Journal of Intelligent & Fuzzy Systems 26 (2014) 2731–2746
DOI:10.3233/IFS-130942
IOS Press

Development of a fuzzy entropy based method for detecting altered gene–gene interactions in carcinogenic state

Anupam Ghosh[a] and Rajat K. De[b,*]
[a]*Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India*
[b]*Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India*

Abstract. In this article, we propose a methodology for identifying the interactions among the genes in terms of dependencies (named as gene–gene interaction) that have altered quite significantly from normal stage to diseased stage with respect to their expression patterns. This idea leads to predict the disease mediating genes along with their altered interactions. The proposed methodology involves measuring information content of individual genes using fuzzy entropy, conditional fuzzy entropy of a gene on another, dependencies (interactions) of a pair of genes in both normal and diseased states, detecting the dependencies being deviated from normal to carcinogenic state and finally identifying the influential genes from altered dependencies. Thus the gene–gene interactions for normal state and diseased state are represented separately by the gene dependency networks (GDN). The altered interactions among the genes have been represented using a network, called altered gene dependency network ($AGDN$), in which each node represents a gene and a directed edge signifies altered dependency between a pair of nodes (genes). The methodology has been demonstrated on five gene expression data sets dealing with human lung cancer, colon cancer, sarcoma, breast cancer and leukemia. The results are appropriately validated, in terms of gene–gene interactions, using biochemical pathways, t-test, p-value, NCBI database and earlier investigations in terms of gene regulation. We have also used sensitivity to validate the results. For a comparative study, we have used some existing association rule mining algorithms and frequent pattern mining algorithms like Fuzzy Cluster-Based Association Rules, Apriori, T-Apriori in terms of gene–gene interactions. In addition, we have implemented Significance Analysis of Microarray, Signal-to-Noise Ratio, Neighborhood analysis, Bayesian regularization and frequent pattern mining algorithms for a comparison with $AGDN$ in terms of ability to identify the important genes mediating the cancers.

Keywords: Gene dependency networks, altered gene dependency networks, lung cancer, colon cancer leukemia, sarcoma, breast cancer, t-test, p-value

1. Introduction

Transcriptional regulatory networks are crucial in the understanding of fundamental cellular processes and functions. The determination of factors that control expression level can offer further insight into the miss regulated expression that is common in many human diseases [1, 2]. There exist many investigations on identifying transcription factors including those through sequence similarity [3, 4], motif binding [5–8] and through the dynamics of gene expression patterns [9, 10].

Various statistical approaches have been developed to capture gene regulations using dynamic Bayesian networks [11, 12], vector autoregressive models [13], and state space models [14, 15] based on statistical

Corresponding author. Rajat K. De, Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India. E-mails: rajat@isical.ac.in (Rajat K. De); anupam.ghosh@rediffmail.com (Anupam Ghosh).

1064-1246/14/$27.50 © 2014 – IOS Press and the authors. All rights reserved

causality, among others. Traditional approaches include formulation of a set of coupled differential equations and their solutions, with the objective to obtain a deeper understanding of the exact nature of the regulatory circuits and their regulation mechanisms [16]. Various alternative methods, like relevance networks [17] and graphical Gaussian models [18] have been proposed and applied to the inference of gene regulatory networks from gene expression data. The identification of several disease-associated polymorphisms by whole genome analysis is now leading to the detection of interactions among genetic (or environmental) factors [19]. However, this type of investigations is till now in nascent stage.

In the literature, we have found some investigations on association rule mining for building large-scale gene regulation networks from microarray data [20]. Here an association rule mining technique has been used to generate the patterns for probing the spatially-mapped gene expression domains. Prediction gene networks from microarray gene expression data using the association rule mining techniques include dynamic Bayesian network [21], a novel algorithm by [22], a temporal association rule mining (TARM) technique that can extract temporal dependencies among genes from time-series microarray data sets [23]. In the above study, the researchers tried to find out the associations among the genes from microarray gene expression data sets by applying a set association rule mining algorithms.

But it is hard to find investigations on the dependencies among the genes, in terms of regulation, which have changed from normal to diseased state of cancer datasets. Association rule network provides a graphical view of the associations among various objects [24] have developed the concept of association rule networks as a structure for synthesizing, pruning and analyzing a collection of association rules to construct hypothesis to draw inferences on the dependence of certain objects on the others. In the present manuscript, we provide the notion of gene dependency networks to show pictorially the dependence of gene(s) on the other, in terms of transcriptional regulations. Thus there is a correspondence between association rule networks and gene dependency networks. In another study, association rule mining and network analysis have been applied on a vast amount of prescription data that is generated from the book of oriental medicine to identify the relationships between the symptoms and the associated medicines [25]. However, we could not find any algorithm or technique like ARN that can construct association networks among the genes from gene expression data. Moreover, the altered of the associations as well as the corresponding networks between the different states of a disease, is one of the most challenging task nowadays that motivates us to work on the present issue.

Thus, in the present article, we concentrate on identifying dependencies (interactions) among the genes obtained from microarray gene expression patterns. Here we develop a methodology for identifying the dependencies among the genes (gene–gene interaction), which have changed from normal to diseased state. In this way, we can predict disease mediating genes along with their altered dependencies (interactions). The methodology involves measuring information content of individual genes using fuzzy entropy, conditional fuzzy entropy of a gene on another, dependencies of a pair of genes in both normal and diseased states, and finally identifying the dependencies that have altered from normal to diseased state. A fuzzy entropy is a "measure of the degree of fuzziness" of a generalized set [26]. It is introduced using non-probabilistic concepts in order to obtain a global measure of indefiniteness connected with the situations described by fuzzy set theory. All the dependencies (interactions) among the genes are represented by networks, called gene dependency networks (GDNs) for normal and diseased states separately. Similarly, altered dependencies are represented by altered gene–gene dependency networks ($AGDN$s).

There exist several approaches of frequent pattern mining algorithms [27] for discovering such dependencies/associations, although none of them has been applied to gene expression profiles. These include Fuzzy Cluster-Based Association Rules (FCBAR) [28], Apriori [29], T-Apriori [30], DGN [31], Max-conf [32], Pattern Fusion [33], TD-CLOSE [34], TOPKRGS [35] algorithms. We have considered these algorithms for the comparison of the associations among the genes resulted in by them with the resulting GDN obtained from normal gene expression profiles.

In the present study, we have proposed a technique to find out the influential genes from $AGDN$. Several attempts have been made during the past several years for developing methodologies or using feature selection algorithms that select informative genes from microarray gene expression data. These informative genes improve the efficiency of the system in terms of disease prediction accuracy. These attempts include Significance Analysis of Microarray (SAM) [36], Signal-to-Noise Ratio (SNR) [37], Neighborhood analysis (NA) [38], Bayesian regularization (BR) [39, 40] and GeneSelector [41]. we have consid-

ered these algorithms for comparison. The effectiveness of the methodology has been demonstrated on five gene expression data dealing with human lung cancer, colon cancer, sarcoma, breast cancer and leukemia. The results are appropriately validated in terms of gene–gene interaction, using biochemical pathways, *t*-test, *p*-value, NCBI database. The results have been also validated with some existing investigations in terms of gene regulation.

2. Methodology

Here, we formulate the methodology for developing a gene dependency network based on the gene expression patterns of normal and diseased samples. First of all, two gene dependency matrices (GDMs) are formed for gene expression profiles of normal and diseased samples respectively. Based on these two matrices, a prediction network is created involving the genes, where dependency has been changed from normal to diseased state. This prediction network provides an idea on the genes along with their dependencies for mediating carcinogenic development. In other words, the network may predict the responsible genes and their altered behavior. Then we find out some possible genes mediating the development of cancer. We have used fuzzy set theoretic entropy to determine the gene dependency matrices. Using these matrices, we build the altered gene dependency network.

The term *gene dependency* may be defined as follows. Let us consider two genes g_i and g_j. If the change in expression level of gene g_j causes the change in expression level of gene g_i, then we say that gene g_i depends on gene g_j. In other words, gene g_j regulates the expression level of gene g_i. Thus we have an $n \times n$ matrix GDM corresponding to the gene dependencies involving n genes. The (i, j)th entry of GDM represents the degree of dependency of g_i on g_j. In this way, GDMs are formed for normal samples as well as for diseased samples. In diseased state, this dependency may change from normal ones. From these two gene dependency matrices, we may have an idea of a gene regulatory network depicting transcriptional regulation for normal and diseased states.

The methodology involves six steps. In Step 1, the information content of a gene is computed using the notion of fuzzy entropy from gene expression data set. The conditional entropy of a pair of genes is computed in Step 2. In Step 3, the information gain is computed pairwise and also generates the gene dependency matrix using the concept of symmetrical uncertainty. The values of the gene dependency matrix are quantized, in Step 4, using a threshold. In Step 5, the gene dependency network (GDN) and the altered gene dependency network ($AGDN$) are generated. Finally, in Step 6, the influential genes are identified from altered gene dependency network ($AGDN$).

