Generation of Linguistic Rules on the Genes Mediating the Development of Lung Adenocarcinoma

Rajat K. De[1] and Anupam Ghosh[2]

[1] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India.
rajat@isical.ac.in
[2] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata 700152, India.
anupam.ghosh@rediffmail.com

Abstract. In the present article, we propose a method for generation of rules on genes mediating the development of human lung adenocarcinoma. The method involves the application of cyclic loess normalization technique followed by the incorporation of the fuzzy sets *low*, *medium* and *high*. Linguistic rules are generated on the gene expression values. The system has been successfully applied on a microarray gene expression data consisting of expression values of 7129 genes in 10 normal and 86 tumor samples. In our results we have found that nine genes, including *RPLP0*, *ADH1*, *UGB*, *FMO2*, *HBA2*, *SFTPA1*, *SFTPA2*, *HBB* are the most significant mediating the development of lung adenocarcinoma. The results are in accordance with a number of earlier investigations.

1 Introduction

Lung cancer continues to be the most common cause of cancer related mortality in men and women. The treatments of lung cancer are primarily based on the broad classification of tumors into small cell, non-small cell types 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.

Recent studies [12, 6, 18] involving gene expression profiling of clinical specimens have had a profound impact on cancer research. In some examples [12, 6], correlations have been made between the expression levels of a gene or set of genes and clinically relevant subclassifications of specific tumor subtypes. These results have shown that true molecular classification and substaging of multiple tumor types may be possible, leading to taking effective measures in prognosis and patient management. Microarray Technology can be used to correlate the gene-expression patterns with numerous clinical parameters including patient outcome to better predict tumor behavior in individual patients [18]. Analysis of lung cancers using array technologies has identified subgroups of tumors that differ according to tumor types and histological subclasses, and to lesser extent, survival among adenocarcinoma patients.

Advances in Computer Science and Engineering.
Research in Computing Science 23, 2006, pp. 87-98

Received 17/07/06
Accepted 04/10/06
Final version 13/10/06

Fuzzy set theory is capable of handling uncertainty in the gene expression values arising due to incompleteness, imprecision, noise and experimental errors. The notion of fuzzy sets has been used in the domain of gene expression analysis. These include identifying interacting genes that fit a known "fuzzy" model of gene interaction by testing all combinations of gene expression determining profiles [29], a list of differentially expressed genes [7], a knowledge-based model of Zhang et. al. [31] to determine the influence of genes on classification of a sample into a tumor category, transforming gene expression values into qualitative descriptors using a set of linguistic rules involving fuzzy logic [29], fuzzy inferencing for classification of tumor samples [20], application of fuzzy ARTMAP to identify normal and tumor patients [4]. Recently, Sokhansanj et. al. [27] have demonstrated an approach with exhaustive search for fuzzy gene interaction models that best fit transcription measurements by microarray technology.

In this article, we have applied linguistic fuzzy sets on gene expression data of lung adenocarcinoma to identify a set of genes mediating the development of lung adenocarcinoma. The method involves a normalization method of cyclic loess [11] to reduce the variation among the expression levels of the gene over different samples. Then we have represented the whole-normalized data set in form fuzzy linguistic variables [23]. In this way, we have found three different classes corresponding to *low*, *medium* and *high* on gene expression values, for normal and tumor samples separately. In the next step, we have performed the matching operation with normal to tumor samples, which has led to identify 293 genes that have changed significantly from normal samples to tumor samples. Finally, rules are generated using the technique involving confidence factor for these genes. We report here the nine significant rules corresponding to nine genes among these 293 genes that have changed their expression values most significantly from normal samples to tumor samples. The gene expression data we have considered here are oligonucleotide arrays containing gene expression profiles for 10 normal and 86 lung adenocarcinoma including 67 stage I and 19 stage III tumor samples on 7129 genes.

