j; \text{ and} \quad (13)$$

$$\eta_i = \frac{\sum_{j=1}^{n} (v_{ij})^{\acute{m}_2} \mathscr{N}(v_i, x_j)}{\sum_{j=1}^{n} (v_{ij})^{\acute{m}_2}}; \quad (14)$$

which represents the zone of influence or size of the cluster β_i.

The centroid is calculated based on the weighting average of the possibilistic lower approximation and probabilistic boundary. Computation of the centroid is modified to include the effects of both fuzzy memberships, probabilistic and possibilistic, and lower and upper bounds. The centroid calculation for the proposed clustering algorithm is obtained by solving equation (7) with respect to v_i:

$$v_i = \begin{cases} \omega\mathscr{C}_1 + (1-\omega)\mathscr{D}_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset,\ B(\beta_i) \neq \emptyset \\ \mathscr{C}_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset,\ B(\beta_i) = \emptyset \\ \mathscr{D}_1 & \text{if } \underline{A}(\beta_i) = \emptyset,\ B(\beta_i) \neq \emptyset \end{cases} \quad (15)$$

$$\text{where } \mathscr{C}_1 = \frac{\sum_{x_j \in \underline{A}(\beta_i)} (v_{ij})^{\acute{m}_2} x_j}{\sum_{x_j \in \underline{A}(\beta_i)} (v_{ij})^{\acute{m}_2}}; \quad (16)$$

$$\text{and } \mathscr{D}_1 = \frac{\sum_{x_j \in B(\beta_i)} (\mu_{ij})^{\acute{m}_1} x_j}{\sum_{x_j \in B(\beta_i)} (\mu_{ij})^{\acute{m}_1}}. \quad (17)$$

Hence, the cluster prototypes or centroids depend on the parameter ω, and fuzzifiers $\acute{m}_1$ and $\acute{m}_2$ rule their relative influence. The performance of the proposed clustering algorithm also depends on the values of two thresholds δ_1 and δ_2, which determine the cluster labels of all the miRNAs. In other words, the proposed clustering algorithm partitions the data set into two classes, namely, lower approximation and boundary, based on the values of δ_1 and δ_2. The thresholds δ_1 and δ_2 control the size of granules of the proposed clustering algorithm. In practice, the following definitions work well:

$$\delta_1 = \frac{1}{n} \sum_{j=1}^{n} (v_{ij} - v_{kj}) \quad (18)$$

where n is the total number of miRNAs, v_{ij} and v_{kj} are the highest and second highest memberships of object x_j. That is, the value of δ_1 represents the average difference of two highest possibilistic memberships of all the miRNAs in the data set. A good clustering procedure should make the value of δ_1 as high as possible. On the other hand, the miRNAs with $(v_{ij} - v_{kj}) \leq \delta_1$ are used to calculate the threshold δ_2:

$$\delta_2 = \frac{1}{\acute{n}} \sum_{j=1}^{\acute{n}} v_{ij} \quad (19)$$

where $\acute{n}$ is the number of miRNAs those do not belong to lower approximations of any cluster and v_{ij} is the highest membership of miRNA x_j. That is, the value of δ_2 represents the average of highest memberships of $\acute{n}$ miRNAs in the data set. The main steps of the proposed clustering algorithm proceed as follows:

1. Select c initial cluster prototypes using the NRNCBD based initialization method.
2. Choose values for fuzzifiers $\acute{m}_1$ and $\acute{m}_2$, and calculate thresholds δ_1 and δ_2. Set iteration counter $t = 1$.
3. Compute v_{ij} by equation (12) for c clusters and n objects.
4. If v_{ij} and v_{kj} are the highest and second highest possibilistic memberships of object x_j and $(v_{ij} - v_{kj}) > \delta_1$ then $x_j \in \underline{A}(\beta_i)$.
5. Otherwise, $x_j \in B(\beta_i)$ and $x_j \in B(\beta_k)$ if $v_{ij} > \delta_2$. Furthermore, x_j is not part of any lower bound.
6. Compute μ_{ij} for the objects lying in boundary regions for c clusters using equation (10).
7. Compute new centroid as per equation (15).
8. Repeat Steps 3 to 7, by incrementing t, until no more new assignments can be made.
9. Stop.

In this regard, it should be noted that different distance measures such as the Pearson distance and Euclidean distance can also be used in equation (5) for the selection of initial cluster prototypes as well as in equations (8), (9), (10), (12), and (14) for rough-fuzzy clustering of miRNA data sets. In general, the square of Euclidean distance is used in rough-fuzzy clustering[24], while the normalized Euclidean distance is used for the selection of initial clusters.

The Euclidean distance between two objects x_i and x_j is defined as

$$d_E(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}, \quad (20)$$

while the Pearson distance is defined as

$$d_P(x_i, x_j) = 1 - |\rho|, \quad (21)$$

where ρ represents the Pearson correlation coefficient, which is the ratio between the covariance of two vectors (x_i, x_j) of

expression values of two objects and product of their standard deviations and is given by

$$\rho(x_i,x_j) = \frac{Cov(x_i,x_j)}{\sigma_{x_i}\sigma_{x_j}}; \quad (22)$$

$$\text{that is, } \rho(x_i,x_j) = \frac{\sum_{k=1}^{m}(x_{ik}-\bar{x_i})(x_{jk}-\bar{x_j})}{\sqrt{\sum_{k=1}^{m}(x_{ik}-\bar{x_i})^2}\sqrt{\sum_{k=1}^{m}(x_{jk}-\bar{x_j})^2}}, \quad (23)$$

where $\bar{x_i}$ and $\bar{x_j}$ are the means of x_{ik} and x_{jk}, respectively. It considers each miRNA as a random variable with m observations and measures the similarity between the two miRNAs by calculating the linear relationship between the distributions of the two corresponding random variables.

4 Results and Discussions

In this section, the performance of the proposed method is compared with that of hard c-means (HCM)[12], fuzzy c-means (FCM)[31], rough-fuzzy c-means (RFCM)[26], cluster identification via connectivity kernels (CLICK)[15], and self organizing map (SOM)[32]. The performance of the NRNCBD over Pearson distance and Euclidean distance is also presented. The results are reported on four miRNA microarray data sets, namely, GSE16473, GSE17155, GSE29495, and GSE35074. For each data set, the number of clusters c is decided by using the CLICK[15] algorithm. Each miRNA data set is pre-processed by standardizing each feature or time point to zero mean and unit variance. The values of two fuzzifiers are set to 2.0, that is, $\acute{m}_1 = 2.0$ and $\acute{m}_2 = 2.0$. All the results are reported using four cluster validity indices, namely, Silhouette index[33], Dunn index[34], Davies-Bouldin index[35], and β index[36]. The biological analysis of the obtained miRNA clusters is also studied using the gene ontology. The source code of the proposed algorithm is available at www.isical.ac.in/∼bibl/results/cbd-rrfcm/cbd-rrfcm.html.

4.1 Quantitative Measures

Following quantitative indices are used to evaluate the performance of different clustering algorithms for grouping functionally similar miRNAs from microarray expression data sets.

4.1.1 Davies-Bouldin Index: The Davies-Bouldin (DB) index[35] is a function of the ratio of sum of within-cluster distance to between-cluster separation and is given by

$$\text{DB} = \frac{1}{c}\sum_{i=1}^{c}\max_{i\neq k}\left\{\frac{S(v_i)+S(v_k)}{d(v_i,v_k)}\right\} \quad (24)$$

for $1 \leq i,k \leq c$. The DB index minimizes the within-cluster distance $S(v_i)$ and maximizes the between-cluster separation $d(v_i,v_k)$. Therefore, for a given data set and c value, the higher the similarity values within the clusters and the between-cluster separation, the lower would be the DB index value. A good clustering procedure should make the value of DB index as low as possible.