2.1. Step 1 – Measuring information content (entropy) of a gene

Let G be the set of n genes $\{g_1, g_2, g_3, \ldots, g_n\}$. For each gene g_i, there is an m-dimensional vector $\mathbf{x}_i$, where x_{il} is the l-th expression value (i.e., in lth sample obtained by lth experiment) of gene g_i. Here we consider a fuzzy set around a gene g_i and the membership function $U_{g_i}(l)$ signifies the degree of membership of lth expression value of gene g_i to this fuzzy set. In other words, $U_{g_i}(l)$ represents the extent by which g_i is expressed in lth sample. Then we compute the entropy (uncertainty) associated with this fuzzy set, i.e., associated with gene g_i, as

$$H(g_i) = \sum_{l=1}^{m} U_{g_i}(l)(1 - U_{g_i}(l)) \tag{1}$$

Here the membership function $U_{g_i}(l) \in [0, 1]$ is defined as

$$U_{g_i}(l) = \exp(-|x_{il} - \bar{x}_i|) \tag{2}$$

where $\bar{x}_i$ is given by

$$\bar{x}_i = \tfrac{1}{m} \sum_{l=1}^{m} x_{il} \tag{3}$$

2.2. Step 2 – Measuring conditional entropy of a gene on another

In the previous step, we have calculated the uncertainty associated with each gene, i.e., information content of an individual gene. Now the entropy associated with gene g_i given that gene g_j has attained some expression value, is defined as

$$H(g_i|g_j) = \sum_{l=1}^{m} U_{g_j}(l) \sum_{l'=1}^{m} U_{g_i|g_j}(l')(1 - U_{g_i|g_j}(l')) \tag{4}$$

where $U_{g_i|g_j}(l) \in [0, 1]$ is defined as

$$U_{g_i|g_j}(l) = \exp(-||\mathbf{s}_{ij}(l) - \bar{\mathbf{s}}_{ij}||)/\exp(-|x_{jl} - \bar{x}_j|) \tag{5}$$

Here $\mathbf{s}_{ij}(l) = [x_{il},\ x_{jl}]^T$ and $\bar{\mathbf{s}}_{ij} = [\bar{x}_i,\ \bar{x}_j]^T$. That is, we are writing $U_{g_i|g_j}(l)$ as $U_{g_ig_j}(l)/U_{g_j}(l)$, where $U_{g_ig_j}(l)$ is a two dimensional membership function of the fuzzy set formed by genes g_i and g_j together.

2.3. Step 3 – Measuring information gain and building gene dependency matrix

The entropy $H(g_i)$ of the gene g_i has two parts. The part $H(g_i|g_j)$ represents the entropy of gene g_i given that gene g_j has attained some expression value. The remaining part represents the information gain of gene g_i provided by gene g_j, and is defined as [42],

$$IG(g_i|g_j) = H(g_i) - H(g_i|g_j) \tag{6}$$

The amount by which the entropy of g_i decreases reflects additional information about g_i provided by g_j and is called information gain. According to this measure, a gene g_j is regarded as more correlated to gene g_i than gene g_k, if $IG(g_i|g_j) > IG(g_k|g_j)$. Information gain may be biased towards the genes with more expression values. We normalize IG-values to get Normalized IG-values or NIG-values as [42],

$$NIG(g_i, g_j) = 2 \times |IG(g_i|g_j)|/(H(g_i) + H(g_j)) \tag{7}$$

$NIG(g_i, g_j)$ signifies the strength of dependency, i.e., to what extent the expression of gene g_i depends on the expression value of g_j. $NIG(g_i, g_j) = 1$ indicates that knowledge of the expression level of either one completely predicts that of the other, and $NIG(g_i, g_j) = 0$ indicating that g_i and g_j are independent. With these NIG-values, we get the gene dependency matrix GDM of order $n \times n$ and with (i, j)th element as $NIG(g_i, g_j)$.

2.4. Step 4 – Quantizing the elements of gene dependency matrix

All entries of GDM are in $[0, 1]$ and provide the degrees of dependency. Here we introduce three types of dependencies (i.e., low, medium, high) between a pair of genes. To represent each type, we have to fix a threshold value to generate the quantized gene dependency matrix $D = [d_{ij}]_{n\times n}$.

The domain of the matrix elements is divided into three partitions: $[0, \alpha]$ as the first partition, $(\alpha, \beta]$ as the second one, and $(\beta, 1]$ as the third partition. In this case, α and β are treated as user defined thresholds. Now d_{ij} values, where $i \neq j$, are defined as

$$\begin{aligned} d_{ij} &= 0.0 \text{ if } 0 \le d_{ij} \le \alpha \\ &= 0.5 \text{ if } \alpha < d_{ij} \le \beta \\ &= 1.0 \text{ if } \beta < d_{ij} \le 1, \end{aligned} \tag{8}$$

and $d_{ii} = 1$, for all i. Thus the quantized gene dependency matrix is formed with the values 0.0, 0.5 and 1.0. Note that, $d_{ij} = 1.0$ means gene g_i is highly dependent on gene g_j. $d_{ij} = 0$ signifies that the dependency of g_i on g_j is low, and $d_{ij} = 0.5$ implies that this dependency is medium.

2.5. Step 5 – Building of gene dependency network (GDN) and altered gene dependency network ($AGDN$)

In the previous step, we have defined gene dependency matrix D. This matrix is computed for control (normal) samples as well as for test (diseased) samples. We denote the gene dependency matrix by $D^{(N)}$ for normal samples, and by $D^{(D)}$ for diseased samples. From each matrix, the gene dependency network (GDN) is constructed. The network (GDN) is constructed in the following way:

- $d_{ij} = 1.0$: In this case, there is a edge directed from node g_j to node g_i (the edge is labeled with 1.0).
- $d_{ij} = 0.0$: In this case, there is no edge between g_i and g_j.
- $d_{ij} = 0.5$: In this case, there is a edge directed from node g_j to node g_i (the edge is labeled with 0.5).

Using these two matrices, we generate a matrix $P = D^{(N)} - D^{(D)}$ of order $n \times n$. The matrix $P = [p_{ij}]_{n\times n}$ consists of the values of 0.0, −1.0, 1.0, −0.5 and 0.5. From matrix P, the altered gene dependency network ($AGDN$) is constructed. Thus, $AGDN$ represents the set deviated gene–gene interactions. The significance of these values are discussed below.

- $p_{ij} = 0.0$: In this case, $d_{ij}^{(N)} = d_{ij}^{(D)}$. That is, the diseased state does not affect the dependency of gene g_i on gene g_j.
- $p_{ij} = -1.0$: In this case, $d_{ij}^{(N)} = 0.0$ and $d_{ij}^{(D)} = 1.0$. It indicates that two genes g_i and g_j of low dependency in normal state become highly dependent in the diseased state.
- $p_{ij} = 1.0$: That is, $d_{ij}^{(N)} = 1.0$ and $d_{ij}^{(D)} = 0.0$ indicates that two highly dependent genes g_i and g_j in normal state become independent in diseased state.

$p_{ij} = 0.5$ (-0.5): In this case, dependency of g_i on g_j in diseased condition is partially lost (gained).

2.6. Step 6 – Selection of Influential genes from altered gene dependency network ($AGDN$)

As already mentioned earlier, we have developed a methodology to establish the concept of gene dependency from gene expression data sets. This concept leads to develop a altered gene dependency network ($AGDN$) that shows the altered dependencies among a set of genes from normal to carcinogenic samples. In this section, we try to validate the results in a different way. From $AGDN$, we find out the set of *influential genes*. Here the term *influential genes* is defined as the genes that are involved in at least one altered dependency in $AGDN$.

3. Results

In this section, the effectiveness of the methodology is demonstrated on five cancer gene expression data sets. The details of these datasets are available in the Supplementary Material. This is followed by validation of the results.

3.1. Analysis of the results

Lung expression data [43] contains 10 normal samples and 86 tumor samples for expression values of 7129 genes. We have applied the methodology to this data set. It has been found that 187 (162) genes have lost (gained) their dependencies from normal samples to tumor samples; 275 (303) genes have lost (gained) their dependencies partially. Similarly for human breast expression data set (expression levels of 22645 genes for 2 normal breast epithelial cells and 4 samples for breast cancer cells) [44], 356 (371) genes have been found to loose (gain) their dependencies from normal to cancer cells; 510 (419) genes have lost (gained) their dependencies partially. For human colon expression data (6600 genes with 18 normal and 18 cancer samples) [45], it has been found that 121 (147) genes have lost (gained) their dependencies from normal samples to tumor samples; 198 (168) genes have lost (gained) their dependencies partially. In the case of human lymphocyte cell expression data (22283 genes with 13 normal samples and 43 cancer samples) [46], 298 (335) genes have been found to loose (gain) the dependencies; 308 (486) genes have lost (gained) their dependencies partially. Finally, considering human sarcoma expression data (22283 genes with 15 normal samples and 39 cancer samples) [47], it has been found that the dependencies have been lost (gained) for 215 (346) genes from normal samples to tumor samples; 510 (379) genes have lost (gained) their dependencies partially.

Figure 3 shows the altered gene dependency network ($AGDN$) for lung expression data. Here we have showed only the altered dependencies that have changed completely from normal to diseased state, i.e., $p_{ij} = 1$ or -1. Figure 1 and 2 depict dependencies among the genes, respectively in normal and diseased states, which are involved in GDN and $d_{ij} = 1$. In order to restrict the size of the article, we have included corresponding figures (Figs. 10–21) for the other data sets in the Supplementary Material.

