Robust RFCM Algorithm for Identification of Co-expressed miRNAs

Sushmita Paul and Pradipta Maji
Machine Intelligence Unit, Indian Statistical Institute, 203, B.T. Road, Kolkata, 700108, India
E-mail: {sushmita_t, pmaji}@isical.ac.in

***Abstract*—MicroRNAs (miRNAs) are short, endogenous RNAs having ability to regulate gene expression at the post-transcriptional level. Various studies have revealed that miRNAs tend to cluster on chromosomes. Members of a cluster that are at close proximity on chromosome are highly likely to be processed as cotranscribed 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 create a big challenge for scientists to decipher this huge expression data. In order to extract meaningful information from expression data, this paper presents the application of robust rough-fuzzy c-means (rRFCM) algorithm to discover co-expressed miRNA clusters. The rRFCM algorithm comprises a judicious integration of rough sets, fuzzy sets, and c-means algorithm. The effectiveness of the rRFCM algorithm and different initialization methods, along with a comparison with other related methods, is demonstrated on three miRNA microarray expression data sets using Silhouette index, Davies-Bouldin index, Dunn index, β index, and gene ontology based analysis.**

***Index Terms*—microRNA, Rough Sets, Fuzzy Sets, Clustering.**

I. Introduction

The microRNAs (miRNAs), a class of short approximately 22-nucleotide non-coding RNAs found in many plants and animals, often act post-transcriptionally to inhibit mRNA expression. Hence, miRNAs are related to diverse cellular processes and regarded as important components of the gene regulatory network. 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]. Existence of co-expressed miRNAs is also demonstrated using expression profiling analysis in [2]. 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 cotranscribed units. Expression data of miRNAs can be used to detect clusters of miRNAs as it is suggested that co-expressed miRNAs are cotranscribed, 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. However, the property of mRNA to get targeted by several miRNAs makes biological networks more complicated and thus greatly increases the challenges of comprehending and interpreting the resulting mass of data. A first step towards addressing this challenge is the use of clustering techniques. The co-expressed miRNAs, that is, miRNAs with similar expression patterns and are cotranscribed, 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. One of the main problems in expression data 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 the fuzzy set theory and rough set theory have gained popularity in modeling and propagating uncertainty. Both fuzzy set and rough set provide a mathematical framework to capture uncertainties associated with human cognition process [3].

In this paper, the application of a newly developed hybrid algorithm, called robust rough-fuzzy c-means (rRFCM) [4], is presented for clustering miRNA expression data. The rRFCM has been used to group functionally similar genes from gene microarray data [4]. Here, it is used to find co-expressed miRNAs. It integrates judiciously the merits of rough sets, and probabilistic and possibilistic memberships of fuzzy sets. 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. 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 lower bound of the rRFCM algorithm differentiates it from the RFCM [3] algorithm used in [5]. The earlier algorithm has possibilistic lower bound, while the later has crisp lower bound. In addition to this, the rRFCM has better mechanism to decide the size of granules used in the clustering algorithm. An efficient method developed in [4] is used to select initial prototypes of different miRNA clusters; thereby circumventing the initialization and local minima problems of c-means algorithm. The performance of different methods is demonstrated on a set of three miRNA expression data sets using some standard validity indices.

II. Robust RFCM Algorithm

This section reports a new c-means algorithm, termed as robust rough-fuzzy c-means (rRFCM). 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

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

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. According to the definitions of lower approximation and boundary of rough sets, 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 ν_{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 fuzzy c-means (FCM). Whereas, 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 rRFCM, the membership values of objects in lower approximation are identical to the possibilistic c-means (PCM), while those in boundary region are the same as the FCM, and are as follows:

$$\mu_{ij} = \left[\sum_{k=1}^{c}\left(\frac{{d_{ij}}^2}{{d_{kj}}^2}\right)^{\frac{1}{\acute{m}_1-1}}\right]^{-1}; \nu_{ij} = \left[1+\left\{\frac{{d_{ij}}^2}{\eta_i}\right\}^{\frac{1}{(\acute{m}_2-1)}}\right]^{-1}$$