4.1.2 Silhouette Index: Let an object $x_i \in \beta_r$, $i = 1,...,n_r$ and n_r is the cardinality of cluster β_r. For each object x_i let a_i be the average distance between object x_i and rest of the objects of β_r, that is,

$$a_i = d_{\text{avg}}(x_i, \beta_r - \{x_i\}) \quad (25)$$

where $d_{\text{avg}}(.,.)$ denotes the average distance measure between an object and a set of objects. For any other cluster $\beta_p \neq \beta_r$, let $d_{\text{avg}}(x_i,\beta_p)$ denote the average distance of object x_i to all objects of β_p. The scalar b_i is the smallest of these $d_{\text{avg}}(x_i,\beta_p), p = 1,\cdots,c, p \neq r$, that is,

$$b_i = \min_{p=1,\cdots,c,p\neq r}\left\{d_{\text{avg}}(x_i,\beta_p)\right\}. \quad (26)$$

The Silhouette width of object x_i is then defined as[33]

$$s(x_i) = \frac{b_i - a_i}{\max\{b_i,a_i\}} \quad (27)$$

where $-1 \leq s(x_i) \leq 1$. The value of $s(x_i)$ close to 1 implies that the distance of object x_i from the cluster β_r where it belongs is significantly less than the distance between x_i and its nearest cluster excluding β_r, which indicates that x_i is well clustered. On the other hand, the value of $s(x_i)$ close to -1 implies that the distance between x_i and β_r is significantly higher than the distance between x_i and its nearest cluster excluding β_r, which indicates that x_i is not well clustered. Finally, the values of $s(x_i)$ close to 0 indicate that x_i lies close to the border between the two clusters. Based on the definition of $s(x_i)$, the Silhouette of the cluster β_k ($k = 1,\cdots,c$) is defined as

$$S(\beta_k) = \frac{1}{n_k}\sum_{x_i\in\beta_k} s(x_i) \quad (28)$$

where n_k is the cardinality of the cluster β_k. The global Silhouette index is defined as

$$\mathscr{S}_c = \frac{1}{c}\sum_{k=1}^{c} S(\beta_k) \quad (29)$$

where $\mathscr{S}_c \in [-1,1]$. Also, the higher the value of $\mathscr{S}_c$, the better the corresponding clustering is.

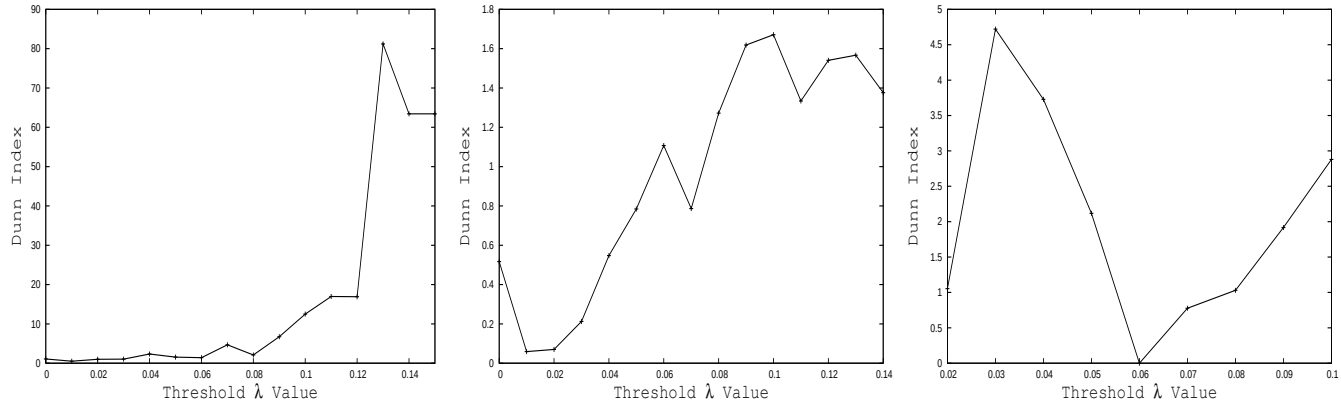


Fig. 1 Variation of Dunn index over different values of threshold λ for GSE16473, GSE17155, and GSE29495 data sets

4.1.3 β Index: The β index[36] is defined as the ratio of total variation and within-cluster variation, and is given by

$$\beta = \frac{\mathrm{N}}{\mathrm{M}}; \text{ where } \mathrm{N} = \sum_{i=1}^{c}\sum_{j=1}^{n_i} ||x_{ij} - \bar{v}||^2;$$

$$\mathrm{M} = \sum_{i=1}^{c}\sum_{j=1}^{n_i} ||x_{ij} - v_i||^2; \text{ and } \sum_{i=1}^{c} n_i = n; \tag{30}$$

n_i is the number of objects in the ith cluster ($i = 1, 2, \cdots, c$), n is the total number of objects, x_{ij} is the jth object in cluster β_i, v_i is the mean or centroid of ith cluster, and $\bar{v}$ is the mean of n objects. For a given data set and c value, the higher the homogeneity within the clusters, the higher would be the β value. The value of β also increases with c.

4.2 Optimum Values of λ and ω Parameters

The threshold λ plays an important role to generate the initial cluster centers. It controls the degree of dissimilarity among the miRNAs present in microarray data. In effect, it has a direct influence on the performance of the initialization method used. Also, the performance of the proposed clustering algorithm depends on the weight parameter ω.

Let $\Phi = \{\lambda, \omega\}$ be the set of parameters and $\Phi^\star = \{\lambda^\star, \omega^\star\}$ is the set of optimal parameters. To find out the optimum set $\Phi^\star$, containing optimum values of $\lambda^\star$ and $\omega^\star$, the Dunn's cluster validity index[34] is used here. Dunn's (D) index[34] is designed to identify sets of clusters that are compact and well separated. Dunn's (D) index maximizes

$$\mathrm{D} = \min_i \left\{ \min_{k \neq i} \left\{ \frac{d(v_i, v_k)}{\max_l S(v_l)} \right\} \right\} \tag{31}$$

for $1 \leq i, k, l \leq c$, where v_i is the centroid of cluster β_i, $S(v_l)$ is the within-cluster distance, $d(v_i, v_k)$ is the between-cluster separation, and c is the number of clusters.

For four miRNA microarray data sets, the value of λ is varied from 0.0 to 0.15, while the value of ω is varied from 0.51 to 0.99. The optimum values of $\lambda^\star$ and $\omega^\star$ for each microarray data set are obtained using the following relation:

$$\Phi^\star = \arg\max_{\Phi} \{\mathrm{D}\}. \tag{32}$$

The proposed clustering algorithm with the NRNCBD distance measure achieves optimum values of λ using equation (32) at 0.13, 0.10, 0.03, and 0.15 for GSE16473, GSE17155, GSE29495, and GSE35074 data sets, respectively, while the optimum value of $\omega = 0.99$ for all the data sets. Fig. 1 represents the variation of Dunn index with respect to different values of λ considering $\omega^\star = 0.99$ on GSE16473, GSE17155, and GSE29495 data sets. From the results reported in Fig. 1, it is seen that as the threshold λ increases, the Dunn index value increases and attains its maximum value at a particular value of $\lambda^\star$. After that the Dunn index value decreases with the increase in the value of λ.

On the other hand, the optimum values of λ and ω for the proposed clustering algorithm with Pearson distance for four data sets, namely, GSE16473, GSE17155, GSE29495, and GSE35074, are $\{0.08, 0.65\}$, $\{0.00, 0.51\}$, $\{0.09, 0.99\}$, and $\{0.03, 0.95\}$, respectively. However, the proposed clustering algorithm with Euclidean distance achieves optimum values of λ using equation (32) at 0.14, 0.15, 0.13, and 0.14 for GSE16473, GSE17155, GSE29495, and GSE35074 data sets, respectively, while the optimum value of $\omega = 0.99$ for all the data sets.