2 Related Work

Zhang has proposed a rule discovery procedure that is based on prior knowledge of the influence of each gene for classification of a sample into a tumor category [31]. Only (gene, expression) descriptors that are consistent with the classification are considered as the antecedent of the rule. For example, if some parameter is positive for gene A, then only (A, $HIGH$) will be retained. On the other hand, a negative value of the parameter indicates that only (A, LOW) would be retained. This is certainly arbitrary and depends on the coding for the class variable. The magnitude of the parameter is ignored in this rule generation process. All possible descriptor sets are considered to compose the rules in the initial step. For example, if parameter for genes A and B are both positive in the logistic regression model, the corresponding rules are (A, $HIGH$), (B, $HIGH$), (A, $HIGH$), and (B, $HIGH$). Every rule thus constructed is considered for

elimination using the impact of its removal in the number of misclassifications. If there is no change or the number of misclassifications goes down, then the rule is eliminated.

A set of heuristic rules in the fuzzy logic framework to transform expression values into qualitative descriptors are evaluated in [29]. This model is used to find triplets of activators, repressors and targets in a gene expression data set. The predictions made by the algorithm agree well with experimental data. The algorithm can also assist in determining the function of uncharacterized proteins and is able to detect a substantially a larger number of transcription factors than that could be found at random. This technique extends current techniques such as clustering to allow the users to generate a connected network of genes using only expression data.

An investigation has been made using information on importance of genes in classification using fuzzy inferencing. This is similar to that of other classifiers, but simpler and easier to interpret [20]. The fuzzy inference system has the theoretical advantage that it does not need to be retrained when using measurements obtained from a different types of microarrays.

Interpretation of classification models derived from gene expression data is usually not simple. Rather, it is an important aspect in the analytical process. The performance of small rule-based classifiers is based on fuzzy sets and distribution of data [32]. The classifiers result in the rules that can be readily examined by biomedical researchers. The fuzzy logic-based classifiers compare favorably with logistic regression in all data sets they have considered.

Sokhansanj demonstrated an approach with exhaustive search for fuzzy gene interaction models that best fit transcription measurements obtained by microarray technology [27]. Applying an efficient, universally applicable data normalization and fuzzification scheme, the search converged to a small number of models that individually predict experimental data within an error tolerance. Although gene transcription levels are only used to develop the models, they include both direct and indirect regulation of genes. Biological relationships in the best-fitting fuzzy gene network models successfully recover direct and indirect interactions predicted from previous knowledge to result in transcriptional correlation. Fuzzy models that fit on data set were used for robust prediction of another experimental data set for the same system. Linear fuzzy gene networks and exhaustive rule search are the first steps towards a framework for an integrated modelling and experiment approach to high-throughput "reverse engineering" of complex biological systems.

Advances in molecular classification of tumors may play a central role in cancer treatment. Using gene expression profiles obtained by cDNA microarrays, a neural network model known as simplified fuzzy ARTMAP has been developed that is able to identify normal and tumor patients [4]. Furthermore, it makes the distinction among patients with molecularly different forms of carcinoma without any previous knowledge of those subtypes.

3 Proposed Method

In this section, we describe the method of generating linguistic rules on genes mediating the development of lung adenocarcinoma. The method has two parts. In the first part, we have used cyclic loess normalization technique [11] for transforming expression values in different normal samples of a gene into one value. Similarly, expression values in different lung adenocarcinoma samples of a gene into one value. The entire method is described in details below.

3.1 Normalization

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 experiment with different samples of either normal or diseased type. The purpose of normalization is to deal with this obscuring variation.

Here we use cyclic loess method [11] to normalize the data set for normal lung samples and as well as tumor samples. This approach is based upon the idea of the M versus A plot, where M is the difference in log expression values and A is the average of the log expression values corresponding to a pair of samples. An M versus A plot for normalized data should show a point cloud scattered about the $M = 0$ axis.

This is due to the fact that the expression values of the pair of samples become closer on application of a normalization method. In particular, for any two arrays i, j with probe intensities x_{ki} and x_{kj} where $k = 1, \ldots, p$ is 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 fitted to these M versus A plot using loess. 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 normalization is carried out in a pair wise manner, recording an adjustment for each of the two arrays in each pair. After looking at all pairs of arrays we have a set of adjustments that can be applied to the set of arrays. Then we repeat the process until the difference in the expression values becomes less than some predefined threshold. Typically only 5 or 6 complete iterations through all pair wise combinations are needed to achieve an acceptable result. 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 steps of the method [11] are provided below for the sake of clarity.