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

$$\text{also } 0 < \sum_{j=1}^{n}\nu_{ij} \leq n, \forall i; \text{ and } \max_i \nu_{ij} > 0, \forall j;$$

where η_i is the scale parameter. The centroid for the rRFCM is computed as:

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

$$\text{where } \mathcal{C}_1 = \frac{\sum_{x_j \in \underline{A}(\beta_i)} (\nu_{ij})^{\acute{m}_2} x_j}{\sum_{x_j \in \underline{A}(\beta_i)} (\nu_{ij})^{\acute{m}_2}}; \quad \mathcal{D}_1 = \frac{\sum_{x_j \in B(\beta_i)} (\mu_{ij})^{\acute{m}_1} x_j}{\sum_{x_j \in B(\beta_i)} (\mu_{ij})^{\acute{m}_1}}.$$

The process starts by choosing c objects as the initial centroids of the c clusters. The possibilistic memberships of all the objects are calculated. Let $\nu_i = (\nu_{i1}, \cdots, \nu_{ij}, \cdots, \nu_{in})$ represents the possibilistic cluster β_i associated with the centroid v_i. After computing ν_{ij} for c clusters and n objects, the values of ν_{ij} for each object x_j are sorted and the difference of two highest memberships of x_j is compared with a threshold value δ_1. Let ν_{ij} and ν_{kj} be the highest and second highest memberships of x_j. If $(\nu_{ij} - \nu_{kj}) > \delta_1$, then $x_j \in \underline{A}(\beta_i)$, otherwise $x_j \in B(\beta_i)$ and $x_j \in B(\beta_k)$ if $\nu_{ij} > \delta_2$. After assigning each object in lower approximations or boundary regions of different clusters based on the thresholds δ_1 and δ_2, the probabilistic memberships μ_{ij} for the objects lying in the boundary regions are computed. The new centroids of different clusters are computed as per (1). The thresholds δ_1 and δ_2 control the size of granules of rough-fuzzy clustering. In practice, the following definitions work well:

$$\delta_1 = \frac{1}{n}\sum_{j=1}^{n}(\nu_{ij} - \nu_{kj}) \qquad \delta_2 = \frac{1}{\acute{n}}\sum_{j=1}^{\acute{n}}\nu_{ij} \quad (2)$$

where n is the total number of miRNAs, ν_{ij} and ν_{kj} are the highest and second highest memberships of object x_j. On the other hand, the miRNAs with $(\nu_{ij} - \nu_{kj}) \leq \delta_1$ are used to calculate the threshold δ_2; where $\acute{n}$ is the number of miRNAs those do not belong to lower approximations of any cluster and ν_{ij} is the highest membership of miRNA x_j.

To generate initial cluster prototypes of c-means algorithms, the initialization method reported in [4] is used, where a quantitative measure, called degree of similarity, is used to evaluate the similarity between two objects. The degree of similarity (DOS) between two objects x_i and x_j is defined as:

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

where m is the number of features of the object x_i, k_{max} and k_{min} denote the maximum and minimum values along the kth feature, respectively. If expression values of two miRNAs are different, the DOS between them is small. A high value of $\text{DOS}(x_i, x_j)$ between two miRNAs x_i and x_j asserts that they may have similar expression patterns and are likely to be involved in same biological process. If two miRNAs are same, the DOS between them is maximum, that is, $\text{DOS}(x_i, x_i) = 1$. Hence, $0 \leq \text{DOS}(x_i, x_j) \leq 1$. Also, $\text{DOS}(x_i, x_j) = \text{DOS}(x_j, x_i)$.

III. Experimental Results and Discussions

In this section, the performance of the rRFCM algorithm is compared with that of hard c-means (HCM), fuzzy c-means (FCM), rough-fuzzy c-means (RFCM) [3], cluster identification via connectivity kernels (CLICK), and self organizing map (SOM) on three miRNA microarray data sets, which are downloaded from *Gene Expression Omnibus* (www.ncbi.nlm.nih.gov/geo/) with accession numbers GSE17155, GSE29495, and GSE35074. For each data set, the number of c is decided by using CLICK algorithm. The weight parameter w for robust rough-fuzzy clustering is set to 0.99, while the values of fuzzifiers $\acute{m}_1 = 2.0$ and $\acute{m}_2 = 2.0$. The major metrics for evaluating the performance of different algorithms are Silhouette index, Davies-Bouldin index, Dunn index, and β index. Also, the biological analysis of the obtained miRNA clusters is studied using GO Term Finder.

TABLE I
COMPARATIVE PERFORMANCE OF DIFFERENT INITIALIZATION METHODS

Different Data Sets	Initial Centers	Silhouette Index				DB Index				Dunn Index			
		HCM	FCM	RFCM	rRFCM	HCM	FCM	RFCM	rRFCM	HCM	FCM	RFCM	rRFCM
GSE17155	Random	0.199	0.180	0.243	0.429	1.638	8.928	1.375	1.203	0.326	0.001	0.693	0.225
	PC	0.224	0.135	0.270	0.196	1.607	2.322	1.168	1.770	0.414	0.001	0.743	0.282
	DOS	0.210	-1.000	0.233	0.303	**1.557**	7.372	**1.131**	**0.885**	**0.606**	**0.002**	**0.899**	**1.252**
GSE29495	Random	0.225	0.365	0.660	0.385	1.293	4.437	0.397	1.493	0.000	0.015	0.127	0.054
	PC	0.608	0.571	0.644	0.853	0.426	0.686	0.525	0.459	0.022	0.076	0.064	0.550
	DOS	**0.820**	0.443	**0.780**	**0.899**	**0.199**	0.916	**0.198**	**0.115**	**2.103**	0.022	**3.899**	**4.500**
GSE35074	Random	0.047	0.365	0.101	0.190	3.944	9.000	2.369	1.848	0.184	0.000	0.211	0.282
	PC	0.139	0.504	0.197	0.338	3.485	9.641	1.774	1.497	0.275	0.000	0.373	0.488
	DOS	0.081	0.365	0.133	**0.374**	3.660	**7.313**	**1.414**	**0.885**	**0.317**	**0.000**	**0.781**	**1.161**

A. Optimum Values of Parameters

The threshold λ plays an important role to generate the initial cluster centers. It controls the degree of similarity among the miRNAs present in microarray data. In effect, it has a direct influence on the performance of the initialization method, used in [4]. Let S=$\{\lambda\}$ represents the set of parameter and S^*=$\{\lambda^*\}$ a set of optimal parameter. To find out the optimum set S^*, containing optimum value of λ^*, the Dunn cluster validity (D) index is used. For three miRNA microarray data sets, the value of λ is varied from 0.50 to 1.0. The optimum value of λ^* for each microarray data set is obtained using the following relation:

$$S^* = \arg\max_{S}\{\mathrm{D}\} \tag{4}$$

The optimum values of λ obtained using (4) are 0.90, 0.92, 0.85 for GSE17155, GSE29495, and GSE35074 data sets.