4.3 Performance of Different C-Means Algorithms and Distance Measures

Tables 1 and 2 provide the comparative performance analysis of different c-means algorithms with respect to three distance measures, namely, Pearson distance (PD), Euclidean distance

Table 1 Comparative Performance of Different C-Means Algorithms and Distance Measures on GSE16473 and GSE17155 Data Sets

Validity Index	Distance Measure	GSE16473				GSE17155			
		HCM	FCM	RFCM	Proposed	HCM	FCM	RFCM	Proposed
Silhouette	PD	-0.011	-0.086	-0.091	0.198	-0.203	-0.184	-0.168	-0.150
	ED	0.314	0.238	0.252	0.687	**0.192**	**0.122**	**0.180**	0.181
	NRNCBD	**0.920**	**0.258**	**0.920**	**0.943**	0.190	0.054	0.117	**0.299**
DB	PD	10.998	14.441	3.208	1.727	26.769	109.406	29.003	22.244
	ED	1.897	**4.406**	3.299	0.206	1.628	**25.020**	1.367	0.793
	NRNCBD	**0.015**	15.856	**0.0156**	**0.007**	**1.562**	176.277	**1.159**	**0.658**
Dunn	PD	0.070	0.032	0.172	0.258	0.030	0.007	0.017	0.021
	ED	0.168	**0.064**	0.207	4.815	0.685	**0.018**	0.869	1.269
	NRNCBD	**34.696**	0.000	**33.558**	**81.200**	**0.702**	0.003	**1.017**	**1.671**
β	PD	1.087	0.995	0.805	**10.843**	1.300	1.297	1.367	1.729
	ED	1.483	**1.528**	1.283	5.918	**8.058**	**6.976**	5.710	12.450
	NRNCBD	**2.204**	1.114	**2.174**	6.281	7.606	4.823	**6.699**	**17.649**

Table 2 Comparative Performance of Different C-Means Algorithms and Distance Measures on GSE29495 and GSE35074 Data Sets

Validity Index	Distance Measure	GSE29495				GSE35074			
		HCM	FCM	RFCM	Proposed	HCM	FCM	RFCM	Proposed
Silhouette	PD	-0.377	-0.365	-0.369	0.214	-0.035	**-0.063**	-0.060	-0.034
	ED	**0.675**	**0.519**	**0.664**	0.796	-0.111	-243	0.045	-0.125
	NRNCBD	0.549	0.080	0.609	**0.907**	**0.051**	-0.212	**0.062**	**0.119**
DB	PD	34.508	18.710	35.781	3.465	10.689	**122.896**	5.941	4.561
	ED	**0.158**	**1.182**	**0.351**	0.122	**3.638**	150.648	2.110	1.379
	NRNCBD	0.385	568.563	0.501	**0.092**	4.112	186.976	**1.360**	**0.772**
Dunn	PD	0.004	0.012	0.011	0.010	0.104	**0.003**	0.148	0.161
	ED	**4.532**	**0.085**	**1.816**	**5.200**	**0.275**	0.000	0.512	0.586
	NRNCBD	0.558	0.000	0.281	4.721	0.268	0.000	**0.905**	**1.402**
β	PD	1.273	0.868	0.101	5.908	1.201	**1.072**	0.846	1.200
	ED	**28.072**	**24.356**	11.669	35.438	**1.641**	0.754	1.481	2.585
	NRNCBD	23.651	9.891	**30.072**	**963.320**	1.596	1.000	**1.158**	**2.769**

(ED), and the NRNCBD, on four miRNA microarray data sets. The results of different *c*-means algorithms are reported for their optimal values of $\lambda^\star$ and $\omega^\star$. In most of the cases, the NRNCBD is found to improve the performance in terms of Silhouette, DB, Dunn, and β indices, irrespective of the *c*-means algorithms. Out of total 64 comparisons, the NRNCBD is found to provide significantly better results in 34 cases compared to both Pearson distance and Euclidean distance. On the other hand, the Pearson distance and Euclidean distance achieve better results in 5 and 25 cases, respectively.

From the results reported in Tables 1 and 2, it can also be seen that, out of total 16 cases, the HCM algorithm with Euclidean distance and the NRNCBD performs better in 9 and 7 cases, respectively. Similarly, the FCM with Euclidean distance attains better results in 11 cases out of total 16 cases, while it achieves better results in 4 and 1 cases, respectively, with Pearson distance and the NRNCBD. On the other hand, the RFCM algorithm with the NRNCBD and Euclidean distance performs better in 12 and 4 cases, respectively. Also, the proposed rough-fuzzy clustering algorithm attains better results in 14, 1, and 1 cases using the NRNCBD, Pearson distance, and Euclidean distance, respectively. Hence, the performance of different *c*-means algorithms deteriorates with Pearson distance. All the results reported above establish the fact that the Euclidean distance is an appropriate choice for both HCM and FCM, while both existing RFCM and proposed rough-fuzzy clustering algorithm perform significantly better using the NRNCBD compared to other two distance measures. Also, the NRNCBD based proposed clustering algorithm achieves better performance in 14 cases out of total 16 comparisons, irrespective of the *c*-means algorithms, cluster validity indices, distance measures, and miRNA data sets used.

Moreover, it is also seen that the proposed rough-fuzzy clustering algorithm achieves better results than that obtained using existing RFCM algorithm, irrespective of the data sets, distance measures, and quantitative indices used. Out of total 48 comparisons, the proposed algorithm attains better results in

46 cases. In existing RFCM, each cluster is represented by a cluster prototype, a crisp lower approximation and a probabilistic boundary. The crisp lower approximation of a miRNA cluster in existing RFCM is usually assumed to be spherical in shape, which restricts to find arbitrary shapes of miRNA clusters and forces to extract circular shaped miRNA clusters. On the other hand, in the proposed rough-fuzzy clustering algorithm, each cluster is represented by a cluster prototype, a possibilistic lower approximation, and a probabilistic boundary. The possibilistic lower approximation of the proposed algorithm helps to extract miRNA groups of any shape. In effect, the chance of inclusion of noisy miRNAs becomes more in the existing RFCM as compare to the proposed rough-fuzzy clustering algorithm. Hence, the possibilistic lower approximation of the proposed algorithm helps in discovering clusters of miRNAs that are highly similar to each other.

Table 3 Performance of Different Clustering Algorithms

Validity Index	Methods/ Algorithms	Data Sets / GSE			
		16473	17155	29495	35074
Silhouette	CLICK	0.005	-0.101	-0.634	0.038
	SOM	0.059	-0.112	-0.540	0.009
	Proposed	0.971	0.471	0.928	0.415
DB	CLICK	2.277	13.016	450.689	8.929
	SOM	10.128	39.558	455.345	19.875
	Proposed	0.007	0.658	0.092	0.772
Dunn	CLICK	0.101	0.003	0.000	0.007
	SOM	0.011	0.001	0.000	0.003
	Proposed	81.200	1.671	4.721	1.402
β	CLICK	0.175	0.090	0.171	0.185
	SOM	0.360	0.205	0.385	0.306
	Proposed	6.281	17.649	963.320	2.769

4.4 Performance of Different Clustering Algorithms

Table 3 presents the performance of different clustering algorithms. The results and subsequent discussions are presented with respect to the Silhouette, DB, Dunn, and β indices. From Table 3, it can be observed that the proposed method outperforms other clustering algorithms, irrespective of the quantitative indices and miRNA data sets used. The best performance of the proposed clustering algorithm is achieved due to the following reasons:

1. the city block distance based dissimilarity measure used for initial partition of data set enables the algorithm to converge to an optimum or near optimum solutions;
2. the city block distance, used to calculate possibilistic and probabilistic membership functions, provides effective values for degree of belongingness of the miRNAs;
3. probabilistic membership function of the proposed clustering algorithm handles efficiently overlapping partitions, while the possibilistic membership function of lower approximation of a cluster helps to discover arbitrary shaped cluster; and
4. the concept of possibilistic lower approximation and fuzzy boundary of the proposed algorithm deals with uncertainty, vagueness, and incompleteness in class definition.