For each gene, do

STEP 1: Choose pair wise samples.

STEP 2: Compute M_k and A_k for each pair using the above method.

STEP 3: If there are n samples for a gene, there should be $\binom{n}{2}$ pairs, so as to get $\binom{n}{2}$ number of M_k and A_k.

STEP 4: Fit M_k with respect to A_k. Here we use the parabolic curve fitting algorithm. In this algorithm we use the formula $\hat{M}_k = a + bA_k + c{A_k}^2$. So for a set of A_k values, we can get a set of $\hat{M}_k$ values. Finally we can get $\binom{n}{2}$ number of $(M_k - \hat{M}_k)$ values. We call these $(M_k - \hat{M}_k)$ values as an adjustment.

STEP 5: Record these adjustments for each sample and compute the resultant adjustment of each sample.

STEP 6: Update the old log expression value for each sample by the following formula,

$$new \log_2 x_{ik} = old \log_2 x_{ik} + resultant\ adjustment$$

STEP 7: Repeat Step 1 to Step 6 until the differences between the log expressions values are less than some threshold values specified by the analyzer (i.e. repeat those steps until the log expression values of different samples are close enough).

3.2 Grouping (into classes) based on fuzzy sets

In conventional statistical methods, the absolute expression pattern of genes is presented to a system for computations. However, in real life situations, gene expression pattern may be uncertain and/or incomplete. In such cases it may become convenient to use linguistic variables such as *low*, *medium*, *high*, *very high*, or *more or less* to replace numerical feature information [23].

The proposed model is capable of handling absolute expression pattern i.e, numerical and inexact i.e, 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*.

Each input expression value x_{jk} of jth gene of kth sample in quantitative form can be expressed in terms of membership values to each of the three linguistic properties *low*, *medium* and *high*. Therefore for n samples, we have an n-dimensional gene expression pattern $\mathbf{x}_j = [x_{j1}, x_{j2}, \ldots, x_{jn}]^T$ for jth gene, which may be represented as a $3n$ dimensional vector

$$\mathbf{v}_j = [U_{low}(x_{j1}), U_{medium}(x_{j1}), U_{high}(x_{j1}), \ldots, U_{high}(x_{jn})]^T. \tag{1}$$

Here $U_{low}(x_{jk})$ is membership value of jth gene with expression value x_{jk} in kth sample, to the fuzzy set *low*. Hence in trying to express input $\mathbf{x}_j$ with the linguistic properties, we are effectively dividing the dynamic range of expression value into three overlapping partitions called low, medium, and high for each gene. Note that, for reducing complexity, we have already applied normalization

technique (described in Section 2.1) for transforming expression values in various samples of a gene into one value. This makes the dimension of $\mathbf{x}_j$ to be one and hence the dimension of $\mathbf{v}_j$ to be three. That is, for a jth gene we have only one expression value x_j.

We now describe the formulation of the membership functions corresponding to the fuzzy sets *low*, *medium* and *high*. These three membership functions are termed as U_{low}, U_{med}, U_{high} corresponding to the fuzzy sets *low*, *medium* and *high*. Here we have considered triangular membership functions for modelling the fuzzy sets. Thus, the membership function U_{low} is defined as

$$\begin{aligned} U_{low}(x_j) &= 1, && if\ x_j \le x_{min} \\ &= 1 + (x_j - x_{min})/(x_{min} - c_{med}), && if\ c_{low} \le x_j < c_{med} \\ &= 0, && otherwise \end{aligned} \tag{2}$$

Similarly, U_{med} and U_{high} are defined as

$$\begin{aligned} U_{med}(x_j) &= (x_{min} - x_j)/(x_{min} - c_{med}), && if\ c_{low} \le x_j < c_{med} \\ &= (x_{max} - x_j)/(x_{max} - c_{med}), && if\ c_{med} < x_j \le c_{high} \\ &= 0, && otherwise \end{aligned} \tag{3}$$

$$\begin{aligned} U_{high}(x_j) &= 1, && if\ x_j > c_{high} \\ &= 1 + (x_j - x_{max})/(x_{max} - c_{med}), && if\ c_{med} < x_j < x_{max} \\ &= 0, && otherwise \end{aligned} \tag{4}$$

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

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

The basic nature of these membership functions is as follows: (i) Maximum value of each function is 1. (ii) Minimum value of each function is 0. (iii) The membership functions corresponding to *low* and *medium*, cut at the point for which $U_{low} = U_{med} = 0.5$. Similar is the case for U_{med} and U_{high} such that at the point of intersections of the membership functions corresponding to *medium* and *high*, $U_{med} = U_{high} = 0.5$. (iv) The membership value corresponding to a gene expression value to a fuzzy set is maximum at the center of the fuzzy set and decreases as it is away from the center of the fuzzy set. It may be noted that one may use other membership functions for modelling the fuzzy sets *low*, *medium* and *high*, keeping the similar basic nature of the membership functions. The choice of c-values automatically ensures that one of the membership values U_{low}, U_{med} or U_{high} 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*. So 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, c-values) of the membership functions, for both normal and tumor samples are computed, based on the normal lung samples.

3.3 Rule generation based on linguistic variables

The membership values of various genes in both normal and tumor samples are used for rule generation in *if-then* form in order to justify any decision on some genes reached. These rules describe the extent to which a gene is responsible for causing adenocarcinoma in lung. The rules generated are in the form of *if-then*, where the antecedent is formed by two conjunctive clauses – one corresponding to the linguistic representation of a gene in normal samples and the other corresponding to that in tumor samples. The consequent part of the rule represents whether the tumor sample is adenocarcinomic or not. In order to generate the antecedent ("if") part of a rule, we compute confidence factor $CONF$ as given by

$$CONF = \frac{1}{2}[{v_{max}}^{n_{max}} + (1/(cl-1)) \times (\Sigma_j(v_{max} - v_j))],\ 0 \leq CONF \leq 1 \quad (5)$$

where $j = 1, 2, \ldots, cl$; cl being the number of classes. (Here $cl = 3$, as the classes are *low*, *medium* and *high*.) Here $v_{max} = \max_{j=1}^{cl}\{v_j\}$, v_j is the membership value to jth class and n_{max} indicates the number of occurrences of v_{max} in vector $\mathbf{v}$. Note that $CONF$ takes care of the fact that the difficulty in assigning a particular gene to a fuzzy class depends not only on the highest entry in the output vector v_{max} but also on its differences from the other entities v_j . It is seen that the higher the value of $CONF$, the lower is the difficulty in deciding a fuzzy set to which the gene belongs, and hence greater is the degree of certainty of the output decision. Based on the value of $CONF$ the system makes the following decisions heuristically while generating a rule. Let $v_k = v_{max}$ such that the pattern under consideration belongs to the class C_k. We have: (i) if $(0.8 \leq CONF_k \leq 1.0)$ then very likely fuzzy set C_k. (ii) if $(0.6 \leq CONF_k < 0.8)$ then likely fuzzy set C_k. (iii) if $(0.4 \leq CONF_k < 0.6)$ then more or less likely fuzzy set C_k. (iv) if $(0.1 \leq CONF_k < 0.4)$ then not unlikely fuzzy set C_k. (v) if $(CONF_k < 0.1)$ then unable to recognize fuzzy set C_k.

4 Results

In this section, the effectiveness of the proposed method is demonstrated on lung adenocarcinoma gene expression data [1, 2, 5].

4.1 Description of the data set

The data set is obtained by microarray experiments of Affymetrix Corporation for Ann Arbor tumors and normal lung samples [1, 2, 5]. In this data set, there

are expression values of 7129 genes (more specifically, Affymetrix probe-sets) for 86 lung tumor and 10 normal lung samples[19, 28].

