Journal of Biomedical Informatics 42 (2009) 1022–1028

Interval based fuzzy systems for identification of important genes from microarray gene expression data: Application to carcinogenic development

Rajat K. De [a,*], Anupam Ghosh [b]

[a] Machine Intelligence Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, West Bengal, India
[b] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata 700152, India

Keywords:
Fuzzy sets
Low
Medium
High
P-value
t-Test

A B S T R A C T

In the present article, we develop two interval based fuzzy systems for identification of some possible genes mediating the carcinogenic development in various tissues. The methodology involves dimensionality reduction, classifying the genes through incorporation of the notion of linguistic fuzzy sets *low*, *medium* and *high*, and finally selection of some possible genes mediating a particular disease, obtained by a rule generation/grouping technique. The effectiveness of the proposed methodology, is demonstrated using five microarray gene expression datasets dealing with human lung, colon, sarcoma, breast cancer and leukemia. Moreover, the superior capability of the methodology in selecting important genes, over five other existing gene selection methods, viz., Significance Analysis of Microarrays (SAM), Signal-to-Noise Ratio (SNR), Neighborhood analysis (NA), Bayesian Regularization (BR) and Data-adaptive (DA) is demonstrated, in terms of the enrichment of each GO category of the important genes based on *P*-values. The results are appropriately validated by earlier investigations, gene expression profiles and *t*-test. The proposed methodology has been able to select genes that are more biologically significant in mediating the development of a disease than those obtained by the others.

1. Introduction

Cancer is a class of diseases for which a group of cells undergoes uncontrolled growth. It causes destruction of adjacent tissues (*invasion*) and sometimes spreads to other locations in the body via lymph or blood (*metastasis*). People at all ages may be affected by cancer, but the risk for most varieties increases with age. About 13% of all deaths in the world are due to cancer. According to the American Cancer Society, 7.6 million people died from cancer in the world during 2007. Nearly all cancers are caused by abnormalities in the genetic material of the transformed cells (http://en.wikipedia.org/wiki/Cancer). Various research efforts, including ones based on surgery, chemotherapy, radiotherapy, are being made to fight against cancer.

Recent studies [12,5,20] involving gene expression profiles, obtained by microarray technology, have a profound impact on cancer research. In some examples [12,5], correlations between the expression levels of a gene or a set of genes, and clinically relevant subclassifications of specific tumor subtypes have been studied. These results show that true molecular classification and substaging of multiple tumor types may be possible, leading to prognosis and patient management. Analysis of cancers using array technologies has identified subgroups of tumors that differ according to tumor types and histological subclasses, and to a lesser extent, survival among carcinogenic patients [20].

Fuzzy set theory is capable of handling uncertainty in the gene expression values arising due to incompleteness, imprecision, noise and experimental errors. The theory provides a tool for natural computing as the systems built on the theory behave like human reasoning process [41,42]. The notion of fuzzy sets has been used in the domain of gene expression analysis. These include, among others, development of rule discovery procedure by Zhang et al. [43], based on knowledge extraction of gene by classification; transformation of gene expression by fuzzy heuristic rule set [40]; classifying fuzzy inference system [22]; development of a fuzzy model for gene regulatory networks [30]; measuring performance of small rule-based classifiers using fuzzy logic [38]; identification of normal and tumor patients using fuzzy neural network model [3].

In microarray gene expression data, genes have expression values that are in some intervals under different conditions. There exists methodology [10] based on these intervals for finding genes responsible for a particular disease. Although each interval has some well defined boundary, they are highly overlapped. Thus it is better to use fuzzy set theory to handle such overlapping intervals. This fact motivates us to develop fuzzy set theoretic methods for identifying genes responsible for a particular disease.

* Corresponding author. Fax: +91 33 25783357.
E-mail addresses: rajat@isical.ac.in (R.K. De), anupam.ghosh@rediffmail.com (A. Ghosh).

Here we develop two interval based fuzzy systems to identify a set of possible genes mediating the development of cancerous growth in cell. The methodologies involve the task of dimensionality reduction, gene selection using interval ratio filter, formulation of linguistic fuzzy sets and their matching, and rule generation and grouping. Dimensionality reduction step is used to reduce the variation among the expression levels of the gene over different samples and is performed using an algorithm similar to the cyclic loess normalization algorithm [7]. Linguistic fuzzy sets, e.g., *low*, *medium* and *high* [41] are modeled using triangular membership functions in the next step, where genes are grouped into these fuzzy sets. Incorporation of linguistic variables provides a natural way of reasoning like human [41,42]. Note that the idea of using normalization algorithm for dimensionality reduction and modeling expression profiles of various genes using linguistic fuzzy sets is novel in this article. An existing rule generation and rule grouping algorithm [10] is used for both the methods in the final step for finding a set of possible genes mediating the development of cancer in human cells. Existing gene selection methods like SAM [13], SNR [37], NA [15], BR [32,8], DA [27] that do not include the notion of fuzzy sets is used for comparative analysis. In this article, we consider five gene expression datasets, viz., human lung gene expression data [4], human colon expression data [1], human breast expression data [24], human soft tissue sarcoma data [11], human lymphocytes and plasma cell expression data [16]. As already discussed, there exist many articles in literature for microarray gene expression analysis. Here we have considered the reference [10] as it deals with association rule mining and provides an algorithm for rule grouping that is effective in the area of gene expression analysis.

Finally, we present a set of possible genes obtained by both these methods, which may be responsible for cancerous growth in human cell. The results are appropriately validated by some earlier investigations and gene expression profiles, and compared using *t*-test and number of enriched attributes. Moreover, the proposed methodology has found more true positives than the existing ones in identifying responsible genes.

2. Methodology

In this paper, we have developed two methods, called Linguistic Fuzzy Rule Generation and Grouping (LFRGG) and Expression Interval based Rule Generation and Grouping (EIRGG) for the interval based fuzzy system and applied to the gene expression datasets to find out some possible genes mediating the development of cancer. In LFRGG, we have used cyclic loess normalization like algorithm for dimensionality reduction, concept of fuzzy sets, and the notion rule generation and grouping. Note that although the purpose of normalization is different from that are widely used in practice, we have used this kind of technique for transforming several expression values of a gene over a number of samples into a single expression value for the said gene. This reduces the computational complexity of the proposed methodology to a great extent. The entire methodology is described in details below.

2.1. Linguistic Fuzzy Rule Generation and Grouping (LFRGG)

The proposed methodology, called LFRGG, has a few stages. Each of these stages is described below.

2.1.1. Stage I — dimensionality reduction

The need for normalization arises naturally when we deal with experiments involving multiple arrays. Here we use an algorithm similar to a normalization algorithm, called cyclic loess, [7] to the dataset for normal samples as well as that for tumor samples. However, the purpose of normalization is to deal with this obscuring variation.

The algorithm transforms expression values of a gene over a number of samples into a single expression value, corresponding to normal as well as to tumor samples separately. It is based on the idea of creating an M versus A plot, where M is the difference in (log) expression values and A is the average of the (log) expression values. An M versus A plot for normalized data should show a point cloud scattered about the $M = 0$ axis. In particular, for any two arrays i, j with expression values x_{ki} and x_{kj}, $k = 1, \ldots, p$ being the gene index, we calculate $M_k = \log_2(x_{ki}/x_{kj})$ and $A_k = 1/2\log_2(x_{ki}x_{kj})$. A normalization curve is used to fit to these M versus A plot. For details of cyclic loess, one may refer to [7]. Note that the idea of using normalization algorithm in dimensionality reduction is novel. The steps of the algorithm is mentioned below.

Algorithm:

For each gene, do

Step 1: choose pair wise samples.

Step 2: compute M_k and A_k for each pair.

Step 3: fit M_k with respect to A_k. Here we are going for parabolic curve fitting. So for a set of A_k values that we have defined in the previous step, we get a set of estimated $\widehat{M}_k$ values. Finally, we get $\binom{m}{2}$ number of $(M_k - \widehat{M}_k)$ values, for m samples. We call these $(M_k - \widehat{M}_k)$ values as *adjustments* for m samples.

Step 4: record these *adjustments* for each sample and compute the *resultant adjustment* for each sample.

Step 5: update the old (log) expression value for each sample by the following formula

$$newx_{ki} = oldx_{ki} + resultant\ adjustment.$$

Step 6: repeat Steps 1 through 5 until the differences among the (log) expression values are less than some threshold values specified by the analyzer (i.e., repeat these steps until the (log) expression values of different samples are close enough).

2.1.2. Stage II — formulation of linguistic fuzzy sets and their matching

This stage is a two-step process — Step 1 and Step 2. In conventional statistical methods, the absolute expression pattern of genes is presented to a system for further computations. However, in real life situations, gene expression pattern may be uncertain and/or incomplete. In such cases it may be convenient to use linguistic variables such as *low*, *medium*, *high* to replace numerical expression values [28,29]. This transformation is capable of handling absolute expression pattern, i.e., numerical and linguistic forms of the input data. Any input expression value can be described through a combination of membership values in the linguistic property sets *low*, *medium* and *high*. Note that incorporation of linguistic fuzzy sets provides a tool for natural computing [41,42] as the resulting system is capable of reasoning like human. This is due to the fact that we (human) often measure a quantity in terms of *low*, *medium* or *high*, and make the subsequent decision accordingly. The idea of modeling gene expression profiles using linguistic fuzzy sets is novel. Using these fuzzy sets, we are partitioning domain of expression values of a gene into three (Fig. 1 in Supplementary material). However, one may consider four, five or even more linguistic fuzzy sets to partition the domain into four, five or more.

Step 1: formulation of linguistic fuzzy sets.

Each input expression value x_j of jth gene in quantitative form can be expressed in terms of membership values to each of the three linguistic properties *low*, *medium* and *high*. That is, a 3-d membership vector for the fuzzy sets *low*, *medium* and *high* corresponding to x_j is given by

$$\mathbf{v}_j = [U_{low}(x_j), U_{medium}(x_j), U_{high}(x_j)]^T.$$

Here $U_{low}(x_j)$ is the membership value of the jth gene with expression value x_j to the fuzzy set *low*. Similarly, $U_{medium}(x_j)$ and $U_{high}(x_j)$ are defined accordingly. Therefore an n-dimensional gene expression pattern for n genes $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ can be represented as a $3n$ dimensional vector

$$\mathbf{v} = [U_{low}(x_1), U_{medium}(x_1), U_{high}(x_1), \ldots, U_{low}(x_n), U_{medium}(x_n), U_{high}(x_n)]^T.$$

Mathematical formulations of the membership functions U's are described in the Supplementary material.

Step 2: matching.

After representing the genes with three linguistic variables, we group the genes based on their membership values into *low*, *medium* or *high*. That is, a gene with membership value to *low* greater than 0.5, is considered as a member of the fuzzy set *low*. Thus we have got three classes of genes in *low*, *medium* and *high*. This process is executed both on normal and tumor samples separately. However, the values of the parameters (of these fuzzy sets), for both normal and tumor samples, are computed based on the normal samples only.

Now we perform matching among these classes. We first match the class *low* of normal with the classes *medium* and *high* of tumor samples. The main significance of this type of matching is that we want to know the genes of class *low* of normal, which move to the class *medium* or *high* of tumor. These genes are identified as the over expressed genes. Similarly, we perform matching of class *medium* (*high*) of normal with the classes of *low* (*low*) and *high* (*medium*) of tumor. Thus we have identified the genes that have changed their classes from normal to tumor samples.

2.1.3. Stage III – rule generation and grouping

Here we describe the technique in [10] that is adopted in this article for rule generation and grouping. Note that it involves grouping of similar rules to get top-k rule groups, where $k = 1$ has been assumed in the present article.

Step 1: generation of intervals, item support set and gene support set.

After selecting the genes in Stage II, we generate intervals of the gene expression values. The corresponding genes form a set $\mathbf{R}$. For a jth gene, the corresponding interval of gene expression values is an item I_j. Let $\mathbf{I} = \{I_1, I_2, \ldots, I_p\}$ be the complete set of such items. Each item could be one of the two types, i.e., normal or tumor. As a mapping between genes and items, we define the item support set, denoted by $\mathbf{R}(\widehat{\mathbf{I}})$ which is the largest set of genes that contain $\widehat{\mathbf{I}}$ (a subset of $\mathbf{I}$). Likewise, we define the gene support set, denoted by $\mathbf{I}(\widehat{\mathbf{R}})$ as the largest set of items common among the genes in $\widehat{\mathbf{R}}$ (a subset of $\mathbf{R}$).

Step 2: rule generation.

In Step 2 of Stage III, we generate some rules, where each rule is of the form $\mathbf{A} \rightarrow C$. Here $\mathbf{A}$ is the subset of $\mathbf{I}$ and forms the antecedent, and C (normal or tumor) forms the consequent.

Step 3: rule grouping.

Here we find a compact set of rules that can describe an exhaustive set of items. If the number of each such rules covering all the items is large, it may create confusion in the decision making process. On the other hand, less number of rules unable to cover all the items exhaustively will be of no use. Thus the rules we have obtained by the above method need to be checked for their appropriateness. They often need to be grouped to form an optimal set of rules, based on certain measures called support (*SUP*) and confidence (*CONF*).

SUP of a rule $\mathbf{A} \rightarrow C$ is the ratio of the number of genes whose expression values lie in an interval in $\mathbf{A}$ for C type (i.e., either normal or tumor) data to the total number of genes in both normal and tumor. *CONF* of a rule $\mathbf{A} \rightarrow C$ is the ratio of the number of genes lying in an interval $\mathbf{A}$ for C type data to those lying in the same interval for both normal and tumor samples [10].

As mentioned earlier, the idea of rule grouping helps one to reduce the number of rules discovered, by identifying rules that come from the same set of genes. For example, if $\mathbf{R}(I_1) = \mathbf{R}(I_2) = \mathbf{R}(I_3) = \mathbf{R}(I_1, I_2) = \mathbf{R}(I_1, I_3) = \mathbf{R}(I_2, I_3) = \mathbf{R}(I_1, I_2, I_3)$, then they make up rule groups

$$I_1 \rightarrow C, I_2 \rightarrow C, \ldots, (I_1, I_2, I_3) \rightarrow C$$

for the same consequent C with the supremum $(I_1, I_2, I_3) \rightarrow C$. It is obvious that all the rules in the same rule group have the same support *SUP* and confidence factor *CONF* since they are essentially derived from the same subset of genes. Based on the supremum of a rule group, it is easy to identify the remaining members. Now to evaluate the significant rule group, we use the criteria that rule group 1, denoted by rg_1, is more significant than rule group 2, denoted by rg_2, if $(rg_1.CONF > rg_2.CONF)$ *OR* $(rg_1.SUP > rg_2.SUP$ *AND* $rg_1.CONF = rg_2.CONF)$. Here, $rg_i.CONF$ and $rg_i.SUP$, indicate values of *CONF* and *SUP* of an ith rule group, respectively. Then we select the genes that are covered by the most significant rule group. For details of this rule grouping algorithm, one may refer to [10].

2.2. Expression Interval based Rule Generation and Grouping (EIRGG)

We now describe another method, called Expression Interval based Rule Generation and Grouping (EIRGG) [10], for the same purpose. EIRGG involves the tasks like interval generation, rule generation using these intervals and grouping of these rules using *SUP* and *CONF* factors [10] defined above. Let us consider a gene expression dataset D consisting of a set $\mathbf{R}$ of rows, $r_1, r_2, \ldots, r_n$ corresponding to n genes. Let $\mathbf{I} = \{i_1, i_2, \ldots, i_m\}$ be the complete set of items of dataset D; each item represents an interval of gene expression levels. Data in D are distributed in k different classes $C_1, C_2, \ldots, C_k$. For example, if there are two types of data, i.e., normal and tumor, then $k = 2$.

2.2.1. Stage I – generation of intervals

Let us assume that the dataset consists of n number of genes (number of rows) and d number of samples (number of columns). Also assume that the number of classes (i.e., C_1 for normal and C_2 for tumor) is two, i.e., $k = 2$. Among these d samples let d_1 be the number of normal samples and d_2 is that of tumor samples.

Step 1: identification of overlapping and non-overlapping intervals.

We can represent each gene by d_1 number of values in normal dataset and d_2 number of values in tumor dataset. We assume that the minimum gene expression value of gene g_i in normal dataset is x_{imin} and the maximum gene expression value of gene g_i in normal dataset is x_{imax}. Similarly, for tumor dataset, the minimum and maximum values of the gene g_i are y_{imin} and y_{imax}, respectively. Now, we examine that whether y_{imin} is in between x_{imin} and x_{imax}. If so then we conclude that the interval of gene g_i has an overlapping region. So the interval may be broken as $y_{imin} - x_{imin}$, $y_{imax} - x_{imax}$ (if $y_{imax} > x_{imax}$) and $x_{imax} - y_{imin}$. These three regions are included into our interval set. It indicates that gene g_i has three intervals. Now, if y_{imin} does not lie between x_{imin} and x_{imax} then the interval for g_i has no overlapping region. So here the intervals are $(y_{imax} - y_{imin})$ and $(x_{imax} - x_{imin})$. Similarly, we can generate intervals for gene g_i if $y_{imax} \leqslant x_{imax}$. Using this procedure we generate all intervals for all genes. For the case $y_{imin} \leqslant x_{imin} \leqslant y_{imax}$, if $y_{imax} < x_{imax}$ then the generated intervals are $(x_{imin} - y_{imin})$, $(y_{imin} - x_{imin})$ and $(x_{imax} - y_{imax})$, and if $y_{imax} > x_{imax}$ then they are $(x_{imin} - y_{imin})$, $(x_{imax} - x_{imin})$ and $(y_{imax} - x_{imax})$.

Step 2: elimination of redundant intervals.

Now we eliminate the redundant intervals from the set of intervals obtained by the above process. If the lower and upper bound values of a pair of intervals match then we delete the interval. In this way, we eliminate the redundant intervals.

2.2.2. Stage II — selection of genes

After generating intervals, we select the genes in which the gene expression values have changed significantly from normal to tumor samples.

Step 1: selection of genes based on expression values.

In order to do this, we first choose those genes whose expression values are in non-overlapping region between normal and tumor samples. Then we define a parameter *ratio* ($0 \leqslant ratio \leqslant 1$) which is basically a ratio of the length of the overlapping region and total length of expression values (both for tumor and normal samples). We also assume that y_{imin} lies between x_{imin} and x_{imax} and $y_{imax} > x_{imax}$. So for this gene g_i an overlapped region is created $x_{imax} - y_{imin}$ and the total length of expression value of gene g_i is $y_{imax} - x_{imin}$. Thus the *ratio* becomes $(x_{imax} - y_{imin})/(y_{imax} - x_{imin})$.

Step 2: generation of intervals of selected genes.

If *ratio* is very close to zero, the overlapping region corresponding to the gene is a very small. We use a threshold value of *ratio* to select the genes. After selection of these two types of genes (i.e., genes with expression values in non-overlapping and low overlapping intervals), we obtain the intervals of the corresponding genes.

2.2.3. Stage III — rule generation and grouping

On selecting the genes and the intervals described by Stage II, we generate the rules based on these intervals. This rule generation and grouping step (i.e., Stage III of EIRGG) is identical to that of LFRGG (described in Section 2.1.3).

3. Results and discussion

In this section, the effectiveness of the proposed methods LFRGG and EIRGG is demonstrated on five cancer gene expression datasets (described in the Supplementary material), viz., human lung gene expression data [4], human colon expression data [1], human breast expression data [24], human soft tissue sarcoma data [11], human lymphocytes and plasma cell expression data [16]. The superior performance of LFRGG and EIRGG over other methods called SAM [13], SNR [37], NA [15], BR [32,8], DA [27] is also established.

3.1. Analysis of the results using LFRGG and EIRGG

Lung expression data contains 10 normal samples for expression values of 7129 genes. So there are $\binom{10}{2}$ pairs, i.e., 45 pairs for each gene. After calculating the first adjustments based on the parabolic curve fitting, we noted the required adjustment. After 5 or 6 iterations, we have got the normalized value of each sample for a gene. Then the log expression values of 10 samples become close enough. In this way we normalized 7129 genes for normal samples and tumor samples separately. After normalization, we have taken mean of the resulting expression values of the genes. Thus we represented each gene with a single value.

In this way we reduced the entire dataset through dimensionality reduction. After this we applied membership functions to these resulting gene expression values to get membership values. In order to do this, we represent each gene by three dimensions (i.e., *low*, *medium* and *high*). Now, we got three different sets of genes that belong to three different classes *low*, *medium* and *high* for normal samples and tumor samples separately. Now we perform the matching operation among the classes separately between normal and tumor samples. We found 45 genes that changed their class from normal to tumor. This is actually Step II of LFRGG.

Finally we used the rule generation followed by grouping technique on these 45 genes. In order to do this, we first generate the intervals based on the entire dataset. It results in 174 intervals. Each interval is an item I corresponding to a gene. Then we generate the rules based on this interval. The general form of the rule is $I \rightarrow C_1$ which indicates that if the gene with expression value is in the interval I then the sample is normal. Similarly, rules with consequent part representing tumor samples are obtained. We found 12 rules for normal class and 7 rules for tumor class. After this, we used the rule grouping algorithm on these two sets of rules separately. We finally found 2 rule groups for normal class and 1 rule group for tumor class. After this, we found the most significant rule group among these 3 rule groups. A total of 27 genes were found to be covered by the most significant rule group.