Now we consider gene expression profiles of cancer samples in different stages. For lung carcinoma data, we have applied the methodology on various diseased states of human lung adenocarcinoma (86 tumor samples including 67 stage I tumor samples and 19 stage III tumor samples). In this case, 32 (41) genes have lost (gained) their dependencies from stage I tumor samples to stage III tumor samples. It has also been noted that 55 (48) genes have lost (gained) their dependencies from stage I to stage III partially.

For human leukemia data set, 43 diseased samples are classified into three diseased states of human lymphocytes and plasma cell expression. Out of these 43 samples, 20 samples for Waldenstrom's macroglobulinemia, 11 for chronic lymphocytic leukemia and the remaining 12 for multiple myeloma [46]. It has been found that 22 (31) genes have lost (gained) their dependencies from Waldenstrom's macroglobulinemia to chronic lymphocytic leukemia. Likewise, 38 (45) genes have lost (gained) their dependencies from Waldenstrom's macroglobulinemia to chronic lymphocytic leukemia partially. Similarly, 27 (19) genes have lost (gained) their dependencies from chronic lymphocytic leukemia to multiple myeloma; 49 (55) genes lost (gained) from chronic lymphocytic leukemia state to multiple myeloma state partially. It has also been reported that 56 (63) genes have lost (gained) their dependencies from Waldenstrom's macroglobulinemia to multiple myeloma state and the dependencies have been lost (gained) for 71 (67) genes partially from Waldenstrom's macroglobulinemia state to multiple myeloma.

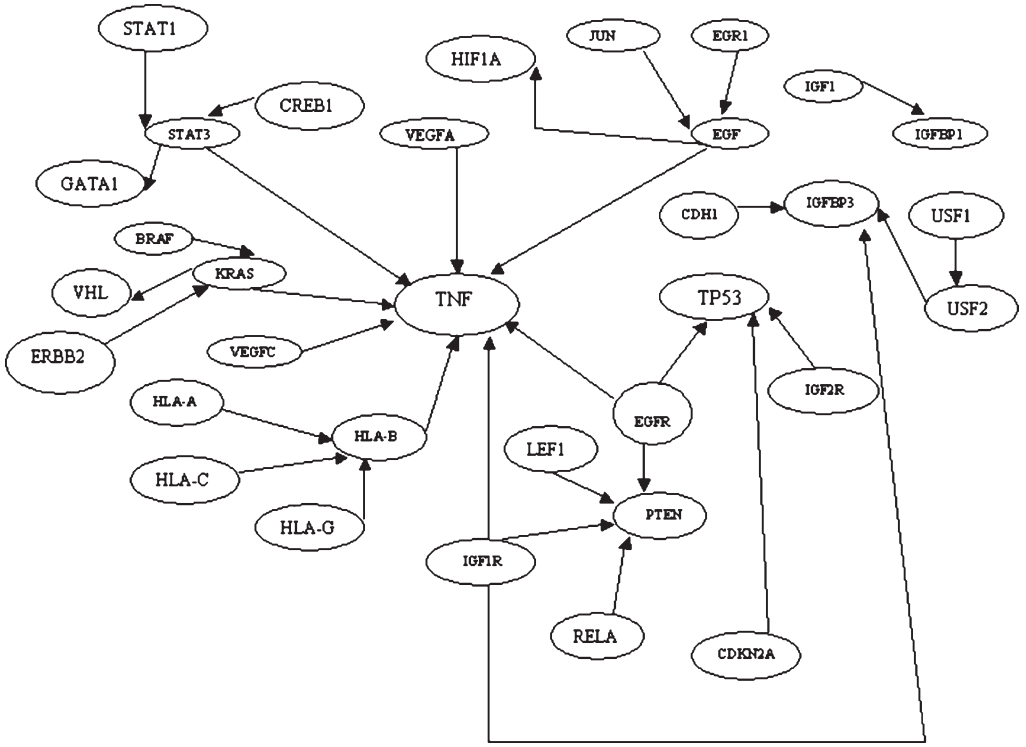

Fig. 1. GDN for lung expression data in normal state. Here we consider the set of dependencies for the genes in GDN, where $d_{ij} = 1.0$.

3.2. Comparison and validation of results

In this section, we compare the results obtained by $AGDN$. The comparison has been done on gene–gene interactions, using biochemical pathways, t-test and sensitivity. Here we have implemented 8 existing association rule mining algorithms (i.e., FCBAR [28], Apriori [29], T-Apriori [30], DGN [31], Max-conf [32], Pattern Fusion [33], TD-CLOSE [34] and TOPKRGS [35]) and 5 other methods (i.e., Significance Analysis of Microarray (SAM) [36], Signal-to-Noise Ratio (SNR) [37], Neighborhood analysis (NA) [38], Bayesian regularization (BR) [39, 40] and GeneSelector [41]) for comparison. Moreover, we have also tried to validate some of our results using some earlier investigations.

3.2.1. On gene–gene interactions

Let us consider y is a function of x, i.e., $y = f(x)$. This means any variation of x will affect y. Thus, we can say that y depends on x. It is represented as $x \rightarrow y$. In other words, we can say there is an association exists between x and y. This means that x interacts with y. Now, consider the above example for gene expression data such that x and y are two genes. Thus there exists an association if $y = f(x)$, and gene x interacts with y. We call this interaction between gene x and y as *gene–gene interaction*.

Here, we validate the results obtained by GDN with various cancer datasets, in terms of gene–gene interactions. By gene–gene interactions, we mean corresponding protein-protein interactions. For this purpose, we have mainly focussed on the fact that how GDN has discovered correctly the associations in terms of gene–gene interactions with respect to various cancer datasets. In this context, we applied the above mentioned association rule mining algorithms for a comparative study with GDN in terms of gene–gene interactions. We get pathway related information from bio-system database that consists of a set of genes and their interactions with other genes. We have found some cancer specific pathways from bio-system (pathway) database of NCBI (www.ncbi.nlm.nih.gov/biosystems/). In the database, we have found non-small cell lung cancer, small cell cancer, colorectal cancer, chronic myeloid and acute myeloid leukemia related pathways. These pathways are related to human lung, colon and lymphocyte and plasma cell. However, we could not find similar

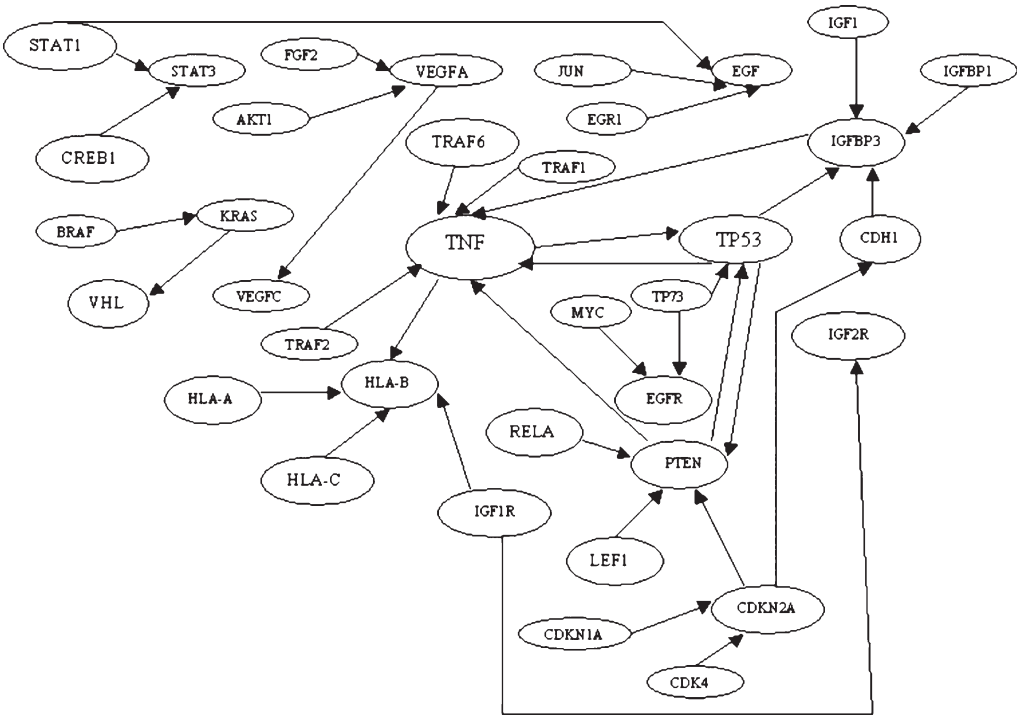

Fig. 2. GDN for lung expression data in diseased state. Here we consider the set of dependencies for the genes in GDN, where $d_{ij} = 1.0$.

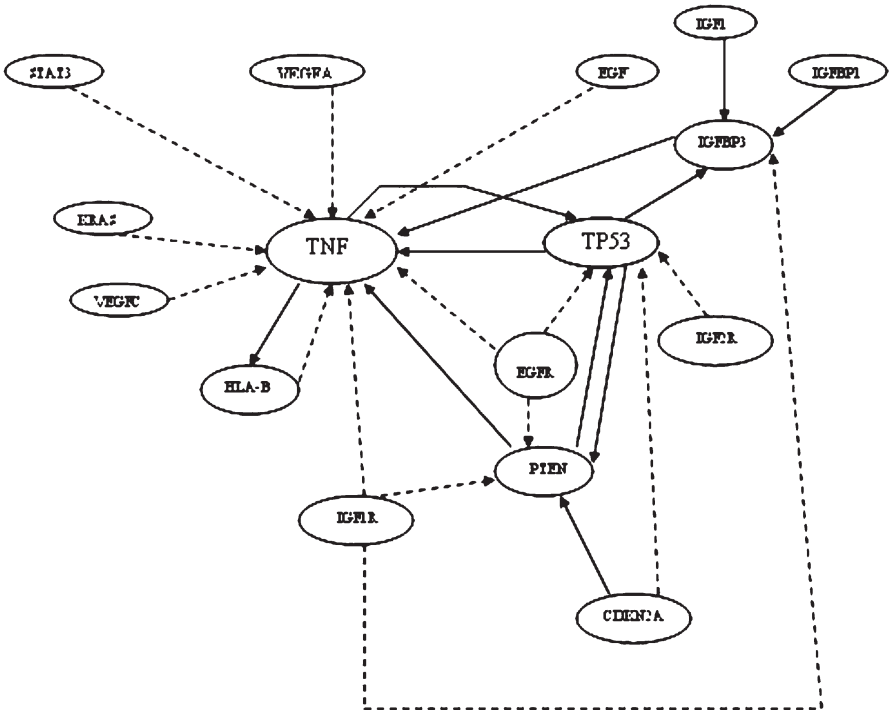

Fig. 3. AGDN for lung expression data. Continuous (dashed) arrow indicates that dependencies between genes in carcinogenic samples have been gained (lost). Here we consider $p_{ij} = 1$ or -1 only.

pathways for human sarcoma and breast cells. From this database, we have collected the information related to gene–gene interaction for leukemia, lung and colon cancers. The list of interactions for three cancers have been compared with associations generated by GDN in normal state of lung, colon, and lymphocyte cell.

For human lung expression dataset, we have identified 132 interactions among the genes in normal state using GDN. The database provides 126 such interactions for human lung. We have found 84 interactions (true positive) that are common in both the sets. Likewise, we have calculated false positive and false negative. Finally, we have calculated these parameters for all the cases except breast and sarcoma (Table 1 in Supplementary Material). In Fig. 4, it is clearly observed that our method GDN produces the highest number of true positives compared to the existing methods for all the three datasets in terms of gene–gene interactions. It is to be noted that GDN (in Fig. 4) also generates less number of false positives and false negatives compared to the existing methods for all the three datasets. In order to validate our results further, we have computed *Sensitivity* for lung expression dataset. The entire result has been shown in Table 1 in supplementary Material. *Sensitivity* is computed using the following equations

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{9}$$

From Fig. 4 (Table 1 in Supplementary Material), it is clearly observed that *Sensitivity* of GDN is much higher with respect to FCBAR, Apriori, T-Apriori, DGN, Maxconf, Pattern Fusion, TD-CLOSE, TOPKRGS for every dataset. In other words, we can say that GDN is able to find out more true positives for all data sets. Thus, we can conclude that GDN performs the best in identifying the associations.

3.2.2. Using biochemical pathways

From the aforesaid pathways, we have identified the genes (protein) involved in these pathways. Now we consider the altered associations in such a way the set of genes involved in these associations has become similar to the set of genes involved in the pathways. For a comparative study, we have applied the aforesaid existing methods on gene expression datasets to find out the informative genes. For lung cancer, we have found non-small cell lung cancer and small cell cancer pathways. A set of 409 genes are involved in these two pathways. We have compared this set of genes with these 409 genes (Table 2 in Supplementary Material). AGDN has identified 412 genes that are involved in altered associations. Here we have identified 338 genes that are common in database information and the results of AGDN. We have called these genes as *true positive* (TP) genes. Thus we have 74 genes that are in the set of 412 genes (obtained by AGDN) but not involved in the pathways. These 74 genes are considered as *false positive* (FP). Similarly, the number of *false negative* genes is 71 for AGDN. In order to validate our results further, we have computed *Sensitivity* for lung expression dataset. The entire result has been shown in Table 2 in Supplementary Material.

For human colon expression data, we have found 62 genes that are present in a colon cancer related pathway (i.e., colorectal cancer pathway). Similarly, we have found 355 genes in leukemia related pathways like chronic myeloid and acute myeloid leukemia. From Fig. 5, it is clear that AGDN have provided the best results to find out true positives with respect to the existing methods for all the three datasets. It is to be mentioned that AGDN is capable of finding out less number of false positive and false negative genes compared to the above mentioned existing methods for all the three datasets. It is to be reported that, *Sensitivity* for AGDN of each data set is better than existing methods. Thus, we can conclude that AGDN finds out more true positive genes for all the datasets compared to aforesaid existing methods.

3.2.3. Using t-test

We apply t-test on the set of influential genes produced by AGDN. Here we report three sets of genes with the levels of significance as 99.9%, 99% and 95%. For lung expression data, we have identified 188 influential genes (Table 3 in Supplementary Material). Out of them 63% (167 out of 263) influential genes result in 99.9% level of significance, and 82% (218 out of 263) and 95% (251 out of 263) correspond to 99% and 95% levels of significance. For human colon expression data, these figures are 76% (99.9% level of significance), 91% (99% level of significance) and 96% (95% level of significance). Similarly, these figures for human breast expression data are 62% (99.9% level of significance), 70% (99% level of significance) and 80% (95% level of significance). Likewise, 71% (99.9% level of significance), 77% (99% level of significance) and 84% (95% level of significance) are the figures for human lymphocyte and cell expression data. Finally, the figures for human sarcoma dataset are 69% (99.9% level of significance), 78% (99% level of significance) and 90% (95% level of significance). All these results (Fig. 6) indicate that these influential genes have changed their expression level from normal state to diseased

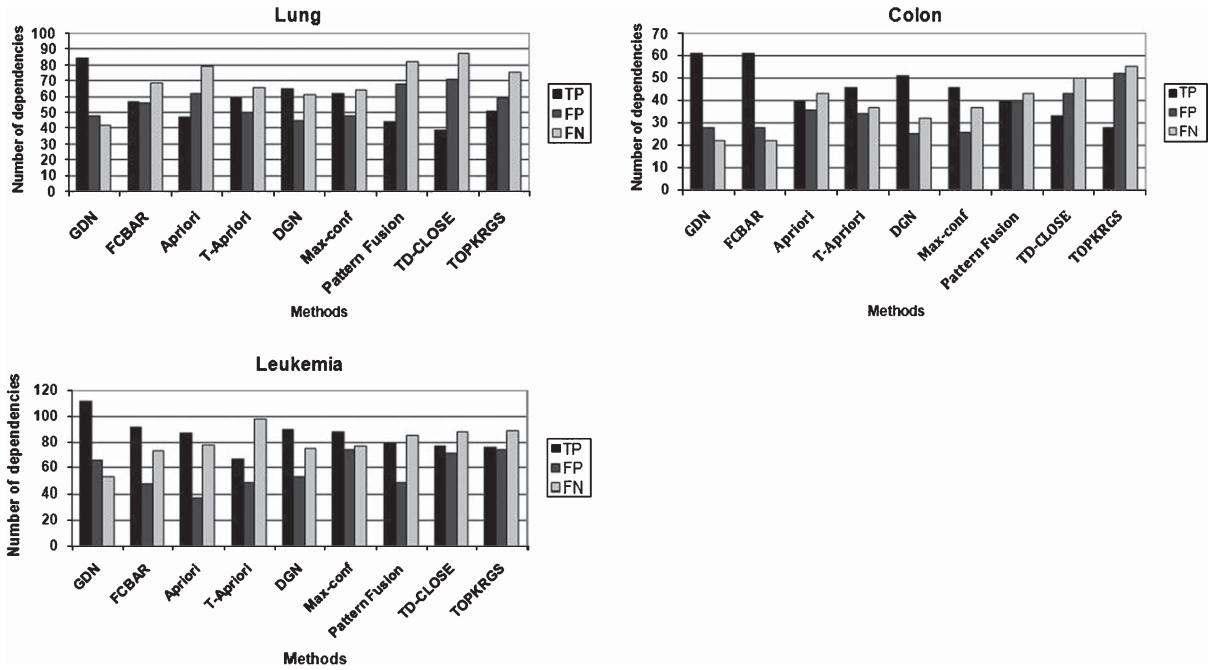

Fig. 4. Comparison among the methods in terms of gene–gene interactions. Here *TP*, *FP*, *FN* indicates true positive, false positive, false negative respectively.

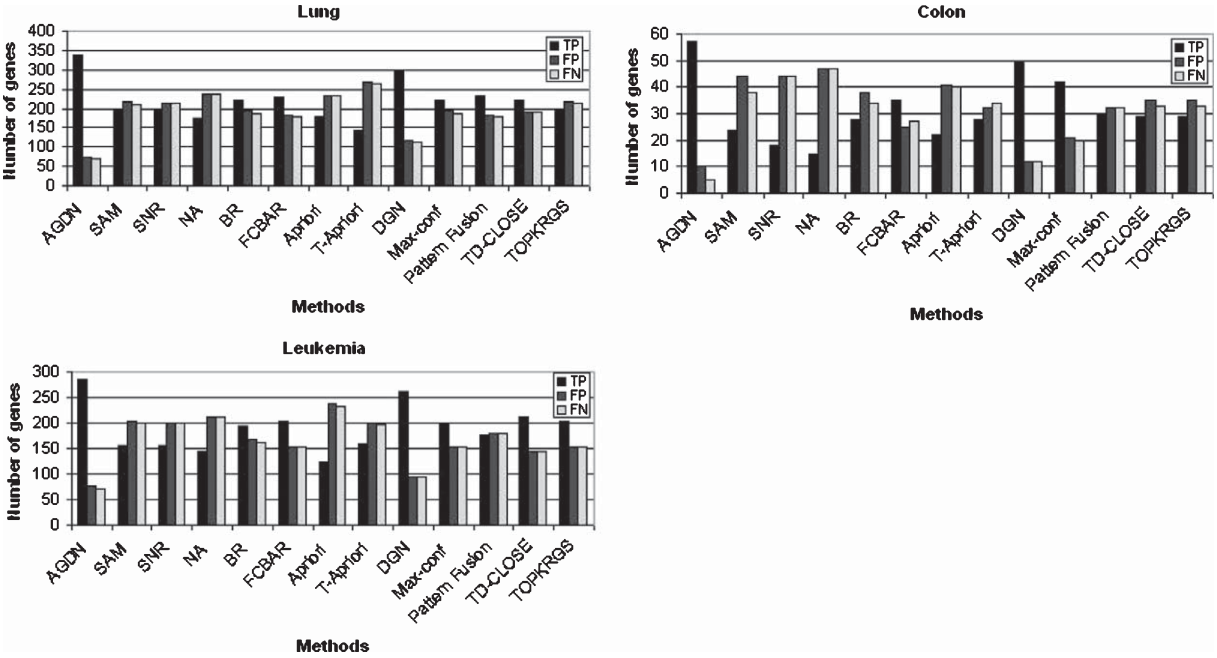

Fig. 5. Comparison among the methods in terms of biochemical pathways. Here *TP*, *FP*, *FN* indicates true positive, false positive, false negative respectively.

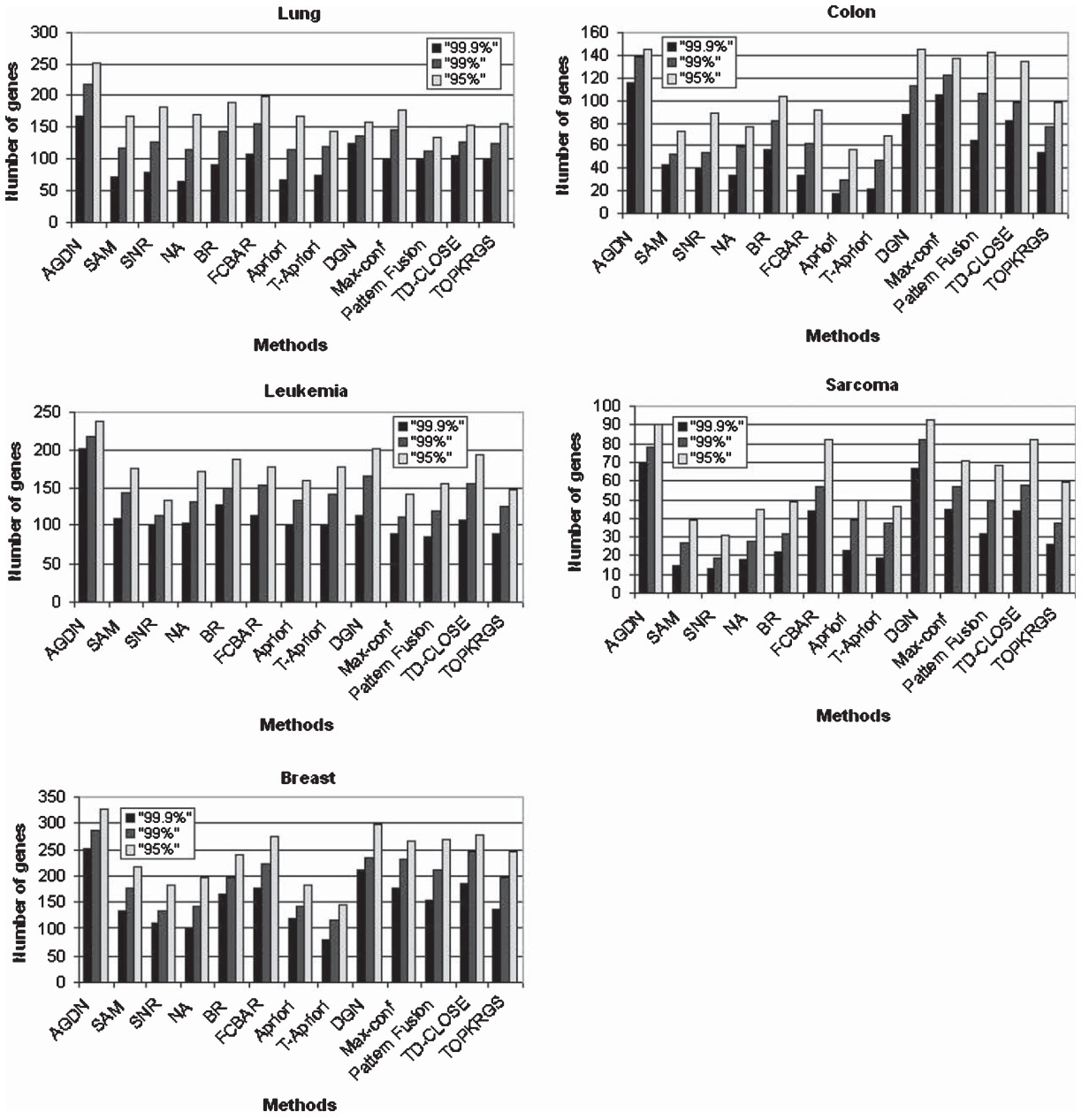

Fig. 6. Comparative study of different methods using *t*-test.

state quite significantly. From Fig. 6 (Table 3 in Supplementary Material) it is clearly observed that *AGDN* is able to find out more significant genes with respect to the above mentioned existing methods for all the five datasets. Thus, we conclude that AGDN is the best method compared to the other exsiting methods to identify the informative genes in terms of level of significance.

3.2.4. Using NCBI database

From AGDN, we have identified 263, 151, 283, 403, 99 influential genes for lung, colon, leukemia, breast and sarcoma respectively. For comparison, we have used the existing methods to identify the disease mediating genes. NCBI provides a gene database (http://www.ncbi.nlm.nih.gov/Database) where the disease mediating gene list corresponding to a specific

disease can be obtained. The list is arranged in terms of relevance of the gene. We have got different sets of genes for lung cancer, colon cancer, sarcoma, breast cancer and leukemia. In NCBI, we have found 346, 210, 329, 483, 124 genes that have responsible for lung cancer, colon cancer, leukemia, breast cancer and sarcoma respectively (Fig. 7). To validate the results differently, we compute the genes such that $c' = c$ where NCBI database results top c' genes according to the relevance and AGDN results c influential genes. Thus we made a matching between two sets of results to find out the *true positives*. In order to validate the results, we have computed *Sensitivity* for all five datasets. The entire result has been shown in Table 4 in Supplementary Material. From Fig. 7 (Table 4 in Supplementary Material) it is clearly observed that AGDN produces high number of true positives compared to the aforesaid methods. It is also to be noted that AGDN produces higher sensitivity (Table 4 in Supplementary Material) with respect to the existing methods for each data set. In other words, we can say that AGDN is capable of identifying more true positives for gene expression datasets.

3.2.5. Using p-values

Since the influential genes have been ranked based on their importance in AGDN, top ranked genes are expected to mediate the respective carcinoma, being involved in some particular biological functions. Thus the functional categories, related to these biological functions, of these genes should be enriched.

In our study, the enrichment of each GO (Gene Ontology) category [48] for each of the genes has been calculated by its p-value. For comparison with $AGDN$, we applied SAM, SNR, NA, BR, GeneSelector on five gene expression datasets. Here only functional categories with p-value $\leq 5 \times 10^{-5}$ has been considered. Figure 8 and 9 show the number of functionally enriched attributes corresponding to AGDN, SAM, SNR, NA, BR, GeneSelector for the set of important genes of all the five datasets (Tables 5 and 6 in Supplementary Material). Higher number of enriched attributes for a set of top ranked genes indicates that the resulting genes are belonging to the same functional categories. In other words, this group of genes are performing the same set of functions. This means, if one of the genes from the pool is responsible for cancer then the other genes may have a strong influence in mediating the disease.

In order to demonstrate the ability to identify cancer mediating genes correctly, we have computed the number of enriched attributes of the first 5, 10, 15, 20 important genes (influential genes for $AGDN$) for all the five datasets. From Fig. 8 (Table 5 in Supplementary Material), it is clearly observed that gene ranking resulted in by AGDN is the best compared to SAM, SNR, NA, BR and GeneSelector of all the five datasets. Similarly, we have calculated the number of enriched attributes for the last 5, 10, 15, 20 gene sets to establish the fact that how correctly the method AGDN is capable of ranking the genes according to their importance. From Fig. 9 (Table 6 in Supplementary Material), it is clearly observed that AGDN performs the best in terms of identifying less important genes compared to SAM, SNR, NA, BR and GeneSelector for all the five datasets. In Figs. 8 and 9, it is clearly viewed that AGDN provides the best results with respect to aforesaid existing methods for lung, colon, sarcoma, leukemia and breast cancer datasets.

3.2.6. Validation based on some earlier investigations

The present method finds some genes along with the dependencies among them, which have changed from normal to carcinogenic samples. We could not find any article in literature, which deals with similar investigation. That is why, we have tried to validate our results based on gene regulation. Tables 7-11 (in Supplementary Material) show some of the dependencies that have changed from normal to carcinogenic state, along with some articles that support the fact on gene regulation in normal and carcinogenic states. These tables include a column containing some references. In order to restrict the size of the article, we have kept these tables as well as references in Supplementary Material. We now mention some of these dependencies that have changed from normal to diseased state.

For lung expression dataset (Table 7 in Supplementary Material), gene like HBB have shown stronger dependency on HBA1 in carcinogenic state. It is also noticed that genes like TP53, IGF1, IGFBP1 have strong influence in regulating gene IGFBP3 in cancer state whereas there is no dependency among them in normal state. In other words, TP53, IGF1, IGFBP1, IGF1R may regulate the expression value of IGFBP3 in tumor samples. It has been found that gene TP53 also regulates TNF, and gene TNF regulates genes TP53, HLA-B and PTEN in lung adenocarcinoma samples. In this way, we have identified a set of genes that have shown their strong dependencies in cancer state of lung expression data. Likewise, genes KRAS, EGFR, VEGFA regulate the gene TNF in normal state, whereas in carcinogenic state the expression levels of

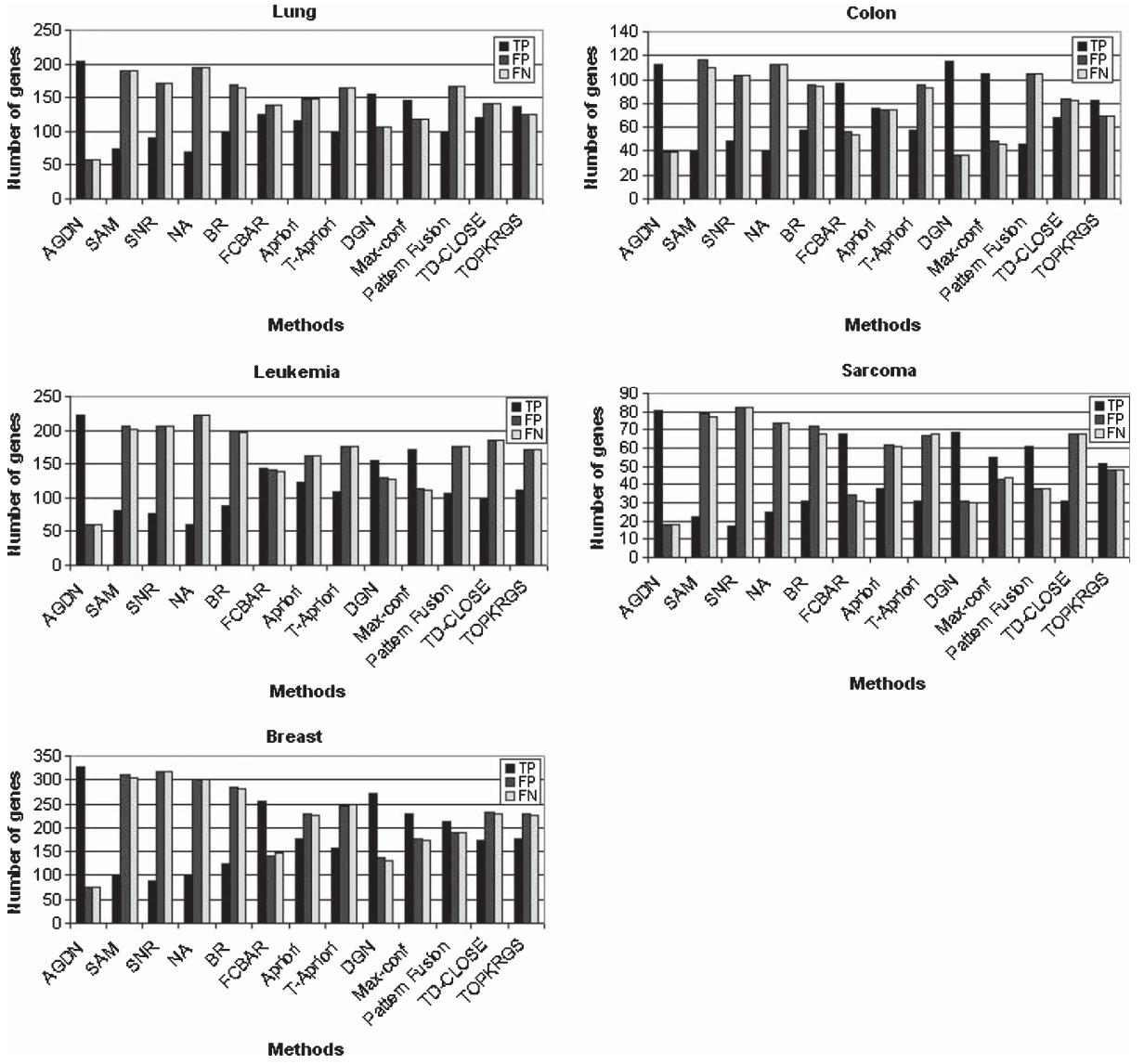

Fig. 7. Comparative study with other methods using NCBI database.

these genes may be almost independent of the expression level of TNF.

Similarly, for colon expression data (Table 8 in Supplementary Material), we have found that genes HLA-B, PTX3 have shown stronger dependency on TNF in cancer state. It is also reported that gene like TP53 has strong influence in regulating the genes IGFBP3, TNF and MBP in malignant tumor state, whereas there is no dependency among them in normal state. It also reflects that gene TP53 may regulate the expression of the genes IGFBP3, TNF and MBP in colon cancer. Some earlier investigations support this behavior. Likewise, we have identified a set of genes that have shown strong association in terms of dependency in normal state. For example, EGFR, IGF1R, STAT3, MAPK8 and IL6 strongly influence the expression of gene TNF in normal state, whereas no dependency has been found among them in malignant state. Gene PDGFRA has shown strong dependency on genes like E2F1 and BRAF in normal state whereas there is no such dependency in cancer state. However, there is no information in literature to our knowledge about these genes. This result suggests that the aforesaid genes may have impact on human colon carcinoma.

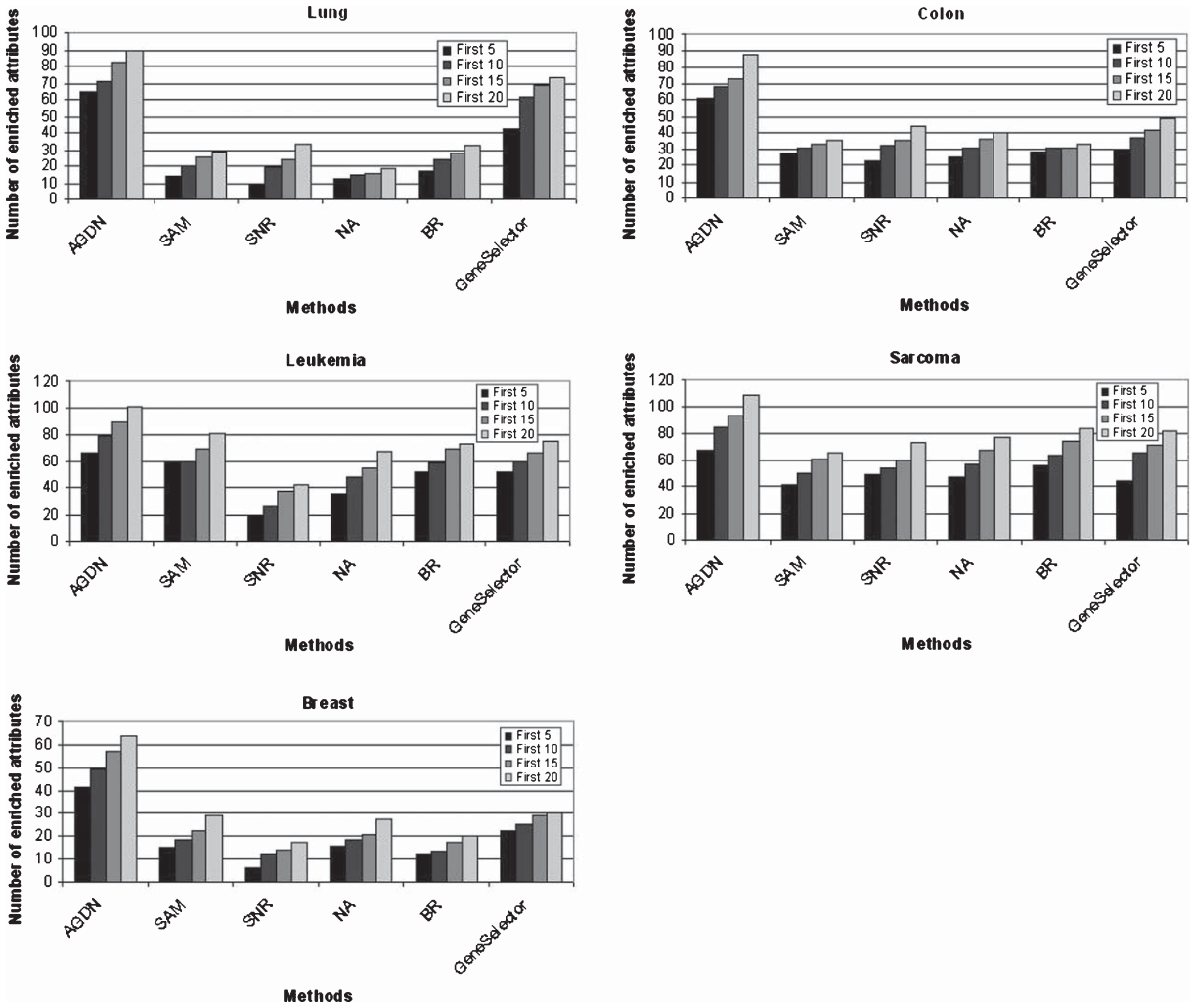

Fig. 8. Comparative results on number of enriched attributes for various sets (first 5 to first 20 gene set) of influential genes corresponding to different methods.

Likewise for human sarcoma cell expression data (Table 9 in Supplementary Material), we have identified that gene BAX has shown strong dependency, in terms of regulation, on BCR, TP53 and TNF in diseased state, and on STAT3 and VEGFA in normal state. On the other hand, genes like CDKN2A, IGFBP3, BCR, PTEN have shown strong dependency on TP53 in carcinogenic state whereas there is no such association among them in normal state. Gene CDKN2A has strong influence in regulating gene TP53 in normal state. Gene TNF influences of the expression the genes like BAX, TP53, HLA-B, PTEN in cancer state.

For human lymphocyte cell expression data (Table 10 in Supplementary Material), IGF1R, VEGFA, BRCA1, MAPK3 have strong influence in regulating the gene IGF1 in normal state, whereas there is no such dependency among them in diseased state. In other words, IGF1R, VEGFA, BRCA1, MAPK3 may regulate the expression value of IGF1 in normal state. It is also reported that gene BRCA1 has shown strong dependency on genes like TP53, PTEN in diseased state, and STAT3 and CDKN2A in normal state. Gene BCL2 influences strongly the expression of the genes BAX and BCR in cancer state. We have also identified that genes like BAX, IGFBP3, BCR, CDKN2A, BRCA1, KRAS have shown their strong dependency on TP53 in diseased sate whereas there is no dependency among them in normal state.

Regarding human breast expression data (Table 11 in Supplementary Material), we have identified the gene

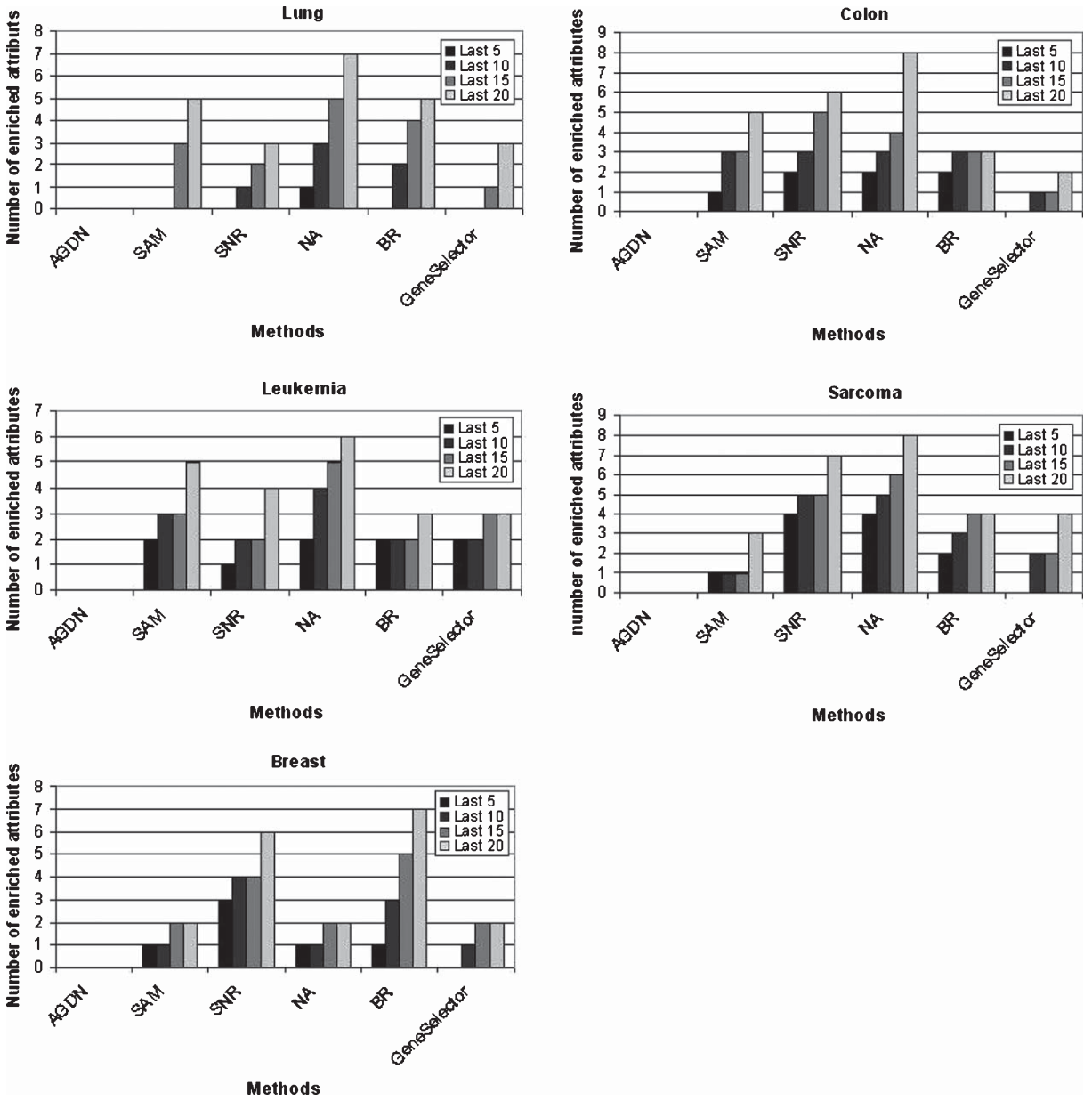

Fig. 9. Comparative results on number of enriched attributes for various sets (last 5 to last 20 gene set) of influential genes corresponding to different methods.

BRCA1 that has shown strong association in terms of dependency on genes like PTEN, TP53 in malignant tumor state. But in normal state, there is no dependency among them. Similarly, genes like STAT3 and CDKN2A have strong influence in regulating the gene BRCA1 in normal state. In other words, STAT3 and CDKN2A may regulate the expression of BRCA1 in normal state. On the other hand, NPM1 has shown strong dependency on gene BRCA1 in normal state. Gene KRAS has shown strong dependency on genes like TP53, PTEN in cancer state, whereas in normal state CDKN2A, BCL2, ERBB2, BRAF have strong influence on gene KRAS. It has been found that genes like HNF1A, HNF4A, CREBBP have strong influence in regulating the gene H3F3A in normal state, whereas in diseased state there is no such dependency. But there

is no information in literature to our knowledge about these genes. This result suggests that the aforesaid genes may have impact on human breast cancer.

4. Conclusion

In this article, we have developed a methodology that shows how the dependencies (interactions) among the genes have changed from normal state to diseased state. The algorithm has identified the dependencies among the genes, which have altered quite significantly from normal state to diseased state. The methodology involves measuring information content of individual genes using fuzzy entropy, dependencies of a pair of genes in both normal and diseased states using conditional fuzzy entropy. Finally, the dependencies that have altered from normal to carcinogenic state have been identified. The interactions among the genes for either state (i.e., normal or disease) of gene expression data is represented by the gene dependency network (GDN). The altered dependencies among the genes have been represented using network, called altered gene dependency network ($AGDN$), in which each node represents a gene and a directed edge signifies altered dependency between a pair of nodes (genes). In addition, we proposed a technique that identifies the influential genes from AGDN.

In this way, we have identified the responsible genes as well as the altered interactions among them. We have applied the algorithm on five cancer data sets (lung, colon, sarcoma, leukemia and breast). As a result, we have identified the altered gene dependency network for each of the data sets. We could not find any article in literature, which deals with similar investigation. So we have tried to validate our results based on gene regulation. In this context, we have used gene–gene interaction, biochemical pathways, t-test, p-value, NCBI database to validate the results. We have used some existing association rule mining algorithms like FCBAR, Apriori, T-Apriori, DGN, Max-conf, Pattern Fusion, TD-CLOSE, TOPKRGS for a comparison with GDN in terms of gene–gene interactions. Likewise, to compare with $AGDN$, we have implemented aforesaid methods in terms of identifying the important genes mediating cancers. From all the results, we have found that the present method has been able to correctly identify many true positive genes as well as gene–gene interactions from gene expression datasets. As a consequence, we can say that these set of identified genes along with their altered interactions, have a significant role of mediating the disease. Hence, these results may facilitate the biologists and researchers carrying out the biochemical analysis to do further study on gene regulatory networks and how the entire network structure changes from normal state to diseased state.

References

[1] R. Tupler, G. Perini, M.A. Pellegrino and M.R. Green, Profound misregulation of muscle-specific gene expression in facioscapulohumeral muscular dystrophy, *Proc Natl Acad Sci* **96** (1999), 12650–12654.

[2] D.H. Ly, D.J. Lockhart, R.A. Lerner and P.G. Schultz, Mitotic misregulation and human aging, *Science* **287** (2000), 2486–2492.

[3] J.L. Riechmann, J. Heard, G. Martin, L. Reuber, C. Jiang, J. Keddie, L. Adam, O. Pineda, O.J. Ratcliffe and R.R. Samaha, Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes, *Science* **290** (2000), 2105–2110.

[4] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael and R. Ohnhauser, The TRANSFAC system on gene expression regulation, *Nucleic Acids Res* **29** (2001), 281–283.

[5] E. Roulet, I. Fisch, T. Junier, P. Bucher and N. Mermod, Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA, *In Silico Biol* **1** (1998), 21–28.

[6] W. Krivan and W.W. Wasserman, A predictive model for regulatory sequences directing liver-specific transcription, *Genome Res* **11** (2001), 1559–1566.

[7] N. Grabe, AliBaba2: Context specific identification of transcription factor binding sites, *In Silico Biol* **2** (2002), S1–S15.

[8] M.S. Halfon, Y. Grad, G.M. Church and A.M. Michelson, Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model, *Genome Res* **12** (2002), 1019–1028.

[9] O. Schuldiner, C. Yanover and N. Benvenisty, Computer analysis of the entire budding yeast genome for putative targets of the GCN4 transcription factor, *Curr Genet* **33** (1998), 16–20.

[10] K. Tan, G. Moreno-Hagelsieb, J. Collado-Vides and G.D. Stormo, A comparative genomics approach to prediction of new members of regulons, *Genome Res* **11** (2001), 566–584.

[11] S. Kim, S. Imoto and S. Miyano, Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, *Biosystems* **75** (2004), 57–65.

[12] I. Nachman, A. Regev, N. Friedman, Inferring quantitative models of regulatory networks from expression data, *Bioinformatics* **4** (2004), i248–i256.

[13] A. Fujita, J.R. Sato, H.M. Garay-Malpartida, P.A. Morettin, M.C. Sogayar and C.E. Ferreira, Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method, *Bioinformatics* **23** (2007), 1623–1630.

[14] O. Hirose, R. Yoshida, S. Imoto, R. Yamaguchi, T. Higuchi, S.D. Charnock-Jones, C. Print and S. Miyano, Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models, module finder on gene expression profiles, *Bioinformatics* **24** (2008), 932–942.

[15] R. Yamaguchi, R. Yoshida, S. Imoto, T. Higuchi and S. Miyano, Finding modulebased gene networks in time-course

gene expression data with state space models, *IEEE Signal Processing Magazine* **24** (2007), 37–46.
[16] J.D. Storey and R. Tibshirani, Statistical significance for genomwide studies, *Proc Natl Acad Sci* **100** (2003), 9440–9445.
[17] A.S. Butte, I.S. Kohane, Relevance networks: A first step toward finding genetic regulatory networks within microarray data. *The Analysis of Gene Expression Data*, Springer, 2003, 428–446.
[18] J. Schafer and K. Strimmer, An empirical bayes approach to inferring largescale gene association networks, *Bioinformatics* **21** (2005), 754–764.
[19] H.J. Cordell, Detecting gene–gene interactions that underlie human diseases, *Nat Rev Genet* **10** (2009), 392Ü–402.
[20] X. Shang, Q. Zhao and Z. Li, Mining high-correlation association rules for inferring gene regulation networks, *Bioinformatics Research and Development* **5691** (2009), 244–255.
[21] H. Wang and Y. Lee, Gene network prediction from microarray data by association rule and dynamic bayesian network, *Computational Science and Its Applications* **3482** (2005), 309–317.
[22] C. Creighton and S. Hanash, Mining gene expression databases for association rules, *Bioinformatics* **19** (2003), 79Ű–86.
[23] H. Nam, K. Lee, D. Lee, Identification of temporal association rules from time-series microarray data sets, *BMC Bioinformatics* **10** (2009), 1–9.
[24] G. Pandey, S. Chawla, S. Poon, B. Arunasalam and J.G. Davis, Association rules network: Definition and applications, *Journal Statistical Analysis and Data Mining archive* **1** (2009), 260–279.
[25] D.H. Yang, J.H. Kang, Y.B. Park, Y.J. Park, H.S. Oh and S.B. Kim, Association rule mining and network analysis in oriental medicine, *PLOS ONE* **8** (2013), e59241.
[26] A. Deluca and S. Termini, A definition of non-probabilistic entropy in the setting of fuzzy set theory, *Information and Control* **20** (1972), 301–312.
[27] R. Alves, D.S. Rodriguez-Baena and J.S. Aguilar-Ruiz, Gene association analysis: A survey of frequent pattern mining from gene expression data, *Briefings in Bioinformatics* **2** (2009), 210–224.
[28] R. Sheibani and A. Ebrahimzadeh, An algorithm for mining fuzzy association rules, *Proceedings of the International MultiConference of Engineers and Computer Scientists 1* (2008).
[29] R. Agrawal, T. Imielinski and A. Swami, Database mining: A performance perspective, *IEEE Transactions on Knowledge and Data Engineering* **5** (1993), 914–925.
[30] M.S. Khan, M. Muyeba, F. Coenen and D. Reid, Mining fuzzy association rules from composite items, (2009), 1–31.
[31] K. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal and A. Barababasi, The human disease network. *PNAS* **104** (2007), 8685Ű–8690.
[32] T. McIntosh and S. Chawla, High confidence rule mining for microarray analysis, *IEEE/ACM Trans Comput Biol Bioinform* **4** (2007), 611Ű–623.
[33] F. Zhu, X. Yan and J. Han, Mining colossal frequent patterns by core pattern fusion, *International Conference on Data Engineering* (2007), 706–715.
[34] H. Liu, J. Han and D. Xin, Top-down mining of interesting patterns from very high dimensional data, *International Conference on Data Engineering* (2006).
[35] G. Cong, K. Tan and A.K. Tung, Mining top-k covering rule groups for gene expression data, *International Conference on Management of Data*, Baltimore, Maryland, USA, 2005, 670Ű–681.
[36] L. Goh, Q. Song and N. Kasabov, A novel feature selection method to improve classification of gene expression data, *Asia-Pacific Bioinformatics Conference* **29** (2004).
[37] V.G. Tusher, R. Tibshirani and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, *Proc Natl Acad Sci USA* **98** (2001).
[38] T.R. Golub, T.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, J.R. Downing, M.A. Caliguri, C.D. Bloomeld and E.S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* **286** (1999), 531–537.
[39] S.K. Shevade and S.S. Keerthi, A simple and efficient algorithm for gene selection using sparse logistic regression, *Bioinformatics* **19** (2003), 2246–2253.
[40] G.C. Cawley and N.L.C. Talbot, Gene selection in cancer classification using sparse logistic regression with Bayesian regularization, *Bioinformatics* **22** (2006), 2348–2355.
[41] A. Boulesteix and M. Slawski, Stability and aggregation of ranked gene lists, *Briefings in Bioinformatics* **10** (2009), 556–568.
[42] J. Quinlan, Programs for machine learning. Morgan Kaufmann, 1993.
[43] G.D. Beer et al., Gene-expression profilespredict survival of patients with lung adenocarcinoma, *Nature Medicine* **8** (2002), 816–823.
[44] B.H. Mecham, G.T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D.Z. Wetmore, T.J. Mariani, I.S. Kohane and Z. Szallasil, Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements, *Nucleic Acids Res* **32** (2004), e74.
[45] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack and A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc Natl Acad Sci USA* **96** (1999), 6745–6750.
[46] N.C. Gutierrez, E.M. Ocio, J. delas Rivas, P. Maiso, M. Delgado, E. Ferminan, M.J. Arcos, M.L. Sanchez, J.M. Hernandez and J.F.S. Miguel, Gene expression profiling of B lymphocytes and plasma cells from WaldenstromŠs macroglobulinemia: Comparison with expression patterns of the same cell counterparts from chronic lymