Appendix: Fuzzy-Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data

I. Gene Expression Data Sets

In [1], publicly available six cancer and two arthritis data sets are used. Since binary classification is a typical and fundamental issue in diagnostic and prognostic prediction of cancer and arthritis, different methods are compared using the following eight binary class data sets.

1) Breast Cancer: The breast cancer data set contains expression levels of 7129 genes in 49 breast tumor samples [2]. The samples are classified according to their estrogen receptor (ER) status: 25 samples are ER positive while the other 24 samples are ER negative.

2) Leukemia: It is an affymetrix high density oligonucleotide array that contains 7070 genes and 72 samples from two classes of leukemia [3]: 47 acute lymphoblastic leukemia and 25 acute myeloid leukemia.

3) Colon Cancer: The colon cancer data set contains expression levels of 2000 genes and 62 samples from two classes [4]: 40 tumor and 22 normal colon tissues.

4) Lung Cancer: This data set contains 181 tissue samples: among them 31 are malignant pleural mesothelioma and rest 150 adenocarcinoma of the lung [5]. Each sample is described by the expression levels of 12533 genes.

5) Breast Cancer (RBreast): In this data set, relapse or non relapse of metastases in patients after initial diagnosis for interval of at least 5 years has been classified in breast cancer patients [6]. Total 97 samples are given: 46 patients developed distance metastases within 5 years (labeled as relapse) while 51 remained healthy (labeled as non-relapse). The data set consists of 24188 genes.

6) Prostate Cancer: In this data set, 136 samples are grouped into two classes: 77 prostate tumor and 59 prostate normal samples [7]. Each sample contains 12600 genes.

7) Rheumatoid Arthritis versus Osteoarthritis (RAOA): The RAOA data set consists of gene expression profiles of thirty patients: 21 with RA and 9 with OA [8]. The Cy5-labeled experimental cDNA and the Cy3 labeled common reference sample were pooled and hybridized to the lymphochips containing $\sim$18,000 cDNA spots representing genes of relevance in immunology [8].

8) Rheumatoid Arthritis versus Healthy Controls (RAHC): The RAHC data set consists of gene expression profiling of peripheral blood cells from 32 patients with RA, 3 patients with probable RA and 15 age and sex matched healthy controls performed on microarrays with a complexity of $\sim$26K unique genes (43K elements) [9].

II. Qualitative Analysis of Supervised Clusters

For six cancer and two arthritis data sets, the best clusters generated by the proposed FRSAC algorithm [1] are analyzed using the Eisen and gene profile plots. In Eisen plot [10], the expression value of a gene in a particular sample is represented by coloring the corresponding cell of the data matrix with a color similar to the original color of its spot on the microarray. The shades of red color represent higher expression level, the shades of green color represent lower expression level and the colors toward black represent absence of differential expression values. On the other hand, the gene profile plot shows for each gene the gene expression values of that gene with respect to the samples.

Fig. 1 represents the expression values of the actual and augmented cluster representatives of the best clusters over the samples for breast, leukemia, colon cancer, RBreast, RAOA, and RAHC data sets. In Fig. 2-4[(a)-(c)], the results of best clusters obtained using the proposed clustering algorithm are reported for lung, prostate, and leukemia data sets considering the value of η as 1.2. Fig. 2-4(a) show the expression values of the actual genes or attributes of the best clusters over the samples for three data sets. Fig. 2-4(b) and Fig. 2-4(c) represent the Eisen and gene profile plots of corresponding finer clusters with gene expression values and the expression values of the augmented cluster representatives of the best clusters for three data sets. All the results reported in Fig. 1-4 establish the fact that the fuzzy-rough set based proposed supervised attribute or gene clustering algorithm can efficiently identify groups of co-regulated genes with strong association to the sample categories.

References

[1] P. Maji, "Fuzzy-Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data," *IEEE Transactions on System, Man and Cybernetics, Part B, Cybernetics*, pp. 1–12, 2010.