Among these 86 tumor samples 67 samples corresponding to stage I and 19 to stage III tumors. There are also 10 neoplastic lung samples. The data set was trimmed of genes expressed at extremely low level. That is, genes were excluded if the measure of their 75th percentile value was less than 100 [2]. Array to array variation in the overall distribution of gene expression values detected by quantile-quantile plots was removed by applying a quantile normalization using a linear spline as a monotone transformation [5]. The gene expression profile of each tumor was normalized to the median gene expression value among all the samples. Features on the oligonucleotide arrays representing the genes in the individual tumors found as outliers were carefully reviewed to confirm expression levels and exclude artifacts. More details on this data set is found in [1, 2, 5].

4.2 Analysis of the results

The above data contains expression value of 10 normal samples of 7129 genes. So there are $\binom{10}{2}$ pairs, i.e., 45 pairs for each gene.

Table 1. Computation of resultant adjustment. Ad_i stands for the adjustment for ith sample.

Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
$+a_{1,2}/2$	$-a_{1,2}/2$	$-a_{2,3}/2$	$+a_{2,4}/2$	$-a_{2,5}/2$	$+a_{2,6}/2$	$-a_{2,7}/2$	$+a_{2,8}/2$	$-a_{2,9}/2$	$+a_{2,10}/2$
$-a_{1,3}/2$	$+a_{2,3}/2$	$+a_{1,3}/2$	$-a_{3,4}/2$	$+a_{4,5}/2$	$-a_{5,6}/2$	$-a_{3,7}/2$	$-a_{5,8}/2$	$+a_{3,9}/2$	$-a_{5,10}/2$
$-a_{1,4}/2$	$-a_{2,4}/2$	$+a_{3,4}/2$	$+a_{1,4}/2$	$+a_{3,5}/2$	$+a_{3,6}/2$	$+a_{4,7}/2$	$+a_{3,8}/2$	$+a_{4,9}/2$	$-a_{6,10}/2$
$+a_{1,5}/2$	$+a_{2,5}/2$	$-a_{3,5}/2$	$-a_{4,5}/2$	$-a_{1,5}/2$	$-a_{4,6}/2$	$+a_{5,7}/2$	$-a_{4,8}/2$	$+a_{5,9}/2$	$+a_{7,10}/2$
$-a_{1,6}/2$	$-a_{2,6}/2$	$-a_{3,6}/2$	$+a_{4,6}/2$	$+a_{5,6}/2$	$+a_{1,6}/2$	$+a_{6,7}/2$	$-a_{6,8}/2$	$+a_{6,9}/2$	$-a_{8,10}/2$
$-a_{1,7}/2$	$+a_{2,7}/2$	$+a_{3,7}/2$	$-a_{4,7}/2$	$-a_{5,7}/2$	$-a_{6,7}/2$	$+a_{1,7}/2$	$+a_{7,8}/2$	$-a_{7,9}/2$	$+a_{9,10}/2$
$-a_{1,8}/2$	$-a_{2,8}/2$	$-a_{3,8}/2$	$+a_{4,8}/2$	$+a_{5,8}/2$	$+a_{6,8}/2$	$-a_{7,8}/2$	$+a_{1,8}/2$	$+a_{8,9}/2$	$-a_{4,10}/2$
$+a_{1,9}/2$	$+a_{2,9}/2$	$-a_{3,9}/2$	$-a_{4,9}/2$	$-a_{5,9}/2$	$-a_{6,9}/2$	$+a_{7,9}/2$	$-a_{8,9}/2$	$-a_{1,9}/2$	$-a_{3,10}/2$
$+a_{1,10}/2$	$-a_{2,10}/2$	$+a_{3,10}/2$	$+a_{4,10}/2$	$+a_{5,10}/2$	$+a_{6,10}/2$	$-a_{7,10}/2$	$+a_{8,10}/2$	$-a_{9,10}/2$	$-a_{1,10}/2$
Ad_1	Ad_2	$-Ad_3$	Ad_4	$-Ad_5$	Ad_6	$-Ad_7$	$-Ad_8$	$-Ad_9$	Ad_{10}