4.5 Qualitative Performance Analysis

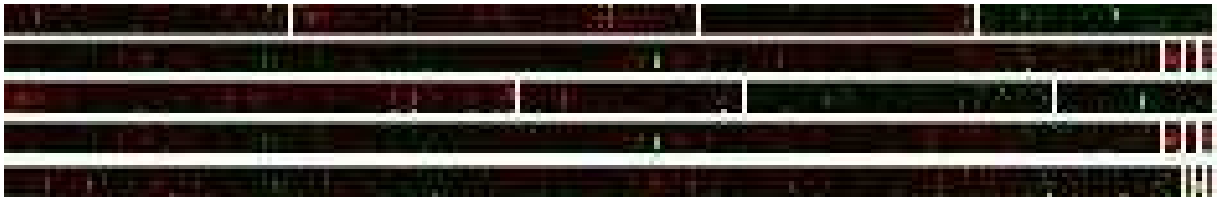
This section presents the visual representation of the clustering solutions obtained by different clustering algorithms. The Eisen plots[37] are generated for each clustering solution of each data set. In the present representation, the miRNAs are ordered before plotting so that the miRNAs that belong to the same cluster are placed one after another. The cluster boundaries are identified by white colored blank rows. The miRNA clusters produced by the SOM, HCM, FCM, RFCM, and proposed algorithms on four data sets are visualized by TreeView software, which is available at http://rana.lbl.gov/EisenSoftware and the plots for four data sets are reported in Fig. 2 as examples.

From the Eisen plots presented in Fig. 2, it is evident that the expression profiles of the miRNAs in a cluster are similar to each other and they produce similar color pattern, whereas the miRNAs from different clusters differ in color patterns. Also, the results obtained by both RFCM and proposed algorithms are more promising than that by both HCM and FCM algorithms. From the plots presented in Fig. 2, it is clearly evident that the proposed method generates the Eisen plots having similar color pattern within the cluster as compare to other clustering algorithms.

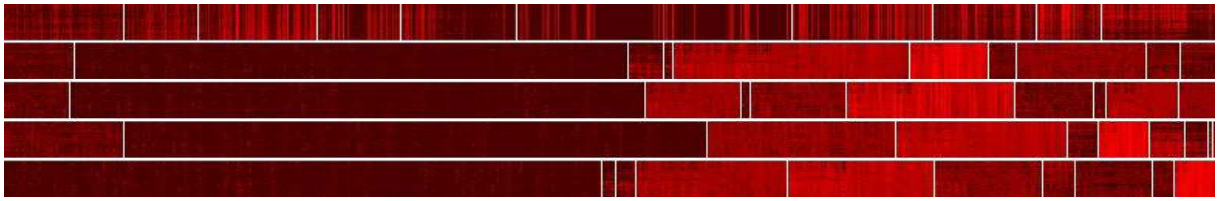
4.6 Functional Consistency of Clustering Result

DIANA microT v3.0[38], a miRNA target prediction algorithm, is used to predict miRNA target genes for all miRNA clusters generated by different clustering algorithms. For each miRNA cluster, genes that are targeted by at least t percentage (%) of miRNAs in a cluster are used for further analysis. Here, the value of t is varied from 10 to 75.

In order to evaluate the functional consistency of the genes targeted by miRNAs of a cluster, the biological annotations of those genes of different clusters are considered in terms of the gene ontology (GO). The annotation ratios of each targeted gene cluster in three GO ontologies are calculated using the GO Term Finder[39]. The GO term is searched in which most of the genes of a particular cluster are enriched[40]. The annotation ratio, also termed as cluster frequency, of a gene cluster is defined as the number of genes in both the assigned GO term and the cluster divided by the number of genes in that cluster. A higher value of annotation ratio indicates that the majority of genes in the cluster are functionally more closer to

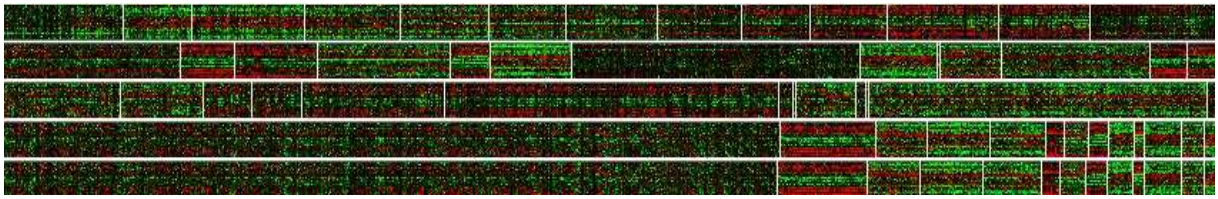


(a) GSE16473: $n = 231, m = 7, c = 4$



(b) GSE17155: $n = 774, m = 38, c = 10$

(c) GSE29495: $n = 574, m = 17, c = 10$



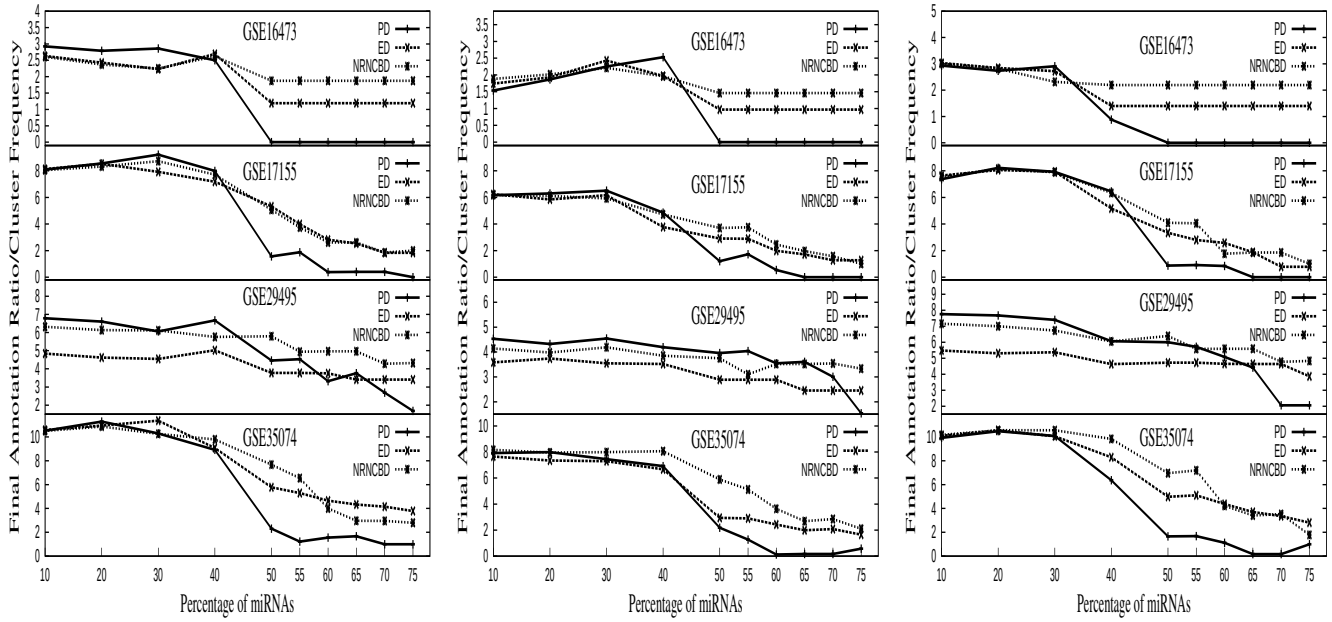
(d) GSE35074: $n = 837, m = 21, c = 13$

Fig. 2 Eisen plots of different clusters generated by SOM, HCM, FCM, RFCM, and proposed algorithms

each other and miRNAs targeting these genes are involved in common cellular processes, while a lower value signifies that the cluster contains much more noises or irrelevant genes and the miRNAs targeting these genes are just randomly clustered. After computing the annotation ratios of all gene clusters for a particular ontology, the sum of all annotation ratios is treated as the final annotation ratio. A higher value of final annotation ratio indicates that the corresponding clustering result is better than other, that is, the genes are better clustered by function, indicating a more functionally consistent clustering result[41].

Fig. 3 presents the comparative performance analysis of the NRNCBD, Pearson distance, and Euclidean distance with respect to the proposed clustering algorithm. The final annotation ratios generated by three distance measures for molecular functions (MF), biological processes (BP), and cellular components (CC) ontologies on four miRNA microarray data sets are shown in this figure. Here, the results are presented for those genes that are targeted by at least 10% to 75% miRNAs in a cluster. From all the results reported in Fig. 3, it is seen that in most of the cases the NRNCBD performs better than both Pearson distance and Euclidean distance. For the proposed clustering algorithm, the NRNCBD performs better than both Pearson distance and Euclidean distance in 65 cases, out of total 120 comparisons. However, the Pearson distance and Euclidean distance perform better in 31 and 24 cases, respectively. The dimension additivity property of the NRNCBD, that is, the total distance is a sum of the distances per dimension, leads to better functionally consistent clustering solutions as compared to Pearson distance and Euclidean distance.

The genes that are targeted by at least 50% miRNAs of a cluster are further analyzed and the results are reported in



(a) Molecular Function (b) Biological Process (c) Cellular Component

Fig. 3 Biological annotation ratios obtained using proposed algorithm with NRNCBD, Pearson and Euclidean distances on four data sets

Fig. 4. The upper portion of Fig. 4 presents the comparative results of the RFCM and proposed clustering algorithm, in terms of final annotation ratio or cluster frequency, for the MF, BP, and CC ontologies on four miRNA expression data sets. All the results reported here confirm that the proposed method provides higher or comparable final annotation ratios than that obtained using the RFCM algorithm in most of the cases. Out of 12 cases, the proposed method provides higher final annotation ratio in 11 cases. On the other hand, the RFCM with Pearson distance attains better result for the BP ontology.

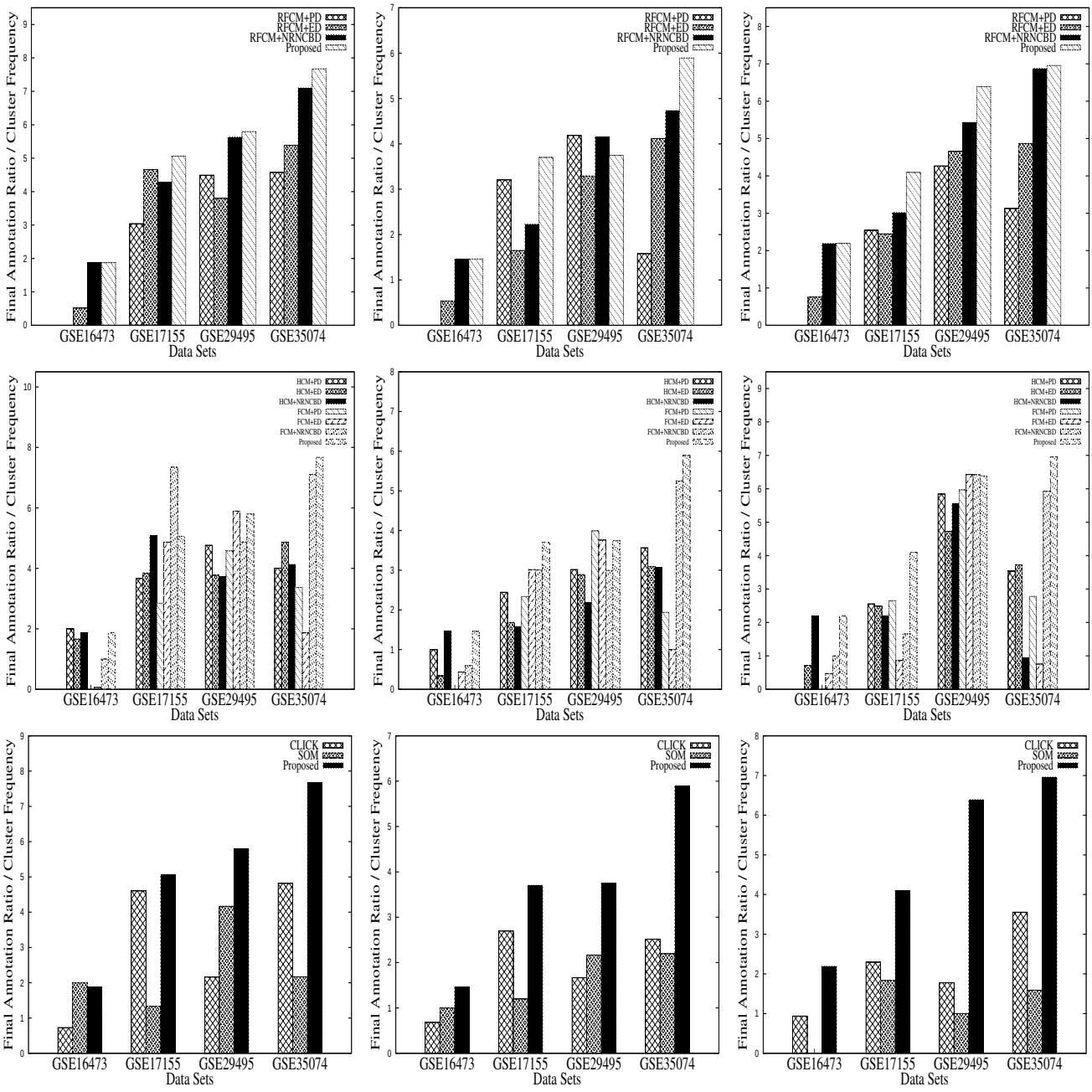
The middle portion of Fig. 4 reports the comparative final annotation ratio of the HCM, FCM, and the proposed algorithm on four data sets. From the results reported in this portion, it is seen that out of total 12 comparisons, the proposed algorithm attains highest final annotation ratio than that obtained using other *c*-means algorithms in 1, 3, and 3 cases for the MF, BP, and CC ontologies, respectively. On the other hand, the HCM provides it in only one case using Pearson distance. Also, the FCM generates higher final annotation ratio in 1, 2 and 1 cases using Pearson distance, Euclidean distance, and the NRNCBD, respectively.

Finally, the lower portion of Fig. 4 compares the final annotation ratios obtained using the CLICK, SOM, and proposed clustering algorithm. From the results reported in this portion, it can be seen that the final annotation ratio obtained using the proposed algorithm is higher than that obtained using both CLICK and SOM in 11 cases out of 12 cases. However, the SOM generates higher final annotation ratio in only 1 case for the MF ontology. Hence, the majority of genes in a cluster produced by the proposed algorithm are functionally more closer to each other than those by other algorithms, while the clusters obtained using existing algorithms include much more noises or irrelevant genes.

4.7 Biologically Significant Gene Clusters

The genes that are targeted by at least 50% miRNAs are used to calculate the number of significant gene clusters. Fig. 5 presents the results for the MF, BP, and CC ontologies on four data sets. The GO Term Finder is used to determine the statistically significant gene clusters produced by different algorithms for all the GO terms from the MF, BP, and CC ontologies. If any cluster of genes generates a p-value smaller than 0.05, then that cluster is considered as a significant cluster. The upper portion of Fig. 5 presents the comparative results of the RFCM and proposed algorithm for the MF, BP, and CC ontologies, respectively. From the results, it is seen that the proposed algorithm generates more or comparable number of significant gene clusters in all the 12 cases.

