Information Fusion 10 (2009) 260–269

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Linguistic recognition system for identification of some possible genes mediating the development of lung adenocarcinoma

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

[a] Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
[b] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India

ARTICLE INFO

Article history:
Received 29 November 2006
Received in revised form 5 September 2007
Accepted 12 November 2008
Available online 7 December 2008

Keywords:
Fuzzy sets
Low
Medium
High
Microarray
Gene expression
p-value

ABSTRACT

In the present article, we develop a linguistic recognition system for identification of some possible genes mediating the development of human lung adenocarcinoma. 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 obtained by a rule generation/grouping technique. The system has been successfully applied on two microarray gene expression data sets. The results are appropriately validated by some earlier investigations, gene expression profiles and *t*-test. The proposed methodology has been able to find more true positives than an existing one in identifying responsible genes. Moreover, we have found some new genes that may have role in mediating the development of lung adenocarcinoma.

1. Introduction

Lung cancer continues to be the most common cause of cancer related to mortality in men and women [17]. The treatments of lung cancer are primarily based on the broad classification of tumors into small cell, non-small cell type and histological subtyping. The heterogeneity of lung cancer patients at each disease stage with respect to outcome and treatment response suggests that additional subclassification and substaging remains possible. Adenocarcinoma is currently the predominant histological subtype of non-small cell lung cancer (NSCLC) [10,19,36]. Pre-operative variables that affect survival of patients with NSCLC are identified. Tumor size, vascular invasion, poor differentiation, high-tumor proliferative index and several genetic alterations, *k*-ras mutations [35,39], p53 mutations [13,16], have prognostic significance.

Recent studies [11,3,27] involving gene expression profiles, obtained by microarray technology, have a profound impact on cancer research. In some examples [11,3], 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 lung 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 adenocarcinogenic patients [27].

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 [55,56]. 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. [57], based on knowledge extraction of gene by classification; transformation of gene expression by fuzzy heuristic rule set [49]; classifying fuzzy inference system [28]; development of a fuzzy model for gene regulatory networks [34]; measuring performance of small rule-based classifiers using fuzzy logic [45]; identification of normal and tumor patients using fuzzy neural network model [1].

In microarray gene expression data, genes have expression values that are in some intervals under different conditions. There exists methodology [6] based on these intervals for finding genes responsible for a particular disease. Although each interval has a 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 a fuzzy set theoretic method for identifying genes responsible for a particular disease.

In this article, we consider two oligonucleotide microarrays; one of them contains gene expression profiles of 86 lung

* Corresponding author. Tel.: +91 33 25753105; fax: +91 33 25783357.
E-mail addresses: rajat@isical.ac.in (R.K. De), anupam.ghosh@rediffmail.com (A. Ghosh).

1566-2535/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2008.11.007

adenocarcinoma including 67 stage I and 19 stage III tumor samples, and 10 normal samples on 7129 genes [2]. The other data includes 139 tumor samples and 17 normal samples on 12,600 genes [3]. Here we develop a linguistic recognition system to identify a set of possible genes mediating the development of lung adenocarcinoma. The methodology involves the task of dimensionality reduction, 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 [8]. Linguistic fuzzy sets, e.g., *low*, *medium* and *high* 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 [55,56]. 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 [6] is used in the final step for finding a set of possible genes mediating the development of lung adenocarcinoma. The method in [6] does not incorporate the notion of fuzzy sets and it is used for comparative analysis. As already discussed, there exist many articles in literature for microarray gene expression analysis. Here we have considered the reference [6] as it deals with association rule mining and provides an algorithm for rule grouping that is effective in the area of gene expression analysis.

The article presents a set of possible genes obtained by both these methods, which may be responsible for lung adenocarcinoma. The results are appropriately validated by some earlier investigation, gene expression profiles and *t*-test. Moreover, the proposed methodology has found more true positives than an existing one in identifying responsible genes. We have also identified some genes like CACNB3, CEACAM4, CLTH, CREB3, D123, DDXL, EIF5, GGCX, HMGCL, HUMPPA, KIAA0057, POLR2B, PRKAG1, SMARCD1, SMCIL1 whose expression values have increased in tumor samples from the normal ones. These genes are newly found, in the present article, to be possible mediators of lung adenocarcinoma. Since we are dealing with a purely computational work, researchers involved in wet-lab based work may consider these genes for further analysis.

Section 2 describes the proposed methodology (Method 1) and another existing method (Method 2) [6] for comparative analysis. Section 3 provides the results along with their validation. The article concludes in Section 4.

2. Methodologies

In this paper, we have proposed a methodology for developing a linguistic recognition system on the gene expression data set to find out some possible genes mediating the development of lung adenocarcinoma. 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 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. We call this methodology as Method 1 throughout the paper.

2.1. Method 1

The proposed methodology, called Method 1, has a few steps. Each of these steps are described below.

2.1.1. Step I – dimensionality reduction

The need of normalization arises naturally when we deal with experiments involving multiple arrays. There may be two broad characterizations one could use for the type of variation in different arrays: interesting variation and obscuring variation. Interesting variation deals with the biological differences, for example [14], when large differences in the expression level of particular genes between a diseased and a normal source are observed. On the other hand, obscuring variation is introduced during the process of carrying out experiment with different samples of either normal or diseased type. The purpose of normalization is to deal with this obscuring variation.

Here we apply an algorithm similar to a normalization algorithm (Cyclic loess [8]) to the data set for normal lung samples as well as for tumor samples. 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 probe intensities x_{ki} and x_{kj}, $k = 1, \ldots, p$ being the probe 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. Note that the idea of using normalization algorithm in dimensionality reduction is novel.

Here we fit a parabolic curve. The fits based on the normalization curve are $\hat{M}_k$ and thus the normalization adjustment is given by $(M_k - \hat{M}_k)$. This adjustment is apportioned equally to x_{ki} and x_{kj}. To deal with more than two arrays, the method is extended to look at all distinct pair wise combinations. The algorithm is carried out in a pair wise manner, recording an adjustment for each of the two arrays in each pair. After looking at all the pairs of the arrays, we have a set of adjustments by which the expression levels of the arrays are updated. Then we repeat the process until the difference in the expression values among the arrays becomes less than some predefined threshold. Typically only 5 or 6 complete iterations through all pair wise combinations are needed to achieve useful results. After getting the normalized values of the genes, we have taken mean of these normalized values of each gene to represent a gene by a single expression value. The algorithm [8] is provided below for convenience of the readers.

Algorithm.

For each gene, do

[1:] choose pair wise samples,
[2:] compute M_k and A_k for each pair,
[3:] fit M_k with respect to A_k. Here we are going for parabolic curve fitting, i.e., for each pair of genes $\hat{M}_k = a + bA_k + cA_k^2$. So for a set of A_k values that we have defined in the previous step, we get a set of estimated $\hat{M}_k$ values. Finally we get $\binom{m}{2}$ number of $(M_k - \hat{M}_k)$ values, for m samples. We call these $(M_k - \hat{M}_k)$ values as adjustment for these m samples,
[4:] record these adjustments for each sample and compute the resultant adjustment for each sample,
[5:] update the old log expression value for each sample using the following formula
new $\log_2 x_{ki}$ = old $\log_2 x_{ki}$ + resultant adjustment
[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).

If the log expression value of sample 1 is less than sample 2, we divide the adjustment (say a) such that sample 1 gets $+a/2$ and sample 2 gets $-a/2$. In this way each sample has nine values after distribution (Table 1). Finally, we calculate the resultant adjustment by taking algebraic sum of these values for each sample. Now we update the old log expression value of a gene for each sample by adding resultant adjustment of the corresponding sample. This completes one iteration.