We have first of all, applied normalization algorithm described in Section 2.1. According to the algorithm, we have consider pairwise samples and calculate adjustment iteratively until the expression values of the two samples become very closed. This is depicted in Table 1. In Table 1, $a_{1,2}$ indicates the adjustment value of pair, sample 1 and sample 2. We have distributed this adjustment value to the sample 1 and sample 2. Here the log expression value of sample 1 is less than sample 2. So we divided the adjustment, say $a_{1,2}$, such that sample 1 got $+a_{1,2}/2$ and sample 2 got $-a_{1,2}/2$. In this way each sample has 9 values after distribution. Finally, we calculate the resultant adjustment by adding those values including sign for each sample. Now for each sample, we update the old log expression value of a gene by adding resultant adjustment of the corresponding sample. This completes one iteration. After 5 or 6 iterations we have got the normalized value of each sample for a gene. That is, log expression value of the 10 samples

become close enough. In this way we normalized 7129 genes for normal and tumor samples. After normalization, we have performed mean operation on the values of the genes. So ultimately we represented each gene with a single value. This would help us a lot for further analysis as well as for implementation with respect to the complexity of the problem is concerned.

The membership values of 7129 genes to the classes *low*, *medium* and *high* were then calculated using equations (2)-(4), and grouped into three fuzzy classes based on these membership values. In case of normal samples there are 6288 genes in class *low*, 835 genes in class *medium*, 6 genes in class *high*. Similar steps were followed for the 86 tumor samples using the parameter values already computed for normal samples. In the case of tumor samples, there are 6251 genes in class *low*, 871 genes in class *medium*, 7 genes in class *high*. It is interesting to note that number of genes of corresponding to the classes of normal and tumor samples changed significantly. In order to determine the extent of changes, we have compared these classes, i.e., between (low_{normal}, low_{tumor}), ($medium_{normal}$, $medium_{tumor}$), ($high_{normal}$, $high_{tumor}$). Based on this comparison, we have identified a set of 293 genes, each of which has changed the corresponding class. That is, one of these 293 genes may belong to the class *low* for normal samples, but is included in the class other than *low* for tumor samples.

Finally, we applied the rule generation technique based on the algorithm specified in Section 2.3 on these 293 genes to generate 293 rules. Among these 293 genes, we have reported the rules for 9 genes that have changed significantly from normal to tumor samples. These rules are provided in Table 2. For example, the rule for the gene FMO2 is as follows:

If FMO2 is very likely in class medium for normal samples and very likely in class low for tumor samples then the tumor sample is adenocarcinomic.

The results are validated by some earlier investigations on these genes. That is, these genes were found to be responsible for lung adenocarcinoma by these investigations. These include the references in [3] for RPL0, [10, 24] for ADH1, [22, 26] for UGB, [17] for FMO2, [8, 13] for HBA2, [30, 15, 25] for SFTPA1, [21] for SFTPA2, [16, 9] for HBB.

5 Conclusions

In this article, we have described a rule generation method for identifying a few genes responsible for a specific disease. First of all, the expression values for genes in different samples were normalized to remove sources of variation between the arrays. Here we have used a cyclic loess normalization method [11] on normal and tumor samples. After normalization, we have performed the mean operation that is mainly used to represent a gene by a single log expression value. We have then applied the concept of fuzzy sets to classify the genes into three fuzzy classes, viz., *low*, *medium*, and *high*. Incorporation of fuzzy set theory makes the system capable of handling uncertainty in the gene expression values arising due to incompleteness, imprecision, noise and experimental errors. Now we have

Table 2. Rules corresponding to the nine most significant genes mediating the development of lung adenocarcinoma. The consequent parts of all these rules are "the tumor sample is adenocarcinomic".

Gene Name	Antecedent clauses corresponding to	
	Normal	Tumor
RPLP0	Likely medium	Very likely high
NULL	Very likely medium	Very likely high
ADH1	Likely medium	Very likely low
UGB	Likely medium	Very likely low
FMO2	Very likely medium	Very likely low
HBA2	Likely medium	Very likely low
SFTPA1	Very likely high	Likely medium
SFTPA2	Very likely high	Likely medium
HBB	Very likely high	Very likely low

performed matching operation with normal classes with the tumor classes. After matching we have identified the genes that moved significantly from one class of normal to another class of tumor or vice versa.