The middle portion of Fig. 5 reports the number of significant gene clusters generated by the HCM, FCM, and proposed algorithm for the MF, BP, and CC ontologies for all microarray data sets, respectively. All the results reported in this portion establish the fact that the proposed algorithm generates more

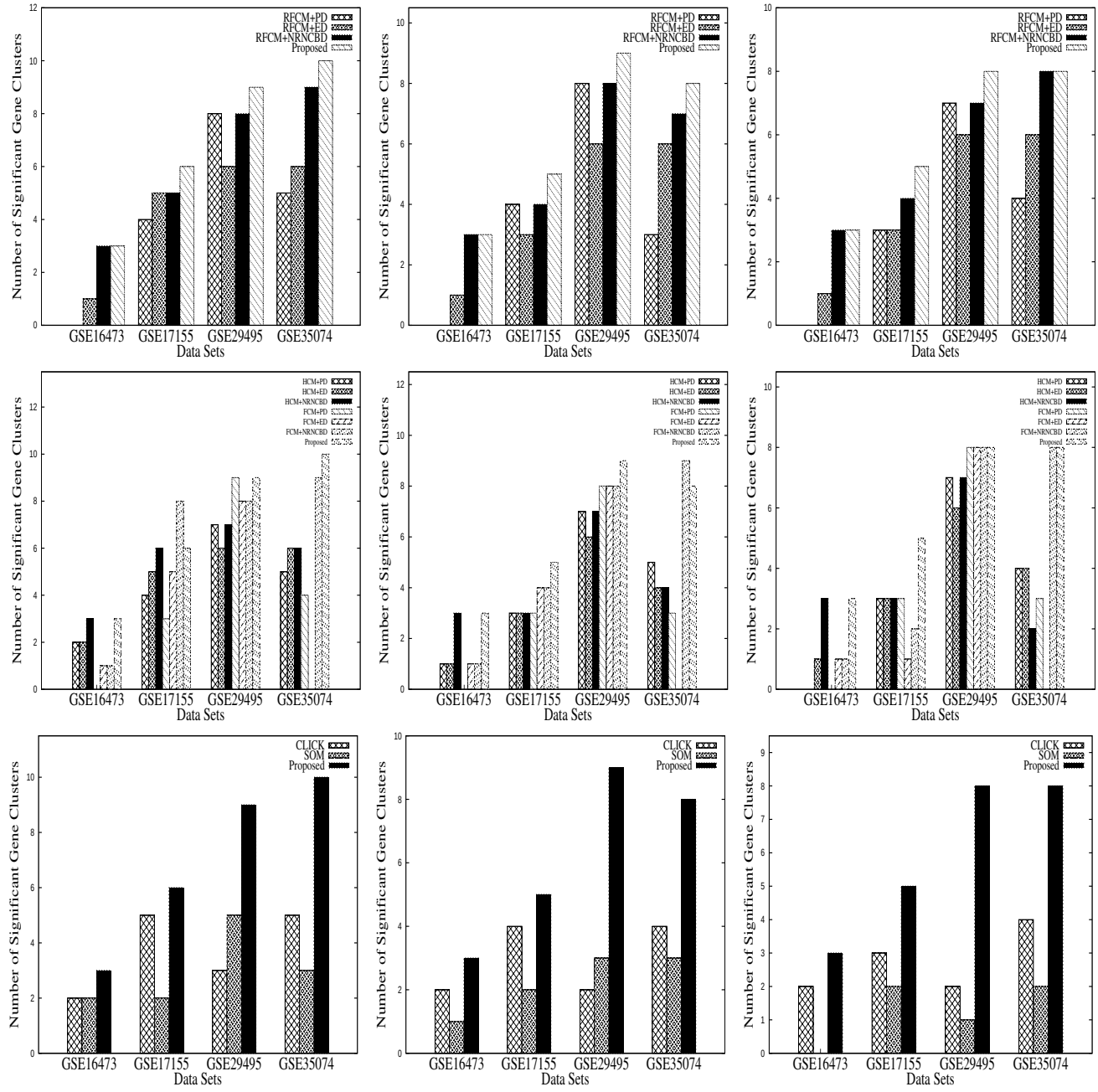


(a) Molecular Function (b) Biological Process (c) Cellular Component

Fig. 4 Biological annotation ratios obtained using different clustering algorithms

or comparable number of significant gene clusters than that of other *c*-means algorithms in most of the cases. For the MF, BP, and CC ontologies, the proposed method generates more or comparable number of significant gene clusters in 3, 3, and 4 cases, respectively. That is, out of total 12 cases, it provides better results in 10 cases. However, the FCM algorithm with the NRNCBD generates better result in one case each for both MF and BP ontologies, respectively.

Finally, the performance of CLICK, SOM, and proposed algorithm is compared in lower portion of Fig. 5 with respect to the number of significant gene clusters generated for MF, BP, and CC ontologies, respectively. From the results reported in this portion, it is seen that the proposed algorithm generates more or comparable number of significant gene clusters com-



(a) Molecular Function (b) Biological Process (c) Cellular Component

Fig. 5 Biologically significant gene clusters obtained using different clustering algorithms

pared to both CLICK and SOM algorithms in all the cases. From Fig. 5, it can also be seen that the proposed clustering algorithm produces better results irrespective of the distance measures, ontologies, and data sets used. Hence, it can be concluded that the proposed clustering algorithm generates highly compact and functionally enriched clusters.

4.8 Biological Interpretation of Gene/miRNA Clusters

This section presents the biological interpretation of some gene clusters those are generated only by the proposed algorithm, but not generated by any other clustering algorithms. Table 4 presents the unique GO terms obtained using the proposed algorithm for GSE17155, GSE29495, and GSE35074 data sets, along with the corresponding cluster index and fre-

Table 4 Unique GO Terms Obtained Using Proposed Algorithm for Different Data Sets

Data Set	Ontology	Cluster	GO Term / Gene Cluster	Frequency	P-Value	FDR (%)
GSE17155	Molecular Function	7	nucleic acid binding transcription factor activity	1.000	1.27E-004	0.00
		9	nucleic acid binding transcription factor activity	0.467	2.28E-007	0.00
	Biological Process	3	regulation of cellular process	0.659	4.08E-011	0.00
	Cellular Component	9	nucleus	0.864	7.15E-007	0.00
GSE29495	Molecular Function	9	beta-catenin binding	0.111	1.41E-002	2.00
	Biological Process	1	RNA metabolic process	0.714	7.73E-003	0.00
		6	multicellular organismal development	0.375	6.26E-007	0.00
		7	regulation of macromolecule metabolic process	0.444	3.80E-017	0.00
		9	regulation of DNA binding	0.167	2.13E-003	0.00
GSE35074	Molecular Function	5	nucleic acid binding	1.000	8.34E-003	4.00
	Biological Process	4	negative regulation of biological process	1.000	2.40E-002	0.00

quency, p-value, and false discovery rate (FDR). The FDR is a multiple-hypothesis testing error measure indicating the expected proportion of false positives among the set of significant results. It is particularly useful in the analysis of high-throughput data such as microarray miRNA expression.

The biological interpretation for GSE17155 data set is reported in this section as an example. This data set contains specific miRNA expression signature, which characterizes male breast cancers. Hence, the unique GO terms generated by the proposed clustering algorithm should reflect the processes related to breast cancer. This section discusses the importance of the genes corresponding to the unique GO term in breast cancer. The relation between the set of miRNAs corresponding to the unique GO term and breast cancer is also established in subsequent discussion.

It has been observed in [42] that the DNA binding activity of few transcriptional factors acts as one of the major causes of breast cancer [43]. The partial or complete loss of a transcription factor is a common event in a breast cancer tumor or cell lines. Transcription factors are gene regulatory proteins endowed with sequence-specific DNA recognition and the ability to positively or negatively influence the rate and efficiency of transcript initiation at a gene containing the factors cognate recognition sequence, or DNA response element. Since transcription factors lie at the heart of almost every fundamental developmental and homeostatic organismal process including DNA replication and repair, cell growth and division, control of apoptosis and cellular differentiation, it is not surprising that inherited or acquired defects in transcription factor structure and function contribute to human carcinogenesis. From several studies, it can be seen that this growing body of transcription factors and the development-specific and issue-restricted gene programs under their control represent a rich and diverse source of mechanisms which, if disrupted, can lead to various types of malignancy including breast cancer.

The unique GO term "nucleic acid binding transcription factor activity" corresponding to clusters 7 and 9 identified by the proposed method reflect this activity of cancerous cell. There are total 10 genes, namely, **NFIA**, **LCOR**, **CSRNP3**, **KLF12**, **ZFHX4**, **KLF3**, **SOX4**, **RUNX1T1**, **TSHZ2**, and **ZEB2**, present in these two clusters. These genes act as transcription factors. The **NFIA** gene was found to involve in an **NFIA/EHF** chimeric fusion in one breast cancer cell line out of 24 breast tumors analyzed: 9 cell lines and 15 primary tumors [44]. However, its role as either a passenger event or a direct, albeit infrequent, contributor to breast cancer development, remains uncertain. The **LCOR** represents a class of corepressor that attenuates agonist-activated nuclear receptor signaling by multiple mechanisms. The **LCOR** transcript raises in breast carcinoma cells [45]. It has been discovered that multiple polymorphic variations in the **KLF12**, which encodes a zinc finger repressor factor and a region surrounding the gene, are associated with the occurrence of cancer, in particular breast cancer [46]. Important function of **SOX4** in the progression of breast cancer by orchestrating the EMT (epithelial-to-mesenchymal transition) has been shown in [47], and this gene product has been implicated as a marker of poor prognosis in this disease. The mechanism of regulation of transcription in breast cancer cell line by **TSHZ2** has been shown in [48]. The EMT plays an important role during normal embryogenesis, and it has been implicated in cancer invasion and metastasis. An aberrant expression of homeobox gene may lead to the activation of a developmentally regulated EMT pathway in human breast cancer. The **ZEB2** regulates developmental EMT and also play roles in tumor progression [49].

On the other hand, there are total 15 miRNAs, namely, **hsa-miR-211**, **hsa-miR-30a**, **hsa-miR-21**, **hsa-miR-299-5p**, **hsa-miR-626**, **hsa-miR-132**, **hsa-miR-32**, **hsa-miR-369-3p**, **hsa-miR-605**, **hsa-miR-616**, **hsa-miR-373**, **hsa-miR-626**, **hsa-miR-622**, **hsa-miR-33b**, and **hsa-miR-138**, which are associated with clusters 7 and 9. The importance of **hsa-miR-211** in the progression of breast cancer has been shown in [50,51]. This miRNA downregulates **RunX2** and **IL11** genes that leads to progression of breast cancer. Zeng et al. [52] have shown that

downregulation of **hsa-miR-30a** in human plasma lead to the overexpression of the genes **CEA** and **CA153**. Hence, the **hsa-miR-30a** acts as a novel marker for breast cancer. In[53], it has been reported that the **hsa-miR-21** regulates breast cancer invasion partly by targeting tissue inhibitor of metalloproteinase 3 **TIMP3** gene expression. Also, the decreased level of **hsa-miR-299-5p** plays a critical role by increasing the level of **OPN** protein that enhance proliferation, tumorigenicity and the ability to display vasculogenic mimicry of the spheroid-forming cells[54]. It has been observed that the **hsa-miR-626** is highly expressed in luminal cell lines lacking **ERBB2** overexpression. Upregulation of **hsa-miR-132** that leads to suppression of **p120RasGAP** in human breast cancer cells has been reported in[55]. Downregulated expression of **hsa-miR-32** in breast carcinoma cells has been observed in[56]. An enrichment test analysis reported in[28] has shown the involvement of **hsa-miR-369-3p**, **hsa-miR-605**, and **hsa-miR-616** in the breast cancer. Huang et al.[57] have demonstrated that **hsa-miR-373** promotes tumor invasion and metastasis in breast cancer. An association between **hsa-miR-626** and breast cancer has been demonstrated in[58]. The miRNA **hsa-miR-622** is linked to enhanced tumorigenesis in breast cancer[58]. Overexpression of **hsa-miR-33b** in breast carcinoma cell lines has been observed in[59]. The miRNA **hsa-miR-138** has been found differentially expressed in human male breast cancer[60].

Hence, the biological interpretation of some unique clusters identified by the proposed rough-fuzzy clustering algorithm reported above establishes the fact that the algorithm generates significant miRNA clusters those are biologically relevant with respect to the given microarray data sets.

5 Conclusion

The paper presents a new miRNA clustering algorithm, integrating judiciously the merits of rough sets, fuzzy sets, *c*-means algorithm, and normalized range-normalized city block distance. The proposed algorithm is used to find groups of co-expressed miRNAs from microarray data. While the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in cluster definition, the membership functions of fuzzy sets enable efficient handling of overlapping clusters in noisy environment. The city block distance is useful to find initial partition of a miRNA data set and helps to handle minute differences between two miRNA expression profiles.

The effectiveness of the proposed clustering algorithm, along with a comparison with other clustering algorithms, has been demonstrated on four miRNA microarray data sets using some cluster validity indices and gene ontology. The extensive experimental results show that the proposed algorithm produces better clustering results than do the conventional algorithms in terms of Silhouette index, DB index, Dunn index, β index, final annotation ratios, and significant gene clusters. The proposed method attains better performance in more than 87.50% cases as compare to other *c*-means algorithms. Also, the dimension additivity property of city block distance leads to better clustering solutions compared to both Pearson and Euclidean distances; thereby successful in effectively circumventing the initialization and local minima problems of iterative refinement clustering algorithms like *c*-means.

Moreover, the city block distance based proposed rough-fuzzy clustering algorithm achieves better results than that obtained using two popular clustering algorithms. The proposed algorithm also generates more number of biologically significant miRNA clusters than the existing *c*-means and other clustering algorithms. The biological interpretation of unique clusters identified by the proposed algorithm also establishes the fact that the algorithm generates significant miRNA clusters those are biologically relevant with respect to the given microarray data sets.

References

1 Y. Altuvia, P. Landgraf, G. Lithwick, N. Elefant, S. Pfeffer, A. Aravin, M. J. Brownstein, T. Tuschl and H. Margalit, *Nucleic Acids Research*, 2005, **33**, 2697–2706.
2 S. Baskerville and D. P. Bartel, *RNA*, 2005, **11**, 241–247.
3 X. Cai, C. H. Hagedorn and B. R. Cullen, *RNA*, 2004, **10**, 1957–1966.
4 Y. Lee, M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek and V. N. Kim, *The EMBO Journal*, 2004, **23**, 4051–4060.
5 E. Enerly, I. Steinfeld, K. Kleivi, S. K. Leivonen, M. R. Aure, H. G. Russnes, J. A. Ronneberg, H. Johnsen, R. Navon, E. Rodland, R. Makela, B. Naume, M. Perala, O. Kallioniemi, V. N. Kristensen, Z. Yakhini and A. L. B. Dale, *PLoS ONE*, 2011, **6**, e16915.
6 P. Maji and S. K. Pal, *Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging*, Wiley-IEEE Computer Society Press, New Jersey, 2012.
7 E. Domany, *Journal of Statistical Physics*, 2003, **110**, 1117–1139.
8 S. Tavazoie, D. Hughes, M. J. Campbell, R. J. Cho and G. M. Church, *Nature Genetics*, 1999, **22**, 281–285.
9 J. Lu, G. Getz, E. A. Miska, E. A. Saavedra, J. Lamb, D. Peck, A. S. Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub, *Nature Letters*, 2005, **435**, 834–838.
10 C. Wang, S. Yang, G. Sun, X. Tang, S. Lu, O. Neyrolles and Q. Gao, *PLoS ONE*, 2011, **6**, 1–11.
11 R. Bargaje, M. Hariharan, V. Scaria and B. Pillai, *RNA*, 2010, **16**, 16–25.
12 L. J. Heyer, S. Kruglyak and S. Yooseph, *Genome Research*, 1999, **9**, 1106–1115.
13 A. Ben-Dor, R. Shamir and Z. Yakhini, *Journal of Computational Biology*, 1999, **6**, 281–297.
14 E. Hartuv and R. Shamir, *Information Processing Letters*, 2000, **76**, 175–181.
15 R. Shamir and R. Sharan, Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, 2000, pp. 307–31.
16 E. P. Xing and R. M. Karp, *Bioinformatics*, 2001, **17**, 306–315.
17 C. Fraley and A. E. Raftery, *The Computer Journal*, 1998, **41**, 578–588.
18 D. Ghosh and A. M. Chinnaiyan, *Bioinformatics*, 2002, **18**, 275–286.
19 G. J. McLachlan, R. W. Bean and D. Peel, *Bioinformatics*, 2002, **18**, 413–422.

