Appendix: 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

1 Functional Consistency of Clustering Result

In this section, the performance of the proposed rough-fuzzy clustering algorithm[1] is compared with that of hard *c*-means (HCM)[2], fuzzy *c*-means (FCM)[3], rough-fuzzy *c*-means (RFCM)[4], cluster identification via connectivity kernels (CLICK)[5], and self organizing map (SOM)[6] with respect to gene ontology. The performance of the normalized range-normalized city block distance (NRNCBD) over Pearson distance (PD) and Euclidean distance (ED) is also presented.

The genes that are targeted by at least 75% miRNAs in a cluster are analyzed and the results are reported in Fig. 1. The final annotation ratios generated by all the algorithms at their optimum values of λ and ω for molecular functions (MF), biological processes (BP), and cellular components (CC) ontologies on four miRNA microarray data sets are shown in this figure. All the results reported here confirm that the proposed clustering algorithm provides higher or comparable final annotation ratios than that obtained using several existing clustering algorithms in most of the cases.

The upper portion of Fig. 1 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 7 cases. On the other hand, the RFCM with the NRNCBD generates better results in 1, 2, and 2 cases for MF, BP, and CC ontologies, respectively.

The middle portion of Fig. 1 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 higher final annotation ratio than that obtained using other *c*-means algorithms in 3, 3, and 2 cases for the MF, BP, and CC ontologies, respectively. However, the FCM with the NRNCBD generates higher final annotation ratio in 1, 1 and 2 cases for the MF, BP, and CC ontologies, respectively.

Finally, the lower portion of Fig. 1 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 all the cases.

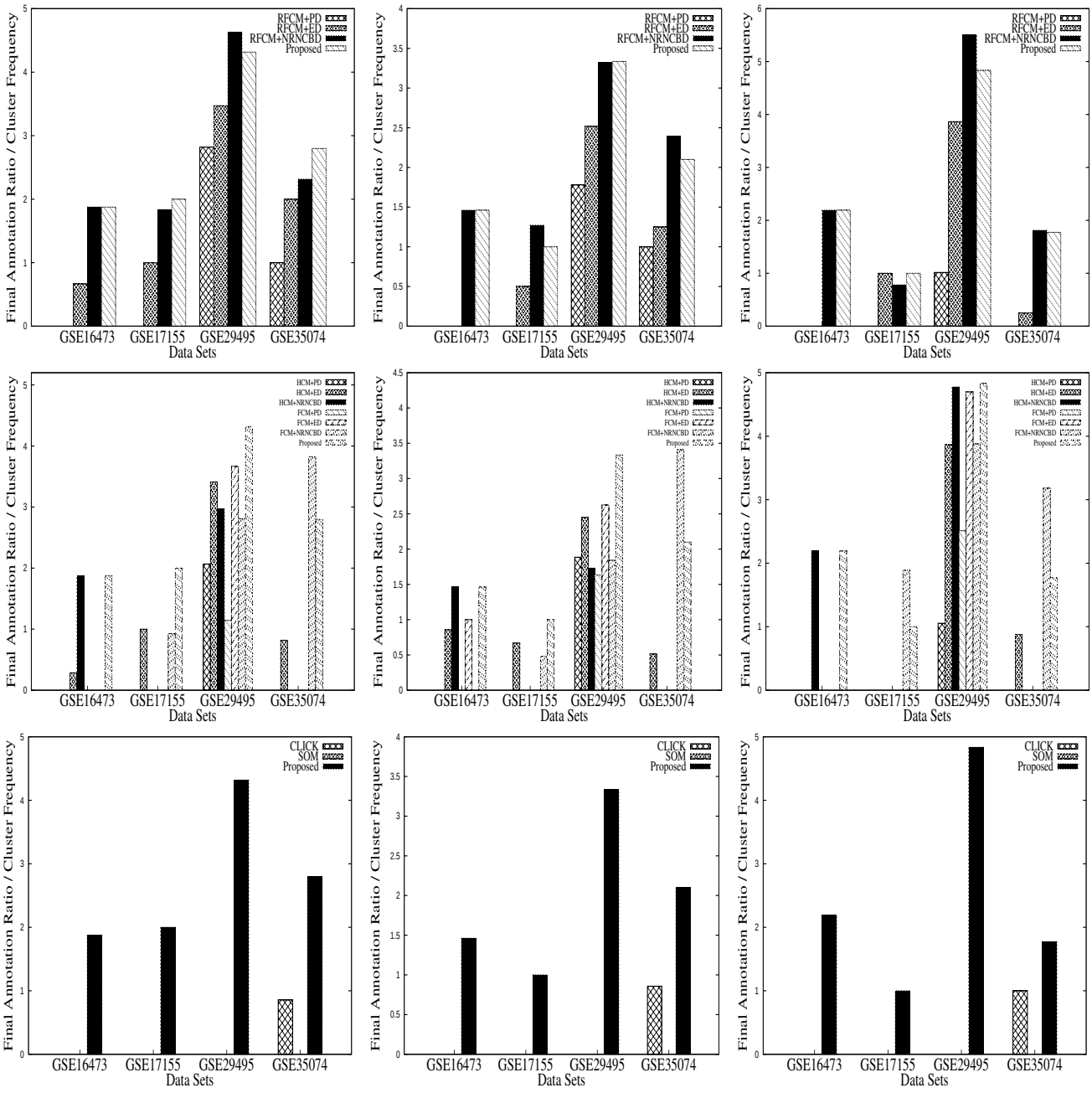
2 Biologically Significant Gene Clusters