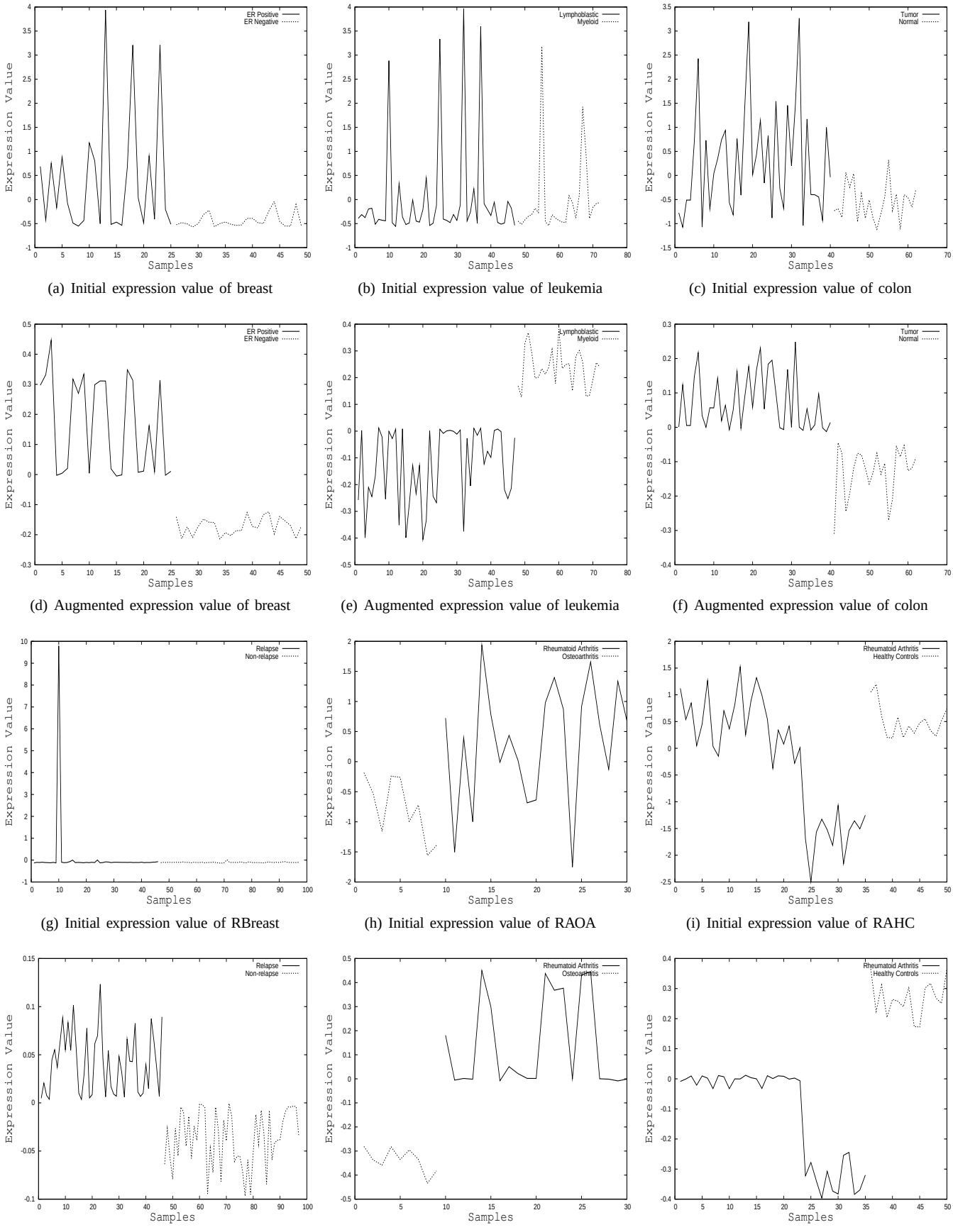
[2] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles," *Proceedings of the National Academy of Science, USA*, vol. 98, no. 20, pp. 11 462–11 467, 2001.

[3] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[4] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proceedings of the National Academy of Science, USA*, vol. 96, no. 12, pp. 6745–6750, 1999.