2.1.2. Step II – formulation of linguistic fuzzy sets and their matching

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* [32,33] to replace numerical expression values. 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 fuzzy sets *low*, *medium* and *high*. Note that incorporation of linguistic fuzzy sets provides a tool for natural computing [55,56] as the resulting system is capable of reasoning like human. 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). However, one may consider four, five or even more linguistic fuzzy sets to partition the domain into four, five or more.

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 generated and is given by

$$v_j = [U_{low}(x_j), U_{medium}(x_j), U_{high}(x_j)]^T.$$

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

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

Let us now describe the formulation of membership functions corresponding to the fuzzy sets *low*, *medium* and *high*. The membership function $U_{low}(x_j)$ is defined as

$$\begin{aligned} U_{low}(x_j) &= 1 && \text{if } x_j \leqslant x_{min}, \\ &= 1 + (x_j - x_{min})/(x_{min} - C_{med}) && \text{if } C_{low} \leqslant x_j < C_{med}, \\ &= 0 && \text{otherwise}. \end{aligned} \quad (1)$$

Similarly, $U_{med}(x_j)$ and $U_{high}(x_j)$ are defined as

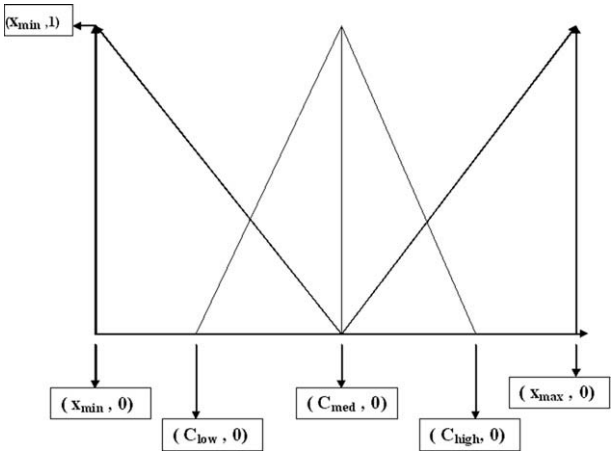

Fig. 1. Triangular membership function.

$$\begin{aligned} U_{med}(x_j) &= (x_{min} - x_j)/(x_{min} - C_{med}) && \text{if } C_{low} \leqslant x_j < C_{med}, \\ &= (x_{max} - x_j)/(x_{max} - C_{med}) && \text{if } C_{med} < x_j \leqslant C_{high}, \\ &= 0 && \text{otherwise} \end{aligned} \quad (2)$$

and

$$\begin{aligned} U_{high}(x_j) &= 1 && \text{if } x_j > C_{high}, \\ &= 1 + (x_j - x_{max})/(x_{max} - C_{med}) && \text{if } C_{med} < x_j < x_{max}, \\ &= 0 && \text{otherwise}. \end{aligned} \quad (3)$$

Here x_{max} and x_{min} denote the upper and lower bounds of the observed range of expression values x_j. The parameters are computed as follows:

$$\begin{aligned} L_{medium} &= 1/2(x_{max} - x_{min}) \\ C_{med} &= x_{min} + L_{medium} \\ L_{low} &= (C_{med} - x_{min}) \\ C_{low} &= (C_{med} - x_{min})/2 + x_{min} \\ L_{high} &= (x_{max} - C_{med}) \\ C_{high} &= (x_{max} - C_{med})/2 + C_{med} \end{aligned}$$

The basic nature of these membership functions is as follows: (i) maximum (minimum) value of each function is 1 (0). (ii) The membership functions corresponding to *low* and *medium*, cut at the point for which $U_{low} = U_{med} = 0.5$. Similar is the case of U_{med} and U_{high} such that at the point of intersections of the membership functions corresponding to *medium* and *high*, $U_{med} = U_{high} = 0.5$. (iii) The membership value of a gene to a fuzzy set is maximum at the center of the fuzzy set and decreases as it is away from the center.

It may be noted that one may use other membership functions (e.g, pi-type membership function [32]) for modeling the fuzzy sets

Table 1
Computation of resultant adjustments.

Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
$+a12/2$	$-a12/2$	$-a23/2$	$+a24/2$	$-a25/2$	$+a26/2$	$-a27/2$	$+a28/2$	$-a29/2$	$+a210/2$
$-a13/2$	$+a23/2$	$+a13/2$	$-a34/2$	$+a45/2$	$-a56/2$	$-a37/2$	$-a58/2$	$+a39/2$	$-a510/2$
$-a14/2$	$-a24/2$	$+a34/2$	$+a14/2$	$+a35/2$	$+a36/2$	$+a47/2$	$+a38/2$	$+a49/2$	$-a610/2$
$+a15/2$	$+a25/2$	$-a35/2$	$-a45/2$	$-a15/2$	$-a46/2$	$+a57/2$	$-a48/2$	$+a59/2$	$+a710/2$
$-a16/2$	$-a26/2$	$-a36/2$	$+a46/2$	$+a56/2$	$+a16/2$	$+a67/2$	$-a68/2$	$+a69/2$	$-a810/2$
$-a17/2$	$+a27/2$	$+a37/2$	$-a47/2$	$-a57/2$	$-a67/2$	$+a17/2$	$+a78/2$	$-a79/2$	$+a910/2$
$-a18/2$	$-a28/2$	$-a38/2$	$+a48/2$	$+a58/2$	$+a68/2$	$-a78/2$	$+a18/2$	$+a89/2$	$-a410/2$
$+a19/2$	$+a29/2$	$-a39/2$	$-a49/2$	$-a59/2$	$-a69/2$	$+a79/2$	$-a89/2$	$-a19/2$	$-a310/2$
$+a110/2$	$-a210/2$	$+a310/2$	$+a410/2$	$+a510/2$	$+a610/2$	$-a710/2$	$+a810/2$	$-a910/2$	$-a110/2$
$Ad1$	$Ad2$	$-Ad3$	$Ad4$	$-Ad5$	$Ad6$	$-Ad7$	$-Ad8$	$-Ad9$	$Ad10$

low, *medium* and *high*, keeping the basic nature of the membership functions identical. Here in trying to express input x_j with the linguistic properties, we are effectively dividing the dynamic range of expression values into three overlapping partitions called *low*, *medium* and *high* corresponding to each gene. However, one may partition the same into four, five or even more to get more number of linguistic variables.

The choices for the L's and C's in Eqs. (1)–(3) automatically ensures that one of the membership values $U_{low}(x_j)$, $U_{med}(x_j)$ and $U_{high}(x_j)$ of each gene in the corresponding three dimensional linguistic space should be greater than 0.5, and among the other two one should be zero. This allows a gene to have a strong membership to at least one of the properties *low*, *medium*, *high*.

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 (i.e., L's and C's), for both normal and tumor samples are computed based on the normal lung samples only.

Thus we get the three classes of genes *low*, *medium* and *high* for normal samples and tumor samples separately. 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 classes *medium* and *high* of tumor. These genes are identified as the over expressed genes. Similarly, we perform matching of class *medium* of normal with the classes of *low* and *high* of tumor. Lastly, we perform matching of class *high* of normal with classes *low* and *medium* of tumor. Thus we have identified the genes that have changed their classes from normal to tumor samples.

2.1.3. Step III – rule generation and grouping

Here we describe the technique in [6] 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. After selecting the genes in Step 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 $\mathbf{R}(\hat{\mathbf{I}})$ which is the largest set of genes that contain $\hat{\mathbf{I}}$ (a subset of $\mathbf{I}$). Likewise, we define the gene support set, denoted by $\mathbf{I}(\hat{\mathbf{R}})$ as the largest set of items common among the genes in $\hat{\mathbf{R}}$ (a subset of $\mathbf{R}$). After generating the intervals of expression values of these genes, 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.