20 K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery and W. L. Ruzz, *Bioinformatics*, 2001, **17**, 977–987.

21 D. Jiang, J. Pei and A. Zhang, Proceedings of the 3rd IEEE International Symposium on Bioinformatics and Bioengineering, 2003, pp. 393–400.

22 L. A. Zadeh, *Information and Control*, 1965, **8**, 338–353.

23 Z. Pawlak, *Rough Sets: Theoretical Aspects of Resoning About Data*, Kluwer, Dordrecht, The Netherlands, 1991.

24 P. Maji and S. K. Pal, *IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics*, 2007, **37**, 1529–1540.

25 J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum, 1981.

26 P. Maji and S. K. Pal, *Fundamenta Informaticae*, 2007, **80**, 475–496.

27 Y. Chen, J. Gelfond, L. M. McManus and P. K. Shireman, *Physiological Genomics*, 2011, **43**, 621–630.

28 M. Fassan, R. Baffa, J. Palazzo, J. Lloyd, M. Crosariol, C. G. Liu, S. Volinia, H. Alder, M. Rugge, C. Croce and A. Rosenberg, *Breast Cancer Research*, 2009, **11**, R58.

29 M. Moes, A. L. Bechec, I. Crespo, C. Laurini, A. Halavatyi, G. Vetter, A. d. Sol and E. Friederich, *PLoS ONE*, 2012, **7**, e35440.

30 R. Krishnapuram and J. M. Keller, *IEEE Transactions on Fuzzy Systems*, 1993, **1**, 98–110.

31 D. Dembele and P. Kastner, *Bioinformatics*, 2003, **19**, 973–980.

32 P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander and T. R. Golub, *Proceedings of the National Academy of Sciences, USA*, 1999, **96**, 2907–2912.

33 J. P. Rousseeuw, *Journal of Computational and Applied Mathematics*, 1987, **20**, 53–65.

34 J. C. Dunn, *Journal of Cybernetics*, 1974, **3**, 32–57.

35 D. L. Davies and D. W. Bouldin, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979, **1**, 224–227.

36 S. K. Pal, A. Ghosh and B. U. Shankar, *International Journal of Remote Sensing*, 2000, **21**, 2269–2300.

37 M. B. Eisen, P. T. Spellman, O. Patrick and D. Botstein, *Proceedings of the National Academy of Sciences, USA*, 1998, **95**, 14863–14868.

38 M. Maragkakis, P. Alexiou, G. L. Papadopoulos, M. Reczko, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, V. A. Simossis, P. Sethupathy, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas and A. G. Hatzigeorgiou, *BMC Bioinformatics*, 2009, **10**, 295.

39 E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry and G. Sherlock, *Bioinformatics*, 2004, **20**, 3710–3715.

40 J. L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, L. A. Martinez-Cruz, F. J. Corrales and A. Rubio, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2005, **2**, 330–338.

41 H. Wang, Z. Wang, X. Li, B. Gong, L. Feng and Y. Zhou, *Algorithms for Molecular Biology*, 2011, **6**, 14.

42 P. M. Ismail, T. Lu and M. Sawadogo, *Oncogene*, 1999, **18**, 5582–5591.

43 C. C. Benz, *Endocrine-Related Cancer*, 1998, **5**, 271–282.

44 P. J. Stephens, D. J. McBride, M.-L. L. Lin, I. Varela, E. D. Pleasance, J. T. Simpson, L. A. Stebbings, C. Leroy, S. Edkins, L. J. Mudie, C. D. Greenman, M. Jia, C. Latimer, J. W. Teague, K. W. W. Lau, J. Burton, M. A. Quail, H. Swerdlow, C. Churcher, R. Natrajan, A. M. Sieuwerts, J. W. Martens, D. P. Silver, A. Langerod, H. E. Russnes, J. A. Foekens, J. S. Reis-Filho, L. van 't Veer, A. L. Richardson, A.-L. L. Borresen-Dale, P. J. Campbell, P. A. Futreal and M. R. Stratton, *Nature*, 2009, **462**, 1005–1010.

45 I. Fernandes, Y. Bastien, T. Wai, K. Nygard, R. Lin, O. Cormier, H. S. Lee, F. Eng, N. R. Bertos, N. Pelletier, S. Mader, V. K. Han, X.-J. Yang and J. H. White, *Molecular Cell*, 2003, **11**, 139–150.

46 A. Braun, M. Denissenko, S. Kammerer, M. Nelson, R. Reneland, C. Rosette and R. Roth, *Methods for Identifying Risk of Breast Cancer and Treatments Thereof*, 2004, WO Patent App. PCT/US2003/037,989.

47 J. Zhang, Q. Liang, Y. Lei, M. Yao, L. Li, X. Gao, J. Feng, Y. Zhang, H. Gao, D. X. Liu, J. Lu and B. Huang, *Cancer Research*, 2012, **72**, 4597–4608.

48 M. Yamamoto, E. Cid, S. Bru and F. Yamamoto, *PLoS ONE*, 2011, **6**, e17149.

49 M. Yu, G. A. Smolen, J. Zhang, B. Wittner, B. J. Schott, E. Brachtel, S. Ramaswamy, S. Maheswaran and D. A. Haber, *Genes and Development*, 2009, **23**, 1737–1742.

50 S. Vimalraj, P. J. Miranda, B. Ramyakrishna and N. Selvamurugan, *Disease Markers*, 2013, **35**, 369–387.

51 S. Pollari, S. K. Leivonen, M. Perala, V. Fey, S.-M. Kakonen and O. Kallioniemi, *PLoS ONE*, 2012, **7**, e37361.

52 R. C. Zeng, W. Zhang, X. Q. Yan, Z. Q. Ye, E. D. Chen, D. P. Huang, X. H. Zhang and G. L. Huang, *Medical Oncology*, 2013, **30**, 1–8.

53 B. Song, C. Wang, J. Liu, X. Wang, L. Lv, L. Wei, L. Xie, Y. Zheng and X. Song, *Journal of Experimental and Clinical Cancer Research*, 2010, **29**, 29.

54 L. A. Shevde, B. J. Metge, A. Mitra, Y. Xi, J. Ju, J. A. King and R. S. Samant, *Journal of Cellular and Molecular Medicine*, 2010, **14**, 1693–1706.

55 A. Sudarshan, K. M. Bharat, M. A. Lisette, A. M. Eric, M. Rajesh, S. Lea, H. Miller, J. S. David, N. L. Jeffrey, E. L. Philip, D. K. Philip, M. W. Sara and A. C. David, *Nature Medicine*, 2010, **16**, 909–914.

56 L. Bhushan and R. P. Kandpal, *PLoS ONE*, 2011, **6**, e22484.

57 Q. Huang, K. Gumireddy, M. Schrier, C. le Sage, R. Nagel, S. Nair, D. A. Egan, A. Li, G. Huang, A. J. Klein-Szanto, P. A. Gimotty, D. Katsaros, G. Coukos, L. Zhang, E. Pure and R. Agami, *Nature Cell Biology*, 2008, **10**, 202–210.

58 M. Riaz, M. van Jaarsveld, A. Hollestelle, W. Prager-van der Smissen, A. Heine, A. Boersma, J. Liu, J. Helmijr, B. Ozturk, M. Smid, E. Wiemer, J. Foekens and J. Martens, *Breast Cancer Research*, 2013, **15**, R33.

59 K. D. Gerson, V. S. R. K. Maddula, B. E. Seligmann, J. R. Shearstone, A. Khan and A. M. Mercurio, *Biology Open*, 2012.

60 U. Lehmann, T. Streichert, B. Otto, C. Albat, B. Hasemeier, H. Christgen, E. Schipper, U. Hille, H. H. Kreipe and F. Langer, *BMC Bioinformatics*, 2010, **10**, 109.