B. Performance of Different Initialization Methods

Table I provides the comparative results of different c-means algorithms with random initialization method, initialization of centroids reported in [5] and initialization method described in Section II for three miRNA microarray data sets. Last two initialization methods are similar from algorithmic point of view. The difference lies in quantification of similarity between two miRNAs. The earlier initialization method uses pearson correlation (PC) for similarity measurement, while later uses the DOS. The best results of each c-means clustering algorithm are reported for their optimal λ values. In most of the cases, the DOS based initialization method is found to improve the performance in terms of Silhouette index, DB index, and Dunn index for all c-means algorithms. Out of 36 comparisons, the DOS based initialization method is found to provide significantly better results in 24 cases, which are marked bold, compare to the random initialization method and initialization method developed in [5]. While the random initialization method and PC based initialization method generates better result in 2 and 10 cases, respectively. From the Table I, it can also be seen that the HCM algorithm with the DOS based initialization method outperforms the rRFCM algorithm with random initialization method in 5 cases. The FCM algorithm with the DOS based initialization method performs better in 3 cases compare to the rRFCM algorithm with random initialization method. On the other hand, the RFCM algorithm with the DOS based initialization method performs better in 7 cases compare to the rRFCM algorithm with random initialization method. Also, the HCM algorithm with DOS based initialization method performs better than the rRFCM algorithm with the PC based initialization method in 5 cases, while FCM with DOS based initialization method performs better in only 1 case compare to the rRFCM algorithm with PC based initialization method. Moreover, the RFCM algorithm with DOS based initialization method performs better than the rRFCM with PC based initialization method in 6 cases. The better performance of the DOS based initialization method is achieved due to the fact that it enables the algorithm to converge to an optimum solution.

C. Performance of Different C-Means Algorithms

Table II present the performance of different clustering algorithms for optimum values of λ. The results and subsequent discussions are presented with respect to the performance of Silhouette index, DB index, Dunn index, and β index. From Table II, it can be observed that the rRFCM algorithm outperforms other clustering algorithms in terms of Silhouette index, DB index, Dunn index, and β index. Again, the results establish the fact that the rRFCM algorithm is superior to other c-means clustering algorithms.

The best performance of the rRFCM is achieved due to the following reasons: (1) the DOS based similarity measure used for initialization of centroids enables the algorithm to converge to an optimum solution; (2) probabilistic membership function of the rRFCM handles efficiently overlapping partitions, while the possibilistic membership function of lower region of a cluster helps to discover arbitrary shaped cluster; and (3) the concept of possibilistic lower bound and fuzzy boundary of the rRFCM algorithm deals with uncertainty, vagueness, and incompleteness in class definition.

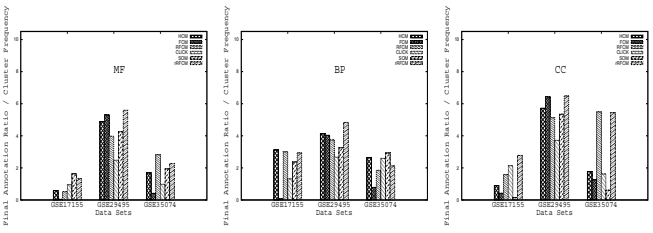

Fig. 1. Biological annotation ratios of different algorithms

D. Functional Consistency of Clustering Result

DIANA microT v3.0 [6], a miRNA target prediction algorithm is used to predict miRNA target genes for all miRNA clusters generated by different clustering algorithms. For each

TABLE II
PERFORMANCE OF DIFFERENT c-MEANS ALGORITHMS

Index	Algorithms	GSE17155	GSE29495	GSE35074
Silhouette Index	HCM	0.210	0.820	0.081
	FCM	-1.000	0.443	0.365
	RFCM	0.233	0.780	0.133
	CLICK	-0.103	-0.633	0.034
	SOM	-0.112	-0.536	0.003
	rRFCM	**0.303**	**0.899**	**0.374**
DB Index	HCM	1.557	0.200	3.660
	FCM	0.000	0.916	6.313
	RFCM	1.131	0.198	1.414
	CLICK	7.722	134.694	4.884
	SOM	34.949	1641.271	13.435
	rRFCM	**0.885**	**0.115**	**0.885**
Dunn Index	HCM	0.606	2.103	0.317
	FCM	1.000	0.022	0.000
	RFCM	0.899	3.899	0.781
	CLICK	0.115	0.000	0.159
	SOM	0.016	0.000	0.087
	rRFCM	**1.252**	**4.500**	**1.161**
β Index	HCM	8.181	22.450	1.632
	FCM	1.000	15.109	1.003
	RFCM	6.689	25.322	1.226
	CLICK	1.809	1.039	1.492
	SOM	1.210	1.033	1.433
	rRFCM	**12.812**	**114.387**	**2.553**