Applying the above method on a lung adenocarcinoma data set containing gene expression values, we have grouped those genes into three different classes *low*, *medium*, *high*. The *low*, *medium* and *high* classes for normal consist of 6288, 835 and 6 genes respectively. Similarly, the *low*, *medium* and *high* classes of tumor consists of 6251, 871 and 7 genes respectively. We have compared all three classes of normal with *low*, *medium* and *high* classes of the tumor. On this comparison, we have found 293 genes that are significantly changed their expression level from normal to the tumor. Among these 293 genes, we have reported nine genes that have changed significantly based on the rule we have got. These 9 genes include RPLP0, ADH1, UGB, FMO2, HBA2, SFTPA1, SFTPA2, HBB. Among those 9 genes we have found, the gene HBB that have moved most significantly from class *high* of normal to the class *low* of the tumor. We have reported that this gene is a down-regulated gene that is mediating the development of lung carcinoma. Therefore, over or under expression of these 9 genes are responsible for the development of lung carcinoma. These results have been validated by a number of earlier investigations.

References

1. http://www.camda.duke.edu/camda03/datasets/
2. http://groups.yahoo.com/group/camdadata/
3. Abo, Y., Hagiya, A., Naganuma, T., Tohkairin, K. S. Y., Kajiura, Z., Hachimori, A., Uchiumi, T., Nakagaki, M.: Baculovirus-mediated expression and isolation of human ribosomal phosphoprotein p0 carrying a gst-tag in a functional state. Biochem Biophys Res Commun. **322** (2004) 814–819

4. Azuaje, F.: A computational neural approach to support the discovery of gene function and classes of cancer. IEEE Transactions on Biomedical Engineering **48** (2001) 332–339
5. Beer, G. D., et. al.: Gene-expression profilespredict survival of patients with lung adenocarcinoma. Nature Medicine **8** (2002) 816–823
6. Bhattacharjee, A., et. al.: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci., USA **98** (2001) 13790–13795
7. Brazma, A., et. al.: Minimum information about a microarray experiment (miame)-toward standards for microarray data. Nat Genet **29** (2001) 365–371
8. Brown, J. M., Leach, J., Reittie, J. E., Atzberger, A., Lee-Prudhoe, J., Wood, W. G., Higgs, D. R., Iborra, F. J., Buckle, V. J.: Coregulated human globin genes are frequently in spatial proximity when active. J Cell Biol. **172** (2006) 177–187
9. Buzina, A., Aladjem, M. I., Kolman, J. L., Wahl, G. M., Ellis, J.: Initiation of DNA replication at the human beta-globin 3' enhancer. Nucleic Acids Res. **33** (2005) 4412–4424
10. Dannenberg, L. O., Chen, H. J., Edenberg, H. J.: Gata-2 and hnf-3beta regulate the human alcohol dehydrogenase 1a (adh1a) gene. DNA Cell Biol. **24** (2005) 543–552
11. Dudoit, S., Yang, Y. H., Callow, M. J., Speed, T. P.: Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Stat. Sin **12** (2002) 111–139
12. Garber, M. E., et. al.: Diversity of gene expression in adenocarcinoma of the lung. Proc. Natl. Acad. Sci. USA **98** (2001) 13784–13789
13. Haider, M., Adekile, A.: Alpha-2-globin gene polyadenylation (AATAAA → AATAAG) mutation in hemoglobin H disease among kuwaitis. Med Princ Pract. **14 Suppl 1** (2005) 73–76
14. Hartemink, A., Gifford, D., Jaakkola, T., Young, R.: Maximum likelihood estimation of optical scaling factors for expression array normalization. SPIE BIOS (2001)
15. Jiang, F., Caraway, N. P., Bekele, B. N., Zhang, H. Z., Khanna, A., Wang, H., Li, R., Fernandez, R. L., Zaidi, T. M., Johnston, D. A., Katz, R. L.: Surfactant protein a gene deletion and prognostics for patients with stage I non-small cell lung cancer. Clin Cancer Res. **11** (2005) 5417–5424
16. Jiang, Y., Xu, X. S., Russell, J. E.: A nucleolin-binding 3' untranslated region element stabilizes beta-globin mrna in vivo. Mol Cell Biol. **26** (2006) 2419–2429
17. Krueger, S. K., Siddens, L. K., Henderson, M. C., Andreasen, E. A., Tanguay, R. L., Pereira, C. B., Cabacungan, E. T., Hines, R. N., Ardlie, K. G., Williams, D. E.: Haplotype and functional analysis of four flavin-containing monooxygenase isoform 2 (FMO2) polymorphisms in hispanics. Pharmacogenet Genomics. **15** (2005) 245–256
18. Liotta, L., Petricion, E.: Molecular profiling of human cancer. Nature Rev. Genet **1** (2000) 48–56
19. Lipshutz, R., Fodor, S., Gingeras, T., Lockart, D.: High density synthetic oligonucleotide arrays. Nat. Genet **21** (1999) 20–24
20. Machado, L., Vinterbo, S., Weber, G.: Classification of gene expression data using fuzzy logic. Journal of Intelligent and Fuzzy Systems **12** (2002) 19–24
21. Madan, T., Saxena, S., Murthy, K. J., Muralidhar, K., Sarma, P. U.: Association of polymorphisms in the collagen region of human sp-a1 and sp-a2 genes with pulmonary tuberculosis in indian population. Clin Chem Lab Med. **40** (2002) 1002–1008

22. Martin, A. C., Laing, I. A., Khoo, S. K., Zhang, G., Rueter, K., Teoh, L., Taheri, S., Hayden, C. M., Geelhoed, G. C., Goldblatt, J., LeSouef, P. N.: Acute asthma in children: Relationships among cd14 and cc16 genotypes, plasma levels, and severity. Am J Respir Crit Care Med. **173** (2005) 617–622
23. Mitra, S., De, R. K., Pal, S. K.: Knowledge-based fuzzy mlp for classification and rule generation. IEEE Transactions on Neural Networks **8** (1997) 1338–1350
24. Rodriguez-Zavala, J. S., Weiner, H.: Structural aspects of aldehyde dehydrogenase that influence dimer-tetramer formation. Biochemistry **41** (2002)
25. Saxena, S., Kumar, R., Gupta, T. M. V., Muralidhar, K., Sarma, P. U.: Association of polymorphisms in pulmonary surfactant protein a1 and a2 genes with high-altitude pulmonary edema. Chest **128** (2005) 1611–1619
26. Shashikant, B. N., Miller, T. L., Welch, R. W., Pilon, A. L., Shaffer, T. H., Wolfson, M. R.: Dose response to RHCC10-augmented surfactant therapy in a lamb model of infant respiratory distress syndrome: physiological, inflammatory, and kinetic profiles. J Appl Physiol. **99** (2005) 2204–2211
27. Sokhansanj, B. A., Fitch, J. P., Quong, J. N., Quong, A. A.: Linear fuzzy gene network models obtained from microarray data by exhaustive search. BMC Bioinformatics **5** (2004) 108–119
28. Warrington, J. A., Dee, S., Trulson, M. (editors): Large-scale genomic analysis using Affymetrix genechip. M. Schena (2000) 119–148
29. Woolf, P. J., Wang, Y.: A fuzzy logic approach to analyzing gene expression data. Physiol Genomics **3** (2000) 9–15
30. Wu, Y. Z., Manevich, Y., Baldwin, J. L., Dodia, C., Yu, K., Feinstein, S. I., Fisher, A. B.: Interaction of surfactant protein a with peroxiredoxin 6 regulates phospholipase a2 activity. J Biol Chem. **281** (2006) 7515–7525
31. Zhang, H. et. al.: Recursive partitioning for tumor classification with gene expression microarray data. Proc Natl Acad Sci USA **98** (2001) 6730–6735
32. Vinterbo, S. A., Kim, E. Y., Machado, L.: Small, fuzzy and interpretable gene expression based classifiers. Bioinformatics **21** (2005) 1964–1970