[5] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer And Mesothelioma," *Cancer Research*, vol. 62, pp. 4963–4967, 2002.

[6] L. J. v. Veer, H. Dai, M. J. v. D. Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. v. d. Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer," *Nature*, vol. 415, pp. 530–536, 2002.



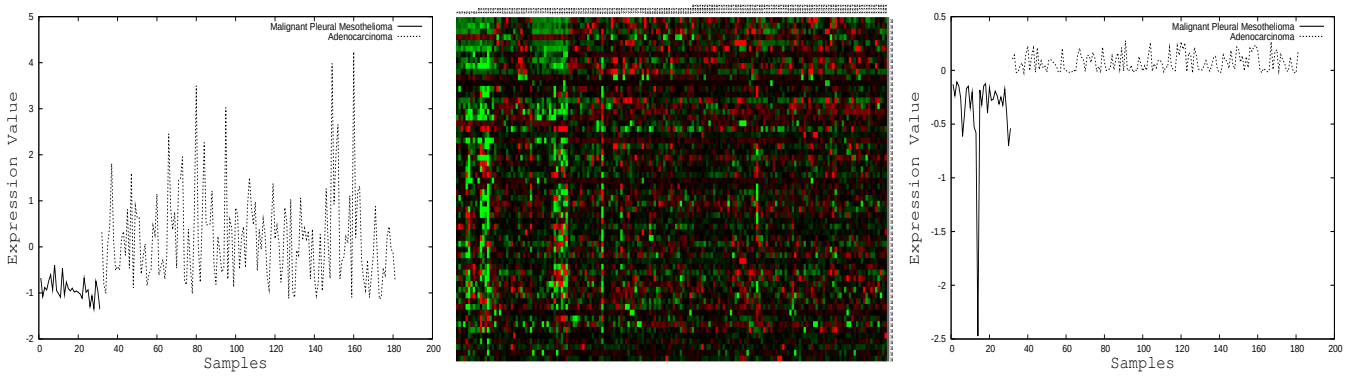
(a) Initial expression value of breast (b) Initial expression value of leukemia (c) Initial expression value of colon

(d) Augmented expression value of breast (e) Augmented expression value of leukemia (f) Augmented expression value of colon

(g) Initial expression value of RBreast (h) Initial expression value of RAOA (i) Initial expression value of RAHC

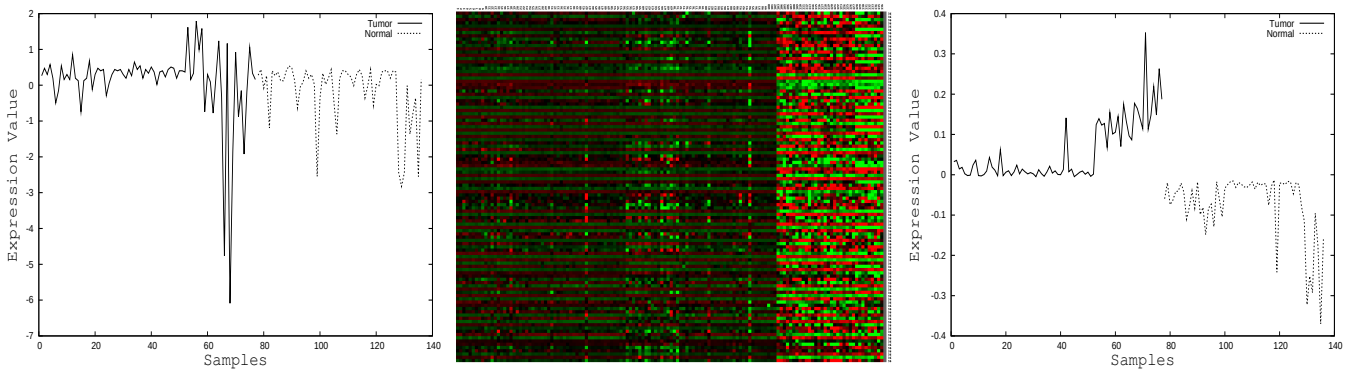
(j) Augmented expression value of RBreast (k) Augmented expression value of RAOA (l) Augmented expression value of RAHC