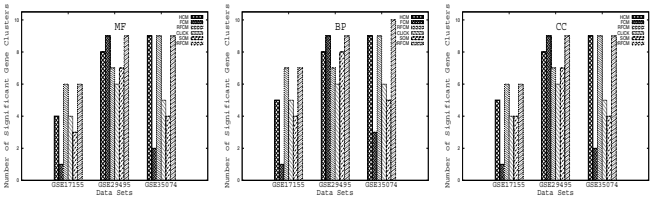

Fig. 2. Significant gene clusters obtained using different algorithms

cluster of miRNAs generated by different algorithms, genes that are targeted by at least 75% miRNAs in a cluster are used for further analysis. In order to evaluate the functional consistency of the genes targeted by miRNAs of a cluster, which are generated by different clustering algorithms, the biological annotations of those genes of different clusters are considered in terms of the GO. The annotation ratios of each targeted gene cluster in three GO ontologies namely, molecular function (MF), biological processes (BP), and cellular components (CC) are calculated using the GO Term Finder. The GO term is searched in which most of the genes of a particular cluster are enriched. 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 closer to 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 represents that the genes are better clustered by function, indicating a more functionally consistent clustering result.

Genes that are targeted by 75% and above miRNAs in a cluster are analyzed and the results are reported in Fig. 1. The final annotation ratios generated by all algorithms for MF, BP, and CC ontologies on three miRNA data sets are shown in this figure. All the results reported here confirm that the rRFCM provides higher or comparable final annotation ratios than that obtained using other algorithms in most of the cases. The rRFCM algorithm attains higher final annotation ratio than that obtained using other clustering algorithms in 1, 1, and 2 cases for the MF, BP, and CC ontologies, respectively. On the other hand, the HCM and SOM generate higher value of final annotation ratio in 2 cases and 1 case for BP and MF ontologies, respectively. Also, the RFCM achieves higher value of final annotation ratio in GSE35074 data set for the MF and CC ontologies.

E. Biologically Significant Gene Clusters

This section presents the comparative performance analysis of different clustering algorithms, in terms of number of significant gene clusters generated. Fig. 2 presents the results for the MF, BP, and CC ontologies on three miRNA microarray data sets. The GO Term Finder is used to determine the statistically significant gene clusters, which are targets of at least 75% miRNAs of that cluster produced by different clustering 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. Fig. 2 presents the comparative results of different clustering algorithms for the MF, BP, and CC ontologies. From the results, it is seen that the rRFCM generates more or comparable number of significant gene clusters compare to other clustering algorithms for all ontologies.

IV. CONCLUSION

In this paper, the application of the rRFCM algorithm in miRNA clustering has been demonstrated. Integration of the merits of rough sets, fuzzy sets, and c-means algorithm generates better results as compared to other c-means algorithms. The effectiveness of the rRFCM algorithm, along with a comparison with other algorithms, is demonstrated on three miRNA 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, "Clustering and Conservation Patterns of Human microRNAs," *Nucleic Acids Research*, vol. 33, pp. 2697–2706.

[2] S. Baskerville and D. P. Bartel, "Microarray Profiling of microRNAs Reveals Frequent Coexpression with Neighboring miRNAs and Host Genes," *RNA*, vol. 11, pp. 241–247.

[3] P. Maji and S. K. Pal, "Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices," *IEEE Transactions on System, Man and Cybernetics, Part B, Cybernetics*, vol. 37, no. 6, pp. 1529–1540.

[4] P. Maji and S. Paul, "Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data," in *Proc. 10th Asia Pacific Bioinformatics Conf.*, Australia, 2012, pp. 307–320.

[5] P. Maji and S. Paul, "Microarray Time-Series Data Clustering Using Rough-Fuzzy C-Means Algorithm," in *Proc. 5th IEEE Intl. Conf. on Bioinformatics and Biomedicine*, Atlanta, USA, 2011, pp. 269–272.

[6] 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, "Accurate microRNA Target Prediction Correlates with Protein Repression Levels," *BMC Bioinformatics*, vol. 10, no. 295.