Fig. 2 presents the comparative performance analysis of the NRNCBD, Pearson distance (PD), and Euclidean distance (ED) with respect to the proposed clustering algorithm. The significant gene clusters generated by the proposed algorithms for molecular functions (MF), biological processes (BP), and cellular components (CC) ontologies on four miRNA microarray data sets are shown in this figure. In Fig. 2, the genes that are targeted by 10 to 75% miRNAs in a cluster are presented. From Fig. 2, 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 Pearson distance and Euclidean distance in 103 cases, out of total 120 comparisons. However, the Pearson distance and Euclidean distance perform better in 4 and 13 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 both Pearson distance and Euclidean distance.

Fig. 3 reports the comparative performance analysis of different clustering algorithms with respect to the number of significant gene clusters. Results are reported for the genes that are targeted by at least 75% miRNAs in a cluster. The upper portion of Fig. 3 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 al-

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).

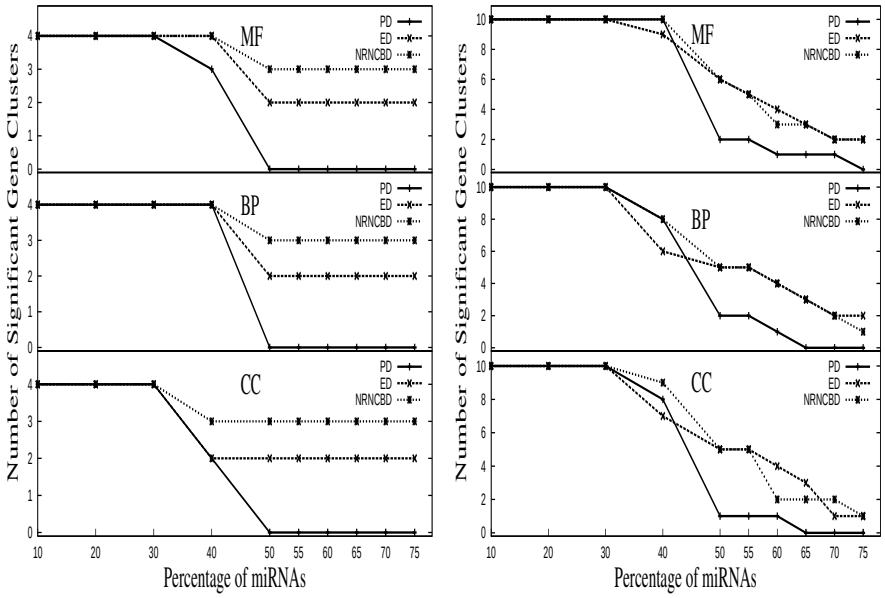


(a) Molecular Function (b) Biological Process (c) Cellular Component

Fig. 1 Biological annotation ratios obtained using different clustering algorithms for $t = 75\%$

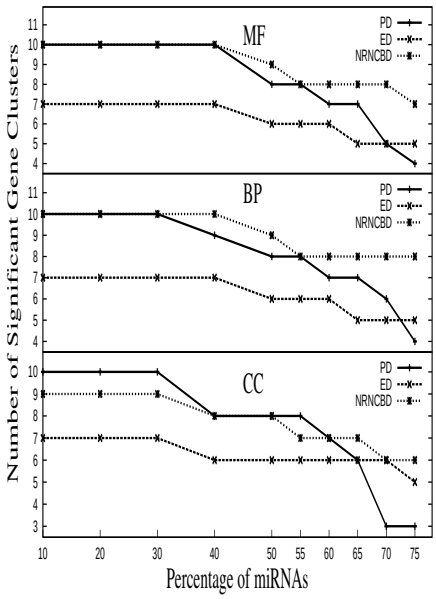
gorithm generates more or comparable number of significant gene clusters in 4, 3, and 3 cases for the MF, BP, and CC ontologies, respectively. On the other hand, the RFCM with the NRNCBD generates more number of significant gene clusters in one case each for BP and CC ontologies.

The middle portion of Fig. 3 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 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

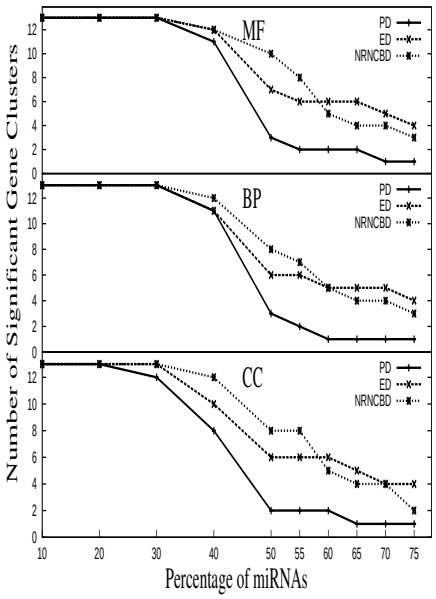


(a) GSE16473

(b) GSE17155



(c) GSE29495

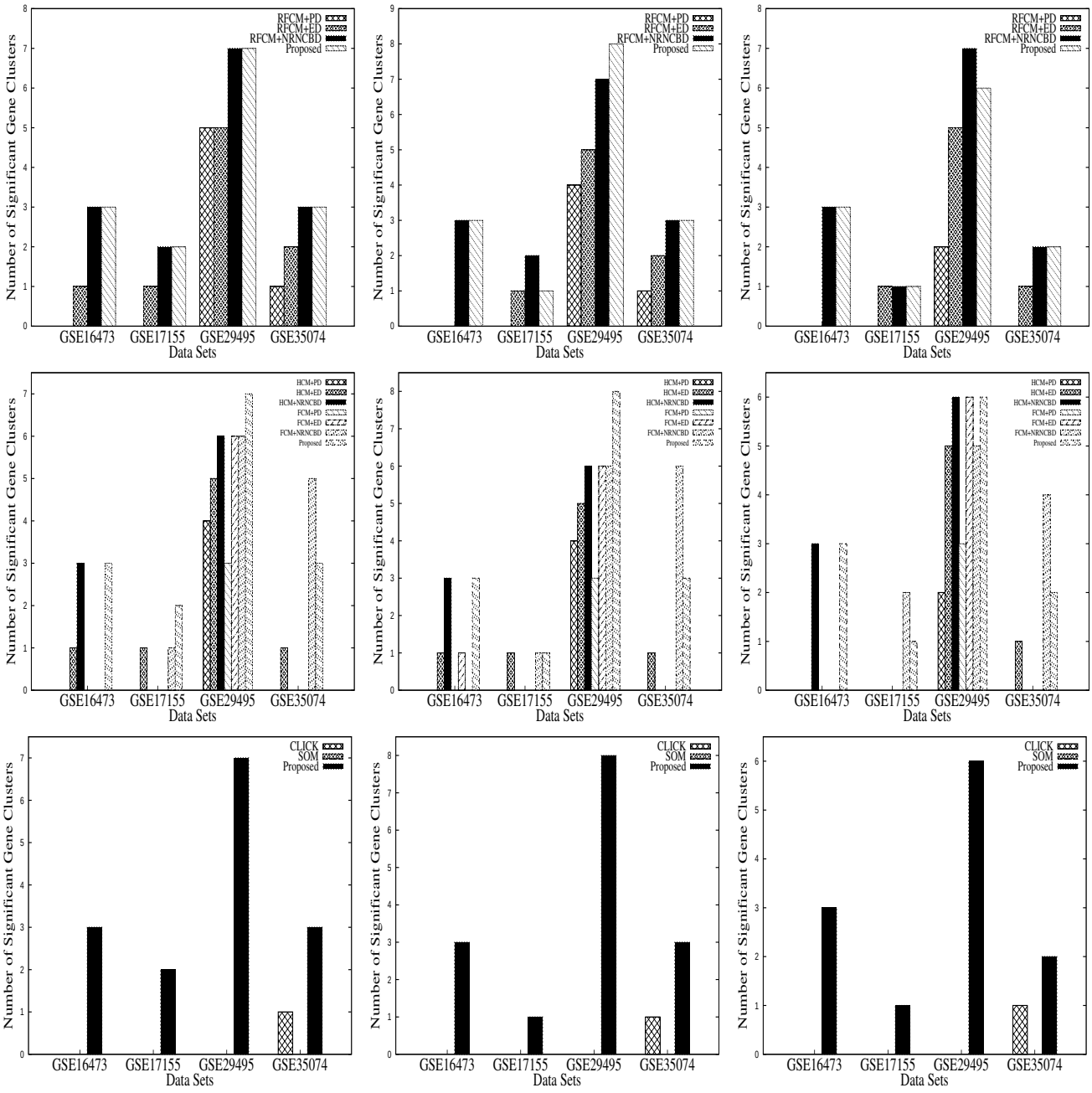


(d) GSE35074

Fig. 2 Biologically significant gene clusters obtained using the NRNCBD, Pearson and Euclidean distances

2 cases, respectively. That is, out of total 12 cases, it provides better results in 8 cases. However, the FCM algorithm with the NRNCBD generates better results in 1, 1, and 2 cases for the MF, BP, and CC ontologies, respectively.

Finally, the performance of CLICK, SOM, and proposed algorithm is compared in lower portion of Fig. 3 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 compared to both CLICK and SOM algorithms in all the cases. From Fig. 3, it can also be seen that the proposed clustering algorithm with the NRNCBD produces better results as compare to other algorithms, irrespective of the data sets, distance measures, and ontologies used. Hence, it can be concluded that the proposed clustering algorithm generates highly compact and functionally enriched clusters.



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

Fig. 3 Biologically significant gene clusters obtained by different algorithms for $t = 75\%$

References

1 S. Paul and P. Maji, *Molecular BioSystems*, 2014, 1–16.

2 L. J. Heyer, S. Kruglyak and S. Yooseph, *Genome Research*, 1999, **9**, 1106–1115.

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

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

5 R. Shamir and R. Sharan, Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, 2000, pp. 307–31.

6 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.