Fig. 1. Results obtained using proposed FRSAC algorithm for breast, leukemia, colon cancer, RBreast, RAOA, and RAHC data sets



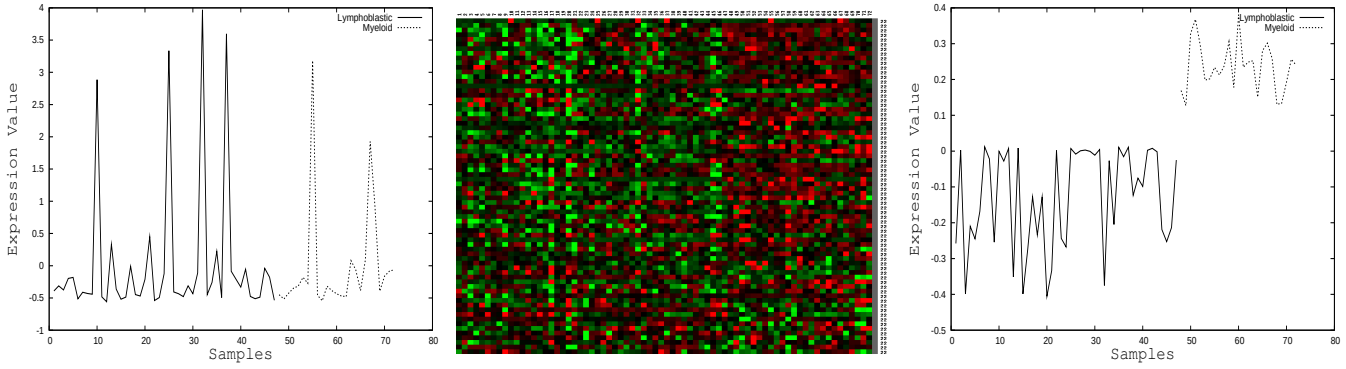
(a) Initial expression value (b) Eisen plot (c) Augmented expression value

Fig. 2. Results obtained using proposed FRSAC algorithm for lung cancer data set considering $\delta = 0.93$ and $\eta = 1.2$



(a) Initial expression value (b) Eisen plot (c) Augmented expression value

Fig. 3. Results obtained using proposed FRSAC algorithm for prostate cancer data set considering $\delta = 0.93$ and $\eta = 1.2$



(a) Initial expression value (b) Eisen plot (c) Augmented expression value

Fig. 4. Results obtained using proposed FRSAC algorithm for leukemia data set considering $\delta = 0.92$ and $\eta = 1.2$

[7] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Research*, vol. 1, pp. 203–209, 2002.

[8] T. C. T. M. van der Pouw Kraan, F. A. van Gaalen, P. V. Kasperkovitz, N. L. Verbeet, T. J. M. Smeets, M. C. Kraan, M. Fero, P.-P. Tak, T. W. J. Huizinga, E. Pieterman, F. C. Breedveld, A. A. Alizadeh, and C. L. Verweij, "Rheumatoid Arthritis is a Heterogeneous Disease: Evidence for Differences in the Activation of the STAT-1 Pathway Between Rheumatoid Tissues," *Arthritis and Rheumatism*, vol. 48, no. 8, pp. 2132–2145, 2003.

[9] T. C. T. M. van der Pouw Kraan, C. A. Wijbrandts, L. G. M. van Baarsen, A. E. Voskuyl, F. Rustenburg, J. M. Baggen, S. M. Ibrahim, M. Fero, B. A. C. Dijkmans, P. P. Tak, and C. L. Verweij, "Rheumatoid Arthritis Subtypes Identified by Genomic Profiling of Peripheral Blood Cells: Assignment of a Type I Interferon Signature in a Subpopulation of Pateints," *Annals of the Rheumatic Diseases*, vol. 66, pp. 1008–1014, 2007.

[10] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Sciences, USA*, vol. 95, no. 25, pp. 14 863–14 868, 1998.