The problem of association rule mining is to 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 are unable to cover all the items exhaustively and 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 [6].

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 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 the values of *CONF* and *SUP*, respectively, of an ith rule group. After this we select the genes that are covered by the most significant rule group. For details of this rule grouping algorithm, one may consult [6].

2.2. Method 2

Here we describe an existing methodology (Method 2) [6] which has been used for comparative analysis. Method 2 includes the tasks like interval generation, rule generation using these intervals and grouping of these rules using support and confidence factors [6] defined above. Let us consider a gene expression data set 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 data set 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 diseased, then $k = 2$.

We use the row ID set $\mathbf{R}$ to represent the set of rows or genes, and the item ID set $\mathbf{I}$ to represent the set of items or intervals. As a mapping between rows and items, the item support set $\mathbf{R}(\hat{\mathbf{I}})$ which is the largest set of rows that contain $\hat{I}$. Likewise, we define row support set $\mathbf{I}(\hat{\mathbf{R}})$. They are already defined in Step III in Section 2.1. As before, rule from the data set takes the form of $\mathbf{A} \rightarrow C$. We now describe the steps involved in Method 2 [6] in details.

2.2.1. Step I – generation of intervals

Let the data set consists of n number of genes (number of rows) and d number of samples (number of columns). Also assume that the number of classes is two, i.e., $k = 2$. In other words, there are

Table 2
SUP and *CONF* values of the rules obtained by Method 1 on Data Set 2.

Rule no.	Rule description	No. of genes (normal)	No. of genes (tumor)	Confidence	Support
1	$(\mathbf{-7}, \mathbf{3061}) \rightarrow N$	30	22	100%	20.8%
2	$(\mathbf{268}, \mathbf{5317}) \rightarrow N$	32	29	100%	22.2%
3	$(\mathbf{1993}, \mathbf{7004}) \rightarrow N$	23	29	100%	15.9%
4	$(\mathbf{401}, \mathbf{4515}) \rightarrow N$	28	24	100%	19.4%
5	$(\mathbf{750}, \mathbf{9142}) \rightarrow T$	40	53	100%	36.8%

Table 3
SUP and *CONF* values of the rules obtained by Method 2 on Data Set 2.

Rule no.	Rule description	No. of genes (normal)	No. of genes (tumor)	Confidence	Support
1	$(\mathbf{25}, \mathbf{5317}) \rightarrow N$	28	32	100%	34.1%
2	$(\mathbf{481}, \mathbf{8773}) \rightarrow T$	30	36	100%	44.4%

normal samples (i.e., samples in class C_1) and tumor samples (i.e., samples in class C_2). Among these d samples let d_1 be the number of normal samples and d_2 be that of tumor samples. Now, we divide this data set in two different subsets (i.e., normal data set and tumor data set). We can represent each gene by d_1 number of values in normal data set and d_2 number of values in tumor data set. We assume that the minimum gene expression value of gene g_i in normal data set is x_{imin} and the maximum gene expression value of gene g_i in normal data set is x_{imax}. Similarly, for tumor data set, 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} or not. If so then we conclude that the interval of gene g_i has an overlapping region. So the interval may be broken into $y_{imin} - x_{imin}$, $y_{imax} - x_{imax}$ (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 $(y_{imax} - x_{imax})$.

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. Step II – selection of genes

After generating intervals, we select the genes for which the expression values have changed significantly from normal to tumor samples. 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* ($0 \leqslant ratio \leqslant 1$) that 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 overlapping region is created $x_{imax} - y_{imin}$ and the total length of expression value of gene g_i is $y_{imax} - x_{imin}$. So the *ratio* becomes $(x_{imax} - y_{imin})/(y_{imax} - x_{imin})$. Now, 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. Step III – rule generation and grouping

After selecting the genes and the intervals described by Step II, we generate the rules based on these intervals. This rule generation and grouping step is identical to Step III of Method 1 (described in Section 2.1.3). For further description of Method 2, one may refer to [6].

2.3. Description of the data sets

We now describe the data sets that we have considered in our analysis.

2.3.1. Data Set 1

Data Set 1 is obtained by microarray experiments of Affymetrix Corporation and contains data for Ann Arbor tumors and normal lung samples [2]. In this data set, there are 7129 genes (more specifically, Affymetrix probe-sets) for 86 lung tumor and 10 normal lung samples. The gene expression profiles represent 86 primary lung adenocarcinomas, including 67 stage I and 19 stage III tumors, as well as 10 neoplastic lung samples. More details on this data set are found in [2].

2.3.2. Data Set 2

Data Set 2 also contains data for both tumor and normal lung samples [3], and is obtained by microarray experiments of Affymetrix Corporation. It consists of 12,600 genes with 203 lung samples. Among them, 186 lung tumors and 17 normal lung specimens were used. Out of them, 125 adenocarcinoma samples were associated with clinical data. 203 specimens include 127 histologically-defined lung adenocarcinoma specimens, 21 squamous cell lung carcinoma specimens, 20 pulmonary carcinoids, six SCLC specimens and 17 normal lung specimens. Other 12 adenocarcinoma specimens were suspected, based on clinical history, to be extra pulmonary metastases. Here we consider 139 adenocarcinoma lung samples (i.e., 127 + 12 = 139) and 17 normal lung samples for our experiment.

3. Results

In this section, the effectiveness of the proposed method (Method 1) is demonstrated on two lung adenocarcinoma gene expression data sets (Data Set 1 and Data Set 2). A comparative analysis with Method 2 [6] is also included.

3.1. Analysis of the results

3.1.1. Using Method 1 and Method 2 on Data Set 1

Data Set 1 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. That is, 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.

Table 4
List of 69 genes obtained by Method 1 on Data Set 1. Underlined genes are also obtained by Method 2 for this data set. (The figures within bracket after each gene are the corresponding *t*-values.)

Over expressed (*t*-value)	KIAA0320 (4.21), CREB3 (3.81), D123 (3.57), KIAA0057 (3.28), SEC14L1 (7.73), AIB3 (3.53), NP220 (3.82), CEACAM4 (2.58), ARRB2 (2.80), CYP2F1 (0.78), SRF (5.48), HUMPPA (3.80), HMGCL (4.06), HRIHFB2206 (3.17), GGCX (7.41), PDK2 (2.92), YES1 (2.59), ADH1(7.00), RPLP0(5.35), HUMMLC2B (3.72), MYL6 (7.93), IGFBP3 (8.05), TNFRSF1A (4.48), TNFAIP3 (3.38), TNFAIP1 (3.77), COX6A2 (2.12), TFAP4 (3.67), SMC1L1 (7.06), ABR (3.49), CACNB3 (5.11), MLH1 (4.22), MAPK9 (3.82), ATRX (3.58), TNFSF6 (3.72), KCNMB1 (2.95), RGS3 (3.33), DUT (3.44), TNFSF10 (2.62), PRKAG1 (5.84), KISS1 (2.41), CLTH (2.67), PIN1 (3.34), EIF5 (3.95), DGKZ (3.18), CUL3 (3.51), TP53BP2 (3.29), MICB (4.10), SMARCD1 (2.55), DDXL (6.20), MEN1 (8.32), PNMT (2.15), PTPN6 (3.87), POLR2B (3.24), CDK5 (2.42), G9A (6.61), DDR2 (3.51), SORT1 (6.81), RBMX (3.22)
Under expressed (*t*-value)	TC10 (2.72), SFTPA2 (9.16), MLLT2 (7.27), ITGA8 (6.41), SFTPA1 (7.00), HBB (6.64), UGB (3.91), PRKACA (6.13), FMO2 (8.83), HBA2 (7.03)

This is how we reduced the entire data set through dimensionality reduction (Step I in Section 2.1). 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 in 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 255 genes that changed their class from normal to tumor. This is actually Step II of Method 1 (Section 2.1).

Finally we used the rule generation followed by grouping technique (Step III of Method 1 described in Section 2.1) on these 255 genes. In order to do this, we first generate the intervals based on the entire data set. It results in 1548 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 in C_1 (normal). Similarly, rules with consequent part representing tumor samples are obtained. We found 42 rules for normal class and 24 rules for tumor class. After this, we used the rule grouping algorithm on these two sets of rules separately. We finally found 23 rule groups for normal class and 3 rule groups for tumor class. Thus a total of 26 rule groups were obtained. After this, we found the most significant rule group among these 26 rule groups. A total of 69 genes were found to be covered by the most significant rule group. They are the possible genes responsible for the development of lung adenocarcinoma and are listed in Table 4.

Among these 69 genes, 59 genes were found to be over expressed and 10 genes under expressed in the tumor samples. Similarly, we have applied Method 2 on the Data Set 1. We found 13 rule groups; among them 11 rule groups were for normal class and 2 for tumor class. The most significant rule group among these 13 rule groups was found. A total of 24 genes were found to be covered by the most significant rule group. It has been found that these 24 genes (underlined in Table 4) are already included in the set of 69 genes as shown in Table 4. Among these 24 genes, 16 were over expressed and 8 were under expressed in the tumor samples. Here we considered 0.013 as the threshold value, since 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.

3.1.2. Using Method 1 and Method 2 on Data Set 2

In Data Set 2, there are 17 normal samples, i.e., $\binom{17}{2}$ pairs containing expression values of 12,600 genes. After calculating the first adjustments based on the parabolic curve fitting, as in the case of Data Set 1, we noted the required adjustment. In this way, we applied Method 1 on Data Set 2. Finally, we found 72 genes that change their classes from normal to tumor. After this we applied the rule generation and rule grouping method on these 72 genes. In order to do this, 1210 intervals were generated. In this way we found 5 rule groups (4 rule groups for normal class and 1 rule group for tumor class). Finally, we found the most significant rule group among these 5 rule groups. A total of 53 genes were found to be covered by the most significant rule group. They are some possible genes responsible for the development of lung adenocarcinoma and are listed in Table 5.

Likewise, we have applied Method 2 on Data Set 2. We have found 2 rule groups; among them one for normal class and the other for tumor class. Finally, a total of 36 genes were found to be covered by the most significant rule group. These 36 genes (underlined) are already included in the set of 53 genes obtained by Method 1 (Table 5). It is to be noted here that unlike Data Set 1, here we used Affymetrix probe-set ID to represent a gene as no gene identifier is there in data set. Among these 53 genes obtained by Method 1 on Data Set 2, 48 were found to be over expressed in the tumor samples. Similarly, 32 genes (underlined) were found to be over expressed and 4 under expressed using Method 2. Due to the same reason for Data Set 1, here we have considered the value of the threshold as 0.05. ♣

We have computed *CONF* and *SUP* values of the rules generated by both Method 1 and Method 2 on both the data sets. In order to restrict the size of the article, the results are presented for Data Set 2 only (Tables 2 and 3). The antecedent part of a rule contains an interval and the consequent part is the type (normal or tumor) of data. The intervals in the antecedent parts are union of some other intervals. Applying Method 1 on 72 genes (obtained by Steps I and II of Method 1) of Data Set 2, we have 5 rules in which rule 5 became the most significant rule (mentioned in Table 2). For rule 5, the number of genes whose interval of expression values over 139 tumor samples being contained in the interval of the antecedent part is 53. Therefore, SUP for this rule is $53/(2 \times 72) = 36.8\%$, the factor 2 in the denominator appears by considering both 72 genes in normal samples as well as that in tumor samples. *CONF* of the rule is 53/53 = 100%, as the number of genes whose interval of expression values over both normal and tumor samples being contained in the interval of the antecedent part is also 53. Applying Method 2 on 41 genes (selected by the interval-ratio filter) of Data Set 2 results in 2 rules in which rule 2 is found to be the most significant (mentioned in Table 3).

3.1.3. *t*-Test

In order to validate the results statistically, we have performed *t*-test on these data sets. The *t*-values obtained for the genes are provided within brackets after corresponding genes in Tables 4 and 5. The *t*-values show that most of these genes are highly significant (p-value < 0.001) in mediating lung adenocarcinoma. For the

Table 5
List of 53 genes obtained by Method 1 on Data Set 2. Underlined genes are also obtained by Method 2 for this data set. (The figures within bracket after each gene are the corresponding *t*-values.) Here the Affymetrix probe IDs have been used due to the unavailability of gene symbols.

Over expressed (*t*-value)	31309_r_at (4.23), 31330_at (3.10), 31331_at (3.79), 31617_at (4.51), 31638_at (3.08), 31956_f_at (4.04),
	32390_at (3.42), 33636_at (6.34), 34637_f_at (8.31), 35537_at (3.11), 34054_at (3.61), 37057_s_at (4.23), 37821_at (4.61), 37599_at (3.48), 37956_at (5.31), 38610_s_at (4.49), 31315_at (4.12), 31477_at (5.31), 32457_f_at (4.86), 32458_f_at (4.99), 33991_g_at (3.56), 37782_at (4.26), 40031_at (3.79), 41096_at (7.90), 34702_f_at (4.48), 34703_f_at (3.54), 35726_at (3.96), 36024_at (4.58), 36883_at (4.40), 37897_s_at (8.11), 39726_at (4.25), 32805_at (7.91), 32821_at (4.97), 33377_at (3.75), 34319_at (10.04), 36105_at (3.79), 36156_at (8.13), 37399_at (3.53), 38469_at (3.61), 38783_at (4.22), 40544_g_at (3.82), 40899_at (3.22), 2027_at (3.58), 1802_s_at (3.98), 1664_at (6.48), 1586_at (3.65), 1371_s_at (4.25), 608_at (5.10)
Under expressed (*t*-value)	31525_s_at (7.56), 38691_s_at (17.10), 31687_f_at (5.89), 39220_at (6.61), 33383_f_at (11.29)

genes CEACAM4, ARRB2, PDK2, YES1, KCNMB1, TNFSF10, KISS1, CLTH, SMARCD1, CDK5 and TC10, p-value < 0.01, and for genes COX6A2 and PNMT, 0.01 < p-value < 0.02. It is to be noted that the p-values mentioned here are unadjusted p-values.

3.2. Validation of the results obtained by both Method 1 and Method 2

On applying the two methods to these data sets, we have found some genes that are common in both the data sets. These genes are either over or under expressed in tumor samples than its 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 some existing ones. It has been found that some of these genes were already found to be responsible for lung adenocarcinoma. For example, genes IGFBP3 [5,23], SFTPA1 [18,41], SFTPA2 [41], TNFAIP3 [12], TNFSF6 [4], TNFSF10 [43], UGB [54,24,22], HBB [31] were already found by some earlier investigations. We now describe, in brief, a few genes that are found to be responsible by the two methods. Some earlier investigations also support this fact.

3.2.1. Tumor necrosis factor (TNF)

In our experiment we report that the gene TNF is highly over expressed in tumor samples. In each of the sample, the expression value of this gene is significantly changed from normal samples to tumor samples in both the data sets. This result confirms that this gene may be responsible for the development of lung adenocarcinoma. Tumor necrosis factor alpha is a member of the TNF/TNFR cytokine super family, which includes TNFRSF1A, TNFAIP3, TNFAIP1, TNFSF6, TNFSF10. It was found by some earlier investigations that this factor is responsible for NSCLC [25,20,52,38,37].

3.2.2. Surfactant protein (SFTP)

Likewise, we report that surfactant protein A1 and A2 are derived from two under expressed genes in tumor samples. Their expression values have changed significantly from normal to tumor samples in both the data sets. Fig. 2 supports this fact. This result along with some earlier investigations [9,51,26,41] ensures that these genes may be responsible to develop tumor cells in lung.

3.2.3. Insulin-like growth factor binding protein (IGFBP)

It has also been found that insulin-like growth factor binding protein-3 (IGFBP3) have changed its expression level significantly in tumor samples (Fig. 3). This gene is identified as over expressed gene in both the data sets. This fact is also supported by some earlier investigations [23,46,30,44,21,15,58].

3.2.4. Hemoglobin beta (HBB)

In our experiment we report that genes of the category hemoglobin are significantly under expressed from normal to tumor

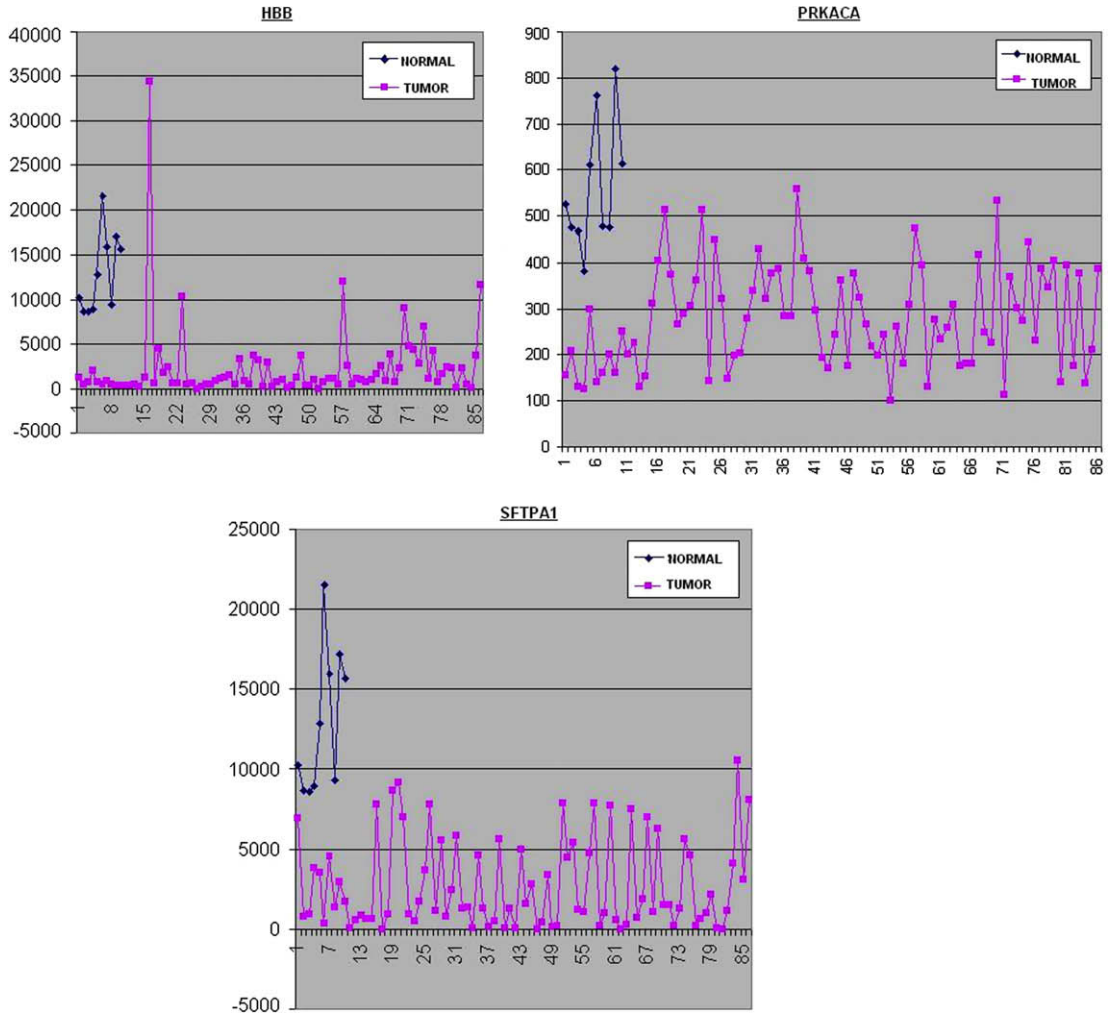

Fig. 2. Expression profiles of some under-expressed genes (HBB, PRKACA, SFTPA1) in normal (shown blue points) and tumor (shown by red points) samples (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

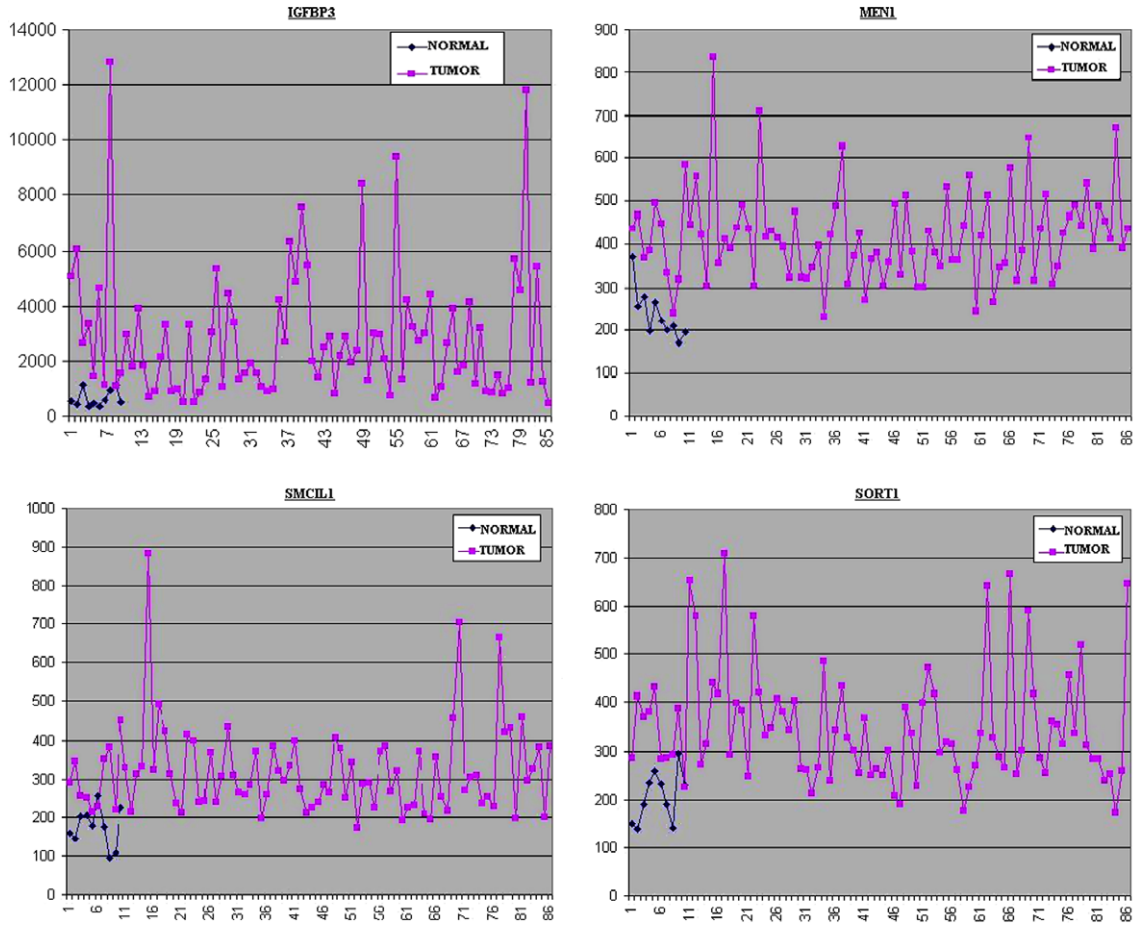

Fig. 3. Expression profiles of some over-expressed genes (IGFBP3, MEN1, SMCIL1, SORT1) in normal (shown blue points) and tumor (shown by red points) samples (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

samples in terms of expression values (Fig. 2). For example, we report that HBB is one of such types of genes. This is purely under expressed gene in tumor samples. An earlier investigation [31] supports this fact.

3.3. Comparative analysis of Method 1 and Method 2

The comparative analysis between the proposed methodology (Method 1) and Method 2 in identifying responsible genes has been demonstrated by experimental results. Method 1 has found more true positives (36 for Data Set 1 and 31 for Data Set 2) than those (10 for Data Set 1 and 19 for Data Set 2) obtained by Method 2. It is difficult to mention the figures corresponding to false positives, and true and false negatives. The reason behind the difficulty is that both Data Set 1 and Data Set 2 consist of a large number of genes. Information on all these genes is not available in literature. Moreover, Data Set 2 contains Affymetrix probe set id instead of gene id. This has created the problem of finding the concerned genes even more difficult.

We now provide a brief account of the genes which have been found responsible in the development of lung adenocarcinoma, and are obtained by Method 1 only. These observations are supported by some earlier investigations and also by gene expression profiles (Fig. 3).

3.3.1. Cullin 3 (CUL3)

In our results we have found that the gene CUL3 is over expressed in tumor samples. This gene is a candidate tumor suppressor located on human chromosome 8p21, a region commonly deleted in cancer [48]. It can be shown that a RhoBTB2 missense mutant identified in a lung cancer cell line is neither able to bind CUL3 nor is it regulated by the ubiquitin/proteasome system, resulting in increased RhoBTB2 protein levels in vivo [48]. Based on these reports we can say that this gene may be responsible for lung adenocarcinoma.

3.3.2. Cyclin-dependent kinase 5 (CDK5)

CDK5 has also been found to be an over expressed gene in tumor samples. Expression pattern of gene CDK5 is matched in analysis of cDNA microarray and proteomics. These powerful approaches using cDNA microarray and proteomics could provide in-depth information on the impact of HPV-16 E6-related genes and proteins. A549E6 human lung adenocarcinoma cell lines, stably expressing the HPV 16-E6 gene, were compared with those of RKO and A549 cell lines to generate a differential protein expression catalog [53]. Based on this result, we may infer that this gene has a major contribution for developing lung tumors.

3.3.3. DUT

DUT is an over expressed gene in tumor samples. Paclitaxel lacked a radiosensitizing effect on DUT cells. The results should be considered while designing clinical trials that use paclitaxel as a potential radiosensitizer of certain human carcinomas [42]. This may confirm that this gene mediates the development of lung adenocarcinoma.

3.3.4. Euchromatic histone–lysine N-methyltransferase 2 (G9A)

Similarly, the gene G9A is an over expressed gene in tumor samples. Small interfering RNA (siRNA)-mediated knockdown of G9A led to increased MASPIN expression in MDA-MB-231 cells, to levels that were supra-additive, verifying the importance of these enzymes in maintaining multiple layers of epigenetic repression in tumor cells. These results highlight an additional and complimentary mechanism of action for 5-aza-CdR in the reactivation of epigenetically silenced genes, in a manner that is independent of its effects on DNA methylation. This further supports an important role for H3 K9 methylation in the aberrant repression of tumor suppressor genes in human cancer [50]. This result validates the fact that the gene G9A may have role in the development of lung adenocarcinoma.

3.3.5. Multiple endocrine neoplasia I (MEN1)

In our results, we have found that the gene MEN1 is highly over expressed in tumor samples. Lung carcinoids occur sporadically and rarely in association with multiple endocrine neoplasia type 1 (MEN1) [7]. This observation along with Fig. 3 supports our results.

3.3.6. RNA binding motif protein, X-linked (RBMX)

RBMX is over expressed in tumor samples. RBM proteins might constitute a novel family of apoptosis modulators. The expression of both RBM10 variants was significantly associated with the expression of the VEGF gene [29]. So based on this observation, we may say that this gene is responsible for lung cancer.

3.3.7. Sortilin 1 (SORT1)

This is also an over expressed gene in tumor samples. Increased numbers of endothelial cells are observed in peripheral blood of cancer patients. These circulating endothelial cells (CECs) may contribute to the formation of blood vessels in the tumor or reflect vascular damage caused by treatment or tumor growth [40]. Thus this observation and Fig. 3 support our results.

3.3.8. Yamaguchi sarcoma viral oncogene homolog 1 (YES1)

The expression values of this gene have changed significantly from normal to tumor samples. This is identified as an over expressed gene. The expression pattern of the gene YES1 was similar to tumor suppressor genes [47].

3.4. Newly identified genes by Method 1

On applying the proposed methodology (Method 1), we have found some genes like CACNB3, CEACAM4, CLTH, CREB3, D123, DDXL, EIF5, GGCX, HMGCL, HUMPPA, KIAA0057, POLR2B, PRKAG1, SMARCD1, SMCIL1 whose expression values have increased in tumor samples from the normal ones. This is evident from Fig. 3 corresponding to a few of these genes. (In order to keep the size of the article small, we have not included the other figures.) Similarly, we have found two genes TC10 and PRKACA whose expression values have significantly decreased in tumor samples. However, there is no information in literature to our knowledge about these genes. Thus these genes are newly identified in this article as possible mediators of lung adenocarcinoma. Since we are dealing with a purely computational work, researchers involved in wet-lab based work may consider these genes for further analysis. Note that most of the genes like CEACAM4, CLTH, CREB3, D123, EIF5, HMGCL, HUMPPA, KIAA0057, POLR2B, SMARCD1, TC10 did not pass through interval-ratio filter of Method 2 [6]; thereby they have not been detected by Method 2.

4. Conclusions

In this article, we have developed a linguistic recognition system that has demonstrated how linguistic variables can be used to select a few possible genes responsible for a specific disease. Note that use of linguistic variables makes it possible to develop the system capable of reasoning like human. Here, first of all, variation among the expression values for genes over different samples has been removed through an algorithm similar to cyclic loess normalization algorithm [8]. This step has reduced the dimension of each gene by which a gene has been represented by a single derived expression value. We have then applied the concept of fuzzy sets to classify the genes into three fuzzy classes, viz., *low*, *medium*, and *high*. That is, genes in both normal and tumor samples have been grouped into these three classes separately. 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 tumor samples. An existing rule generation and grouping algorithm [6] has finally been used to find a set of possible genes responsible for lung adenocarcinoma. A comparative analysis of the performance of the system with an existing one [6] has been provided.

Applying the above methodology on two lung adenocarcinoma data sets, we have found the genes that have changed significantly from normal class to tumor class. The results are appropriately validated by earlier investigations and gene expression profiles. Method 1 has been able to find more true positives than Method 2 in identifying responsible genes. Finally, we have found some new genes like CACNB3, CEACAM4, CLTH, CREB3, D123, DDXL, EIF5, GGCX, HMGCL, HUMPPA, KIAA0057, POLR2B, PRKAG1, SMARCD1, SMCIL1 whose expression values have increased in tumor samples from the normal ones, as possible mediators of the development of lung adenocarcinoma. Note that most of these genes did not pass through the interval-ratio filter of Method 2 [6]. Hence they have not been detected by Method 2. These results facilitate the researchers carrying out wet-lab experiments to do further analysis on these genes instead of on the entire genome.

References

[1] F. Azuaje, A computational neural approach to support the discovery of gene function and classes of cancer, IEEE Trans. Biomed. Eng. 48 (2001) 332–339.
[2] G.D. Beer, S.L.R. Kardia, C.C. Huang3, T.J. Giordano4, A.M. Levin2, D.E. Misek5, L. LIN, G. Chen1, T.G. Gharib1, D.G. Thomas4, M.L. Lizyness4, R. Kuick5, S. Hayasaka3, J.M.G. Taylor3, M.D. Iannettoni1, M.B. Orringer1, S. Hanash5, Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nature Med. 8 (2002) 816–823.
[3] A. Bhattacharjee et al., Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc. Natl. Acad. Sci. USA 98 (2001) 13790–13795.
[4] M. Bjorling-Poulsen, G. Seitz, B. Guerra, O.G. Issinger, The pro-apoptotic fas-associated factor 1 is specifically reduced in human gastric carcinomas, Int. J. Oncol. 23 (2003) 1015–1023.
[5] Y. Chang, L. Wang, D. Liu, L. Mao, W. Hong, F. Khuri, H. Lee, 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. 8 (2002) 3669–3675.
[6] G. Cong, K. Tan, A. Tung, X. Xu, Mining top-k covering rule groups for gene expression data, SIGMOD (2005) 670–681.
[7] L.V. Debelenko, E. Brambilla, S.K. Agarwal, J.I. Swalwell, M.B. Kester, I.A. Lubensky, Z. Zhuang, S.C. Guru, P. Manickam, S.E. Olufemi, S.C. Chandrasekharappa, J.S. Crabtree, Y.S. Kim, C. Heppner, A.L. Burns, A.M. Spiegel, S.J. Marx, L.A. Liotta, F.S. Collins, W.D. Travis, M.R. Emmert-Buck, Identification of MEN1 gene mutations in sporadic carcinoid tumors of the lung, Hum. Mol. Genet. 6 (1997) 2285–2290.
[8] S. Dudoit, Y.H. Yang, M.J. Callow, T.P. Speed, Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments, Stat. Sinica 12 (2002) 111–139.
[9] A.A. Ewis, K. Kondo, F. Dang, Y. Nakahori, Y. Shinohara, M. Ishikawa, Y. Baba, Surfactant protein B gene variations and susceptibility to lung cancer in chromate workers, Am. J. Ind. Med. 49 (2006) 367–373.

[10] W.A. Fry, J.L. Phlillips, H.R. Menck, Ten-year survey of lung cancer treatments and survival in hospitals in united states, Cancer 86 (1999) 1867–1876.
[11] M.E. Garber et al., Diversity of gene expression in adenocarcinoma of the lung, Proc. Natl. Acad. Sci. USA 98 (2001) 13784–13789.
[12] O. Golovko, N. Nazarova, P. Tuohimaa, A20 gene expression is regulated by TNF vitamin D and androgen in prostate cancer cells, J. Steroid Biochem. Mol. Biol. 94 (2005) 197–202.
[13] D.H. Harpole, J.E. Herndon, W.G. Wolfe, J.D. Iglehart, J.R. Marks, A prognostic model of recurrence and death in stage I non-small cell lung cancer utilizing presentation histopathology and oncoprotein expression, Cancer Res. 55 (1995) 51–56.
[14] A. Hartemink, D. Gifford, T. Jaakkola, R. Young, Maximum likelihood estimation of optical scaling factors for expression array normalization, SPIE Bios (2001).
[15] R. Hochscheid, G. Jaques, B. Wegmann, Transfection of human insulin-like growth factor-binding protein 3 gene inhibits cell growth and tumorigenicity: a cell culture model for lung cancer, J. Endocrinol. 166 (2000) 553–563.
[16] Y. Horio, T. Takahashi2, T. Kuroishi, K. Hibi, M. Suyama, T. Niimi, K. Shimokata, K. Yamakawa, Y. Nakamura, R. Ueda, T. Takahashi, Prognostic significance of p53 mutations and 3p deletions in primary resected non-small cell lung cancer, Cancer Res. 53 (1993) 1–4.
[17] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, C. Smigal, M.J. Thun, Cancer statistics, CA Cancer J. Clin. 56 (2006) 106–130.
[18] F. Jiang, N. Caraway, B. Nebiyou, H.Z. Zhang, A. Khanna, H. Wang, R. Li, R.L. Fernandez, T.M. Zaidi, D.A. Johnston, R.L. Katz, Surfactant protein a gene deletion and prognostics for patients with stage I non-small cell lung cancer, Clin. Cancer Res. 11 (2005) 5417–5424.
[19] M.C. Kaisermann et al., Evolving features of lung adenocarcinoma Rio de Janeiro, Braz. Oncol. Rep. 8 (2001) 189–192.
[20] O. Kayacan, D. Karnak, S. Beder, E. Gullu, H. Tutkak, F.C. Senler, D. Koksal, Impact of TNF-alpha and IL-6 levels on development of cachexia in newly diagnosed NSCLC patients, Am. J. Clin. Oncol. 29 (2006) 328–335.
[21] Y. Kodama, R.C. Baxter, J.L. Martin, Insulin-like growth factor-I inhibits cell growth in the a549 non-small lung cancer cell line, Am. J. Resp. Cell Mol. Biol. 27 (2002) 336–344.
[22] G.C. Kundu, Z. Zhang, G. Mantile-Selvaggi, A. Mandal, C.J. Yuan, A.B. Mukherjee, Uteroglobin binding proteins: regulation of cellular motility and invasion in normal and cancer cells, Ann. NY Acad. Sci. 923 (2000) 234–248.
[23] H.Y. Lee, K.H. Chun, B. Liu, S.A. Wiehle, R.J. Cristiano, W.K. Hong, P. Cohen, J.M. Kurie, Insulin-like growth factor binding protein-3 inhibits the growth of non-small cell lung cancer, Cancer Res. 62 (2002) 3530–3537.
[24] J.C. Lee, K.H. Park, S.J. Han, C.G. Yoo, C.T. Lee, S.K. Han, Y.S. Shim, Y.W. Kim, Inhibitory effect of adenovirus-uteroglobin transduction on the growth of lung cancer cell lines, Cancer Gene Ther. 10 (2003) 287–293.
[25] F.J. Lejeune, C.F. Ruegg, Recombinant human tumor necrosis factor: an efficient agent for cancer treatment, Bull. Cancer 93 (August) (2006) 90–100.
[26] H.M. Lin, C. Seifart, U. Seifart, A. Plagens, S. DiAngelo, P. von Wichert, J. Floros, Rare SP-A alleles and the SP-A1-6A(4) allele associate with risk for lung carcinoma, Clin. Genet. 68 (2005) 128–136.
[27] L. Liotta, E. Petricion, Molecular profiling of human cancer, Nature Rev. Genet. 1 (2000) 48–56.
[28] L. Machado, S. Vinterbo, G. Weber, Classification of gene expression data using fuzzy logic, J. Intell. Fuzzy Syst. 12 (2002) 19–24.
[29] F. Martinez-Arribas, D. Agudo, M. Pollan, F. Gomez-Esquer, G. Diaz-Gil, R. Lucas, J. Schneider, Positive correlation between the expression of X-chromosome RBM genes (RBMX, RBM3, RBM10) and the proapoptotic Bax gene in human breast cancer, J. Cell Biochem. 97 (2006) 1275–1282.
[30] J.W. Moon, Y.S. Chang, C.W. Ahn, K.N. Yoo, J.H. Shin, J.H. Kong, Y.S. Kim, J. Chang, S.K. Kim, H.J. Kim, S.K. Kim, Promoter-202 A/C polymorphism of insulin-like growth factor binding protein-3 gene and non-small cell lung cancer risk, Int. J. Cancer 118 (2006) 353–356.
[31] J.F. Morere, Role of epoetin in the management of anaemia in patients with lung cancer, Lung Cancer 46 (2004) 149–156.
[32] S.K. Pal, D. Dutta Majumder, Fuzzy Mathematical Approach to Pattern Recognition, John Wiley (Halsted Press), New York, 1986.
[33] S.K. Pal, D.P. Mandal, Linguistic recognition system based on approximate reasoning, Inform. Sci. 61 (1992) 135–161.
[34] R. Ram, M. Chetty, T.I. Dix, Fuzzy model for gene regulatory network, IEEE Congress Evolution. Comput. (2006) 1450–1455.
[35] S. Rodenhuis et al., Mutational activation of *k*-ras oncogene: a possible pathogenic in adenocarcinoma of lung, New Engl. J. Med. 317 (1987) 929–935.
[36] V.L. Roggli, R.T. Vollmer, S.D. Greenberg, M.H. McGavran, H.J. Spjut, R. Yesner, Lung cancer heterogenity: a blinded and randomized study of 100 consecutive cases, Hum. Pathol. 16 (1985) 569–579.
[37] C. Seifart, A. Plagens, A. Dempfle, U. Clostermann, C. Vogelmeier, P. von Wichert, U. Seifart, TNF-alpha, TNF-beta, IL-6 and IL-10 polymorphisms in patients with lung cancer, Dis. Markers 21 (2005) 157–165.
[38] C.M. Shih, Y.L. Lee, H.L. Chiou, W. Chen, G.C. Chang, M.C. Chou, L.Y. Lin, Association of TNF-alpha polymorphism with susceptibility to and severity of non-small cell lung cancer, Lung Cancer 52 (2006) 15–20.
[39] R.J.C. Slebos, R.E. Kibbelaar, O. Dalesio, A. Kooistra, J. Stam, C.J. Meijer, S.S. Wagenaar, R.G. Vanderschueren, N.V. Zandwijk, W.J. Mooi, K-ras oncogene activation as a prognostic marker in adenocarcinoma of lung, New Engl. J. Med. 323 (1990) 561–565.
[40] D.A. Smirnov, B.W. Foulk, G.V. Doyle, M.C. Connelly, L.W. Terstappen, S.M. O'Hara, Global gene expression profiling of circulating endothelial cells in patients with metastatic carcinomas, Cancer Res. 66 (2006) 2918–2922.
[41] M. Stoffers, T. Goldmann, D. Branscheid, J. Galle, E. Vollmer, Transcriptional activity of surfactant-apoproteins A1 and A2 in non small cell lung carcinomas and tumor-free lung tissues, Pneumologie 58 (2004) 395–399.
[42] J.S. Stromberg, Y.J. Lee, E.P. Armour, A.A. Martinez, P.M. Corry, Lack of radiosensitization after paclitaxel treatment of three human carcinoma cell lines, Cancer 75 (1995) 2262–2268.
[43] X. Tang, W. Wu, S.Y. Sun, I.I. Wistuba, W.K. Hong, L. Mao, 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. 2 (2004) 685–691.
[44] E. Unsal, D. Koksal, A.S. Yurdakul, S. Atikcan, P. Cinaz, Analysis of insulin like growth factor 1 and insulin like growth factor binding protein 3 levels in bronchoalveolar lavage fluid and serum of patients with lung cancer, Resp. Med. 99 (2005) 559–565.
[45] S.A. Vinterbo, E.Y. Kim, L. Machado, Small fuzzy and interpretable gene expression based classifiers, Bioinformatics 21 (2005) 1964–1970.
[46] H. Wang, Y.X. Wan, Q.K. Zhang, Significance and expression of insulin-like growth factor 1 and IGF binding protein 3 in serum of patients with lung cancer, Ai Zheng. 23 (2004) 710–714.
[47] L. Wang, J.S. Zhu, M.Q. Song, G.Q. Chen, J.L. Chen, Comparison of gene expression profiles between primary tumor and metastatic lesions in gastric cancer patients using laser microdissection and cdna microarray, World J. Gastroenterol. 12 (2006) 6949–6954.
[48] A. Wilkins, Q. Ping, C.L. Carpenter, RhoBTB2 is a substrate of the mammalian CUL3 ubiquitin ligase complex, Genes Dev. 18 (2004) 856–861.
[49] P.J. Woolf, Y. Wang, A fuzzy logic approach to analyzing gene expression data, Physiol. Genomics 3 (2000) 9–15.
[50] R.J. Wozniak, W.T. Klimecki, S.S. Lau, Y. Feinstein, B.W. Futscher, 5-Aza-2-deoxycytidine-mediated reductions in G9A histone methyltransferase and histone H3 K9 di-methylation levels are linked to tumor suppressor gene reactivation, Oncogene (2006).
[51] O. Yamamoto, H. Takahashi, M. Hirasawa, H. Chiba, M. Shiratori, Y. Kuroki, S. Abe, Surfactant protein gene expressions for detection of lung carcinoma cells in peripheral blood, Resp. Med. 99 (2005) 1164–1174.
[52] F. Yang, P. Shi, X. Xi, S. Yi, H. Li, Q. Sun, M. Sun, Recombinant adenoviruses expressing trail demonstrate antitumor effects on non-small cell lung cancer (NSCLC), Med. Oncol. 23 (2006) 191–204.
[53] E.K. Yim, J. Meoyng, S.E. Namakoong, S.J. Um, J.S. Park, Genomic and proteomic expression patterns in HPV-16 E6 gene transfected stable human carcinoma cell lines, DNA Cell Biol. 23 (2004) 826–835.
[54] J.M. Yoon, J.J. Lim, C.G. Yoo, C.T. Lee, Y.J. Bang, S.K. Han, Y.S. Shim, Y.W. Kim, Adenovirus-uteroglobin suppresses COX-2 expression via inhibition of NF-kappaB activity in lung cancer cells, Lung Cancer 48 (2005) 201–209.
[55] L.A. Zadeh, The concept of linguistic variable and its applications to approximate reasoning-II, Inform. Sci. 8 (1975) 301–357.
[56] L.A. Zadeh, Precisiated natural language – toward a radical enlargement of the role of natural languages in information processing, decision and control, in: Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP'02), vol. 1, 2002, pp. 1–3.
[57] H. Zhang, C.-Y. Yu, B. Singer, M. Xiong, Recursive partitioning for tumor classification with gene expression microarray data, Proc. Natl. Acad. Sci. USA 98 (2001) 6730–6735.
[58] Y. Zhao, M. El-Gabry, T.K. Hei, Loss of Betaig-h3 protein is frequent in primary lung carcinoma and related to tumorigenic phenotype in lung cancer cells, Mol. Carcinogen. 45 (2006) 84–92.