Similarly, we have applied EIRGG on the lung expression data. We found 2 rule groups; among them 1 rule group were for normal class and 1 for tumor class. A total of 24 genes were found to be covered by the most significant rule group. Here we considered 0.013 as the threshold value, as for this threshold value we got maximum difference in the number of genes for two consecutive threshold values, and the number of genes selected is moderately large. Note that we started with 0.001 with an interval of 0.001 as the threshold values.

Similarly, for breast, leukemia, sarcoma, colon expression data, we found 4, 5, 3 and 2 rule groups by applying LFRGG, respectively. Finally, a total of 28, 32, 31, 21 were found to be covered by most significant rule groups for breast, leukemia, sarcoma, colon expression datasets, respectively.

Likewise, applying EIRGG on breast, leukemia, sarcoma, colon expression data, we found 3, 2, 4 and 3 rule groups, respectively. Finally, a total of 24, 34, 32, 22 were found to be covered by most significant rule groups for breast, leukemia, sarcoma, colon expression datasets, respectively.

3.2. Comparative analysis of LFRGG and EIRGG with other existing methods

3.2.1. Based on GO attributes

In our study, the enrichment of each GO category [18] for each of the genes has been calculated by its P-value. A low P-value indicates that the genes belonging to the enriched functional categories are biologically significant. Here only functional categories with P-value $< 5.0 \times 10^{-5}$ were considered. We have made comparative study, with other methods, viz., SAM [13], SNR [37], NA [15], BR [32,8] and DA [27] in terms of their ability to identify functionally enriched genes. Table 1 shows the number of functionally enriched attributes corresponding to these methods for different sets of genes. It is found that for all the datasets, LFRGG and EIRGG performed the best. These results show that the proposed methodology has been able to select more important genes responsible for mediating a particular type of adenocarcinoma than the other methods considered here. The comparative analysis between the proposed methods (LFRGG) and EIRGG in identifying responsible genes has also been demonstrated. LFRGG has found more true positives than those obtained by EIRGG. It is difficult to mention the figures corresponding to false positives, and true and false negatives. The reason behind the difficulty is that gene expression datasets consist of a large number of genes.

3.2.2. Based on level of significance

We have applied t-test on the genes identified by LFRGG, EIRGG, SAM, SNR, NA, BR, DA on each dataset. We have considered top 20

Table 1
Comparative results on number of enriched attributes of various sets of genes.

Dataset	Gene set	Number of enriched attributes						
		LFRGG	EIRGG	SAM	SNR	NA	BR	DA
Lung	First 5	63	51	14	21	26	27	24
	First 10	87	85	13	9	15	21	27
	First 15	84	82	30	14	16	16	30
	First 20	86	86	28	13	15	16	32
Colon	First 5	61	66	27	34	28	28	30
	First 10	63	74	30	39	30	31	35
	First 15	80	79	30	35	30	31	38
	First 20	78	66	33	44	33	33	41
Sarcoma	First 5	73	63	41	49	67	66	42
	First 10	95	84	50	54	67	53	49
	First 15	75	79	61	40	31	34	51
	First 20	73	68	65	43	41	34	55
Lymphocytes	First 5	63	51	59	19	36	52	41
	First 10	87	85	60	26	48	59	47
	First 15	84	82	70	37	55	70	55
	First 20	86	86	51	13	52	73	58
Breast	First 5	39	28	5	6	16	15	12
	First 10	47	38	3	12	18	13	14
	First 15	45	44	6	9	7	17	14
	First 20	38	42	15	12	10	21	16

genes obtained by each method for each dataset. Here we perform comparison based on the number of genes (out of top 20) that matches with different levels of significance. Higher this number, better is the performance of the method in selecting genes. Tables 3–7 (in Supplementary material) show that LFRGG and EIRGG perform the best as the number of matched genes (shown in bracket in the third column) are the highest over the others.

For human lung expression data (Table 3 in Supplementary material), on applying LFRGG and EIRGG, we have identified some important genes like CALCA (4.02), PFKP (5.78), TYMS (3.98), IGFBP3 (6.98), IARS (5.98), HBB (7.08), HLA-B (5.42), SFTPA2 (6.89), and TNF (4.23). The number in the bracket indicates t-value corresponding to the gene. The t-value of these set of genes exceeds the value for $P = 0.001$. It indicates that these set of genes are highly significant (99.9% level of significance). Similarly, genes like IGHG3 (2.67), PRKACA (2.89), MEN1 (3.15) and IGHM (3.25) exceeds the t-value for $P = 0.01$. This means that these genes are significant at the level of 99%. Likewise, RPLP1 (2.14), LARS (2.22), SMCIL1 (2.07), MGP (2.31), RNASE1 (2.43), SFTPC (2.37), and HLA-DRA (2.27) genes are important at the level of 95% significance. We have performed t-test for the genes identified by some other gene selection algorithms like SAM, SNR, NA, BR and DA. But highly significant (99.9% significance level) genes like PFKP, TYMS, IARS, and HLA-B are not included in the set of first twenty selected genes by these methods. This result suggests that LFRGG and EIRGG are able to find more true positive genes than the existing methods.

Similarly, for human breast expression data (Table 4 in Supplementary material), applying our methods we have identified some important genes like NARS (5.63), BAT1 (6.05), GDI2 (5.89), KRAS (6.88), HAL (6.10), FNTA (5.94), BCAN (5.76). The t-values of this set of genes indicate that these genes are highly significant (99.9% level of significance). Likewise, the t-values of genes FMO2 (3.73), COX4I1 (3.89), SDHC (3.82), TTN (3.69), TRIO (3.60), H3F3A (3.79), IGHG1 (3.98) indicate that their significance levels are 99%. The t-values of LARS (2.14), FADS2 (2.26), NPM1 (2.18), PTPRA (2.20), CANX (2.08) indicate 95% level of significance.

We have applied LFRGG and EIRGG on human colon expression data. As a result (Table 5 in Supplementary material), we have found some genes like microtubule-associated protein-2 (4.26), thymidylate syntase (5.34), phosphofructokinase, platelet type (4.21), Calcitonin (5.04), major histocompatibility complex enhancer-binding protein (4.43), isoleucyl-tRNA synthetase (3.78), hemoglobin beta chain (5.87), insulin-like growth factor-binding protein-6 (6.22), tumor necrosis factor (4.12), whose t-values indicate that these genes are statistically highly significant (level of significance 99.9%). Genes like flavin-containing monooxygenase form II (2.41), colon carcinoma kinase-4 (2.78), methylthioadenosine phosphorylase (2.65) are found to be significant at a level of 99%. The t-values of the genes like pepsinogen C (1.89), cytochrome P450 4F2 (1.68), platelet-derived growth factor receptor alpha (2.07), vasoactive intestinal peptide (1.72) indicate 95% level of significance.

For human lymphocytes and plasma cell expression data (Table 6 in Supplementary material), we have identified some highly significant (99.9% level of significance) genes like BAX (4.15), PFKP (3.98), TYMS (5.67), NARS (4.89), BAT1 (5.08), BCR (4.98), HBB (6.52), HAL (3.89), IGFBP3 (4.79), CALCA (5.34), HLA-B (6.22), IARS (4.79), BRCA1 (5.37). It was also reported that genes like GDI2 (2.71), FNTA (2.56), SDHC (2.78), KRAS (2.67), IGF1 (2.90), H3F3A (2.39) have a significance level of 99%. In our results, the t-values of the genes like LARS (1.98), ATP6V0B (2.02), CDKN2A (2.11) indicate 95% of significance level.

Likewise, applying both the methods on human sarcoma expression data (Table 7 in Supplementary material), we found some highly significant (99.9% level of significance) genes like BAX (3.98), CALCA (4.07), PFKP (3.87), TYMS (4.89), IARS (5.46), NARS (3.77), BRCA1 (3.92), BCR (4.32), HBB (5.76), IGFBP3 (6.23), HLA-B (4.23). It was also reported that genes like FMO2 (2.62), PTEN (2.98), IGF1 (3.02), CDKN2A (2.79), MEN1 (2.56), IGHM (2.40) have a significance level at 99% and genes like VEGFA (1.77), AGER (1.89), LARS (2.12), FHL1 (2.02) have a significance level at 95%.

3.3. Validation of the results obtained by both LFRGG and EIRGG

Here we provide an account for validating the results obtained by LFRGG and EIRGG. On applying the two methods on the five datasets, we have found some genes that are common in some of the datasets. These genes are either over or under expressed in tumor samples than in normal ones. In each case, we have found the same nature of growth and decay in terms of expression values of these genes. Moreover, we have made a broader search through internet to validate our results with the existing ones. It has been found that some of these genes were already found to be responsible for cancer.

3.3.1. Using existing literature

Applying LFRGG and EIRGG to human lung expression data, we have found some important genes, like CALCA [2,23,39], TYMS [21,33], IGFBP3 [9,19], HLA-B [26,34], HBB [25], TNF [14,6,36], IGHG3 [31], SFTPA1 [17,35], and SFTPA2 [35] that have a quite significant number of enriched attributes (Table 1). Genes like PFKP and IARS have a quite a significant number of enriched attributes, but there is no information in literature to our knowledge about these genes. This result suggests that the aforesaid genes may have impact on lung adenocarcinoma.

On applying other existing methods like SAM, SNR, NA, BR and DA on this dataset, we have found a set of important genes (CALCA, IGFBP3, HBB, SFTPA2) that are also present in the results of LFRGG and EIRGG. Thus we may conclude that genes like CALCA, IGFBP3, HBB, SFTPA2 have a quite a significant role to the development of lung adenocarcinoma. It is interesting to note that the proposed LFRGG and EIRGG have been able to find more responsible genes (for mediating lung adenocarcinoma) supported by wide range of earlier investigations than those of other methods. Thus the methodology developed in this article is able to select biologically more sig-

nificant genes than the others. Similar findings (Tables 3–7 in Supplementary material) have been obtained for the other datasets too.

3.3.2. Using expression profile plots

Here we consider some genes that are among the most significant genes of our results. The expression values of these genes have changed significantly from normal samples to diseased samples. Applying LFRGG and EIRGG on human lung expression data, we report that genes like IGFBP3, PFKP, IARS, TYMS, among the top 10 most important genes, are over expressed in tumor samples (Fig. 2 in Supplementary material). On the other hand, the expression value of gene HBB has reduced quite significantly in tumor samples (Fig. 2 in Supplementary material). This gene is identified as under expressed gene.

In the case of human colon expression profile, the genes like calcitonin (CALCA), colon carcinoma kinase-4 (CCK4), isoleucyl-tRNA synthetase (IARS), thymidylate syntase (TYMS), Hemoglobin Beta Chain (HBB), tumor necrosis factor receptor (TNF), insulin-like growth factor-binding protein-6 (IGFBP6) have changed their expression values from normal colon samples to tumor ones. Among these genes, CALCA, CCK4, IARS, TYMS are identified as up regulated genes (Fig. 3 in Supplementary material). On the other hand, HBB, TNF, IGFBP6 are the down regulated (Fig. 4 in Supplementary material).

For human breast cell expression profile, we clearly see that genes like BCAN, GDI2, NARS (among the ten most important genes obtained by LFRGG and EIRGG) change their expression levels quite significantly from normal breast mammary epithelial cell samples to breast cancer ones. The expression value of the gene BCAN increases in breast cancer cell lines whereas the expression value of genes GDI2 and NARS decrease drastically in breast cancer cell lines. The expression profile plot of these genes are depicted in (Fig. 5 in Supplementary material).

Similarly, for human soft tissue sarcoma expression data, genes like BRCA1, TYMS, IARS, HBB have changed their expression values from normal tissue to sarcoma tissue (Fig. 6 in Supplementary material). The expression value of gene HBB drastically decreases in diseased sarcoma samples, whereas that of BRCA1, TYMS, IARS increase in diseased samples.

For human lymphocytes and plasma cell expression data, we report that expression values of the genes like BAX, CALCA, ATP6V0B, NARS have changed significantly from normal B lymphocytes and plasma cells to macroglobulinemia, chronic lymphocytic leukemia, multiple myeloma samples. Fig. 7 (in Supplementary material) clearly indicates that genes like BAX, NARS, ATP6V0B are over expressed in diseased samples, whereas the gene CALCA is under expressed in diseased samples.

4. Conclusions

In this article, we have developed two methods, called Linguistic Fuzzy Rule Generation and Grouping (LFRGG), and Expression Interval based Rule Generation and Grouping (EIRGG), that have demonstrated how fuzzy sets and linguistic variables can be used to select a few possible genes responsible for mediating a specific disease. Note that use of linguistic variables makes it possible to develop the system capable of reasoning like human. Note that, incorporation of fuzzy set theory makes the system capable of handling exact/inexact forms of input data. A small set of possible genes have been identified that have moved from one class of normal to another class of diseased samples. An existing rule generation and grouping algorithm [10] has finally been used to find a set of possible genes responsible for cancer. A comparative analysis of the performance of the methods with some others, viz., SAM [13], SNR [37], NA [15], BR [32,8] and DA [27] has been provided.

Applying the above methods on five cancer datasets, we have found the genes whose over/under expression may cause a particular type of cancer. The results are appropriately validated by some statistical parameters, earlier investigations and gene expression profile plots. It has been demonstrated that the numbers of enriched attributes corresponding to the sets of genes obtained by LFRGG and EIRGG are much higher than those obtained by the aforesaid existing methods. The results of both LFRGG and EIRGG are statistically more significant than those of the others. LFRGG has been able to find more true positives than EIRGG in identifying responsible genes. This results in important genes that may have role in mediating development of a particular disease. The results are verified using some existing results and expression profile plots. It has been found that the methods provided here are able to detect more important genes in mediating a disease than the other existing ones considered here. Note that most of these genes did not pass through the interval ratio filter of EIRGG [10]. Hence they have not been detected by EIRGG. These results facilitate the researchers carrying out biochemical experiments to do further analysis on these genes instead of on the entire genome.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2009.06.003.

References

[1] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999;96:6745–50.
[2] Amatschek S, Koenig U, Auer H, Steinlein P, Pacher M, Gruenfelder A, et al. Tissue-wide expression profiling using cDNA subtraction and microarrays to identify tumor-specific genes. Cancer Res 2004;64:844–56.
[3] Azuaje F. A computational neural approach to support the discovery of gene function and classes of cancer. IEEE Trans Biomed Eng 2001;48:332–9.
[4] Beer DG, Kardia SLR, Huang C-C, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–23.
[5] Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001;98:13790–5.
[6] Bjorling-Poulsen M, Seitz G, Guerra B, Issinger OG. The pro-apoptotic fas-associated factor 1 is specifically reduced in human gastric carcinomas. Int J Oncol 2003;23:1015–23.
[7] Bolstad BM, Irizarry RA, Strand MA, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185–93.
[8] Cawley GC, Talbot NLC. Gene selection in cancer classification using sparse logistic regression with bayesian regularization. Bioinformatics 2006;22:2348–55.
[9] Chang YS, Wang L, Liu D, Mao L, Hong WK, Khuri FR, et al. Correlation between insulin-like growth factor-binding protein-3 promoter methylation and prognosis of patients with stage I non-small cell lung cancer. Clin Cancer Res 2002;8:3669–75.
[10] Cong G, Tan K, Tung A, Xu X. Mining top-*k* covering rule groups for gene expression data. SIGMOD 2005:670–81.
[11] Detwiller KY, Fernando NT, Segal NH, Ryeom SW, DAmore PA, Yoon SS. Analysis of hypoxia-related gene expression in sarcomas and effect of hypoxia on ma interference of vascular endothelial cell growth factor A. Cancer Res 2005;65:5881–9.
[12] Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci USA 2001;98:13784–9.
[13] Goh L, Song Q, Kasabov N. A novel feature selection method to improve classification of gene expression data. In: Asia–Pacific bioinformatics conference; 2004.
[14] Golovko O, Nazarova N, Tuohimaa P. A20 gene expression is regulated by TNF, vitamin D and androgen in prostate cancer cells. J Steroid Biochem Mol Biol 2005;94:197–202.
[15] Golub TR, Slonim TK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
[16] Gutierrez NC, Ocio EM, Rivas GD, Maiso P, Delgado M, Ferminan E, et al. Gene expression profiling of B lymphocytes and plasma cells from Waldenstroms macroglobulinemia: comparison with expression patterns of the same cell

counterparts from chronic lymphocytic leukemia, multiple myeloma and normal individuals. Leukemia 2007;21:541–9.
[17] Jiang F, Caraway N, Nebiyou B, Zhang HZ, Khanna A, Wang H, et al. Surfactant protein a gene deletion and prognostics for patients with stage I non-small cell lung cancer. Clin Cancer Res 2005;11:5417–24.
[18] Kim DW, Lee KH, Lee D. Detecting clusters of different geometrical shapes in microarray gene expression data. Bioinformatics 2005;21:1927–34.
[19] Lee HY, Chun KH, Liu B, Wiehle SA, Cristiano RJ, Hong WK, et al. Insulin-like growth factor binding protein-3 inhibits the growth of non-small cell lung cancer. Cancer Res 2002;62:3530–7.
[20] Liotta L, Petricion E. Molecular profiling of human cancer. Nat Rev Genet 2000;1:48–56.
[21] Liu H, Jin G, Wang H, Wu W, Liu Y, Qian J, et al. Association of polymorphisms in one-carbon metabolizing genes and lung cancer risk: a case-control study in chinese population. Lung Cancer 2008;61:21–9.
[22] Machado L, Vinterbo S, Weber G. Classification of gene expression data using fuzzy logic. J Intel Fuzzy Syst 2002;12:19–24.
[23] Marchevsky AM, Tsou J, Laird-Offringa I. Classification of individual lung cancer cell lines based on dna methylation markers: use of linear discriminant analysis and artificial neural networks. J Mol Diagn 2004;6:28–36.
[24] Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, et al. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. Nucleic Acids Res 2004;32.
[25] Morere JF. Role of epoetin in the management of anaemia in patients with lung cancer. Lung Cancer 2004;46:149–56.
[26] Mottironi VD, Banks SM, Pollara B, Rudofsky UH. Hla and survival in lung cancer. Clin Immunol Immunopathol 1987;45:55–62.
[27] Mukherjee S, Roberts SJ, Van der Laan MJ. Data-adaptive test statistics for microarray data. Bioinformatics 2005;21(Suppl. 2):ii108–14.
[28] Pal SK, Dutta Majumder D. Fuzzy mathematical approach to pattern recognition. New York: Wiley (Halsted); 1986.
[29] Pal SK, Mandal DP. Linguistic recognition system based on approximate reasoning. Inform Sci 1992;61:135–61.
[30] Ram R, Chetty M, Dix TI. Fuzzy model for gene regulatory network. In: Proceedings IEEE congress on evolutionary computation Vancouver, BC, Canada; 2006. p. 1450–5.
[31] Remmelink M, Mijatovic T, Gustin A, Mathieu A, Rombaut K, Kiss R, et al. Identification by means of cDNA microarray analyses of gene expression modifications in squamous non-small cell lung cancers as compared to normal bronchial epithelial tissue. Int J Oncol 2005;26:247–58.
[32] Shevade SK, Keerthi SS. A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics 2003;19:2246–53.
[33] Shi Q, Zhang Z, Neumann AS, Li G, Spitz MR, Wei Q. Case-control analysis of thymidylate synthase polymorphisms and risk of lung cancer. Carcinogenesis 2005;26:649–56.
[34] So T, Takenoyama M, Mizukami M, Ichiki Y, Sugaya M, Hanagiri T, et al. Haplotype loss of HLA class I antigen as an escape mechanism from immune attack in lung cancer. Cancer Res 2005;65:5945–52.
[35] Stoffers M, Goldmann T, Branscheid D, Galle J, Vollmer E. Transcriptional activity of surfactant-apoproteins A1 and A2 in non small cell lung carcinomas and tumor-free lung tissues. Pneumologie 2004;58:395–9.
[36] Tang X, Wu W, Sun SY, Wistuba II, Hong WK, Mao L. Hypermethylation of the death-associated protein kinase promoter attenuates the sensitivity to trail-induced apoptosis in human non-small cell lung cancer cells. Mol Cancer Res 2004;2:685–91.
[37] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98.
[38] Vinterbo SA, Kim EY, Machado L. Small, fuzzy and interpretable gene expression based classifiers. Bioinformatics 2005;21:1964–70.
[39] Virmani AK, Tsou JA, Siegmund KD, Shen LY, Long TI, Laird PW, et al. Hierarchical clustering of lung cancer cell lines using DNA methylation markers. Cancer Epidemiol Biomarkers Prev 2002;11:291–7.
[40] Woolf PJ, Wang Y. A fuzzy logic approach to analyzing gene expression data. Physiol Genomics 2000;3:9–15.
[41] Zadeh LA. The concept of linguistic variable and its applications to approximate reasoning-II. Inform Sci 1975;8:301–57.
[42] Zadeh LA. Precisiated natural language — toward a radical enlargement of the role of natural languages in information processing, decision and control. Proceedings of the 9th international conference on neural information processing (ICONIP'02), vol. 1; 2002. p. 1–3.
[43] Zhang H, Yu CY, Singer B, Xiong M. Recursive partitioning for tumor classification with gene expression microarray data. Proc Natl Acad Sci USA 2001;98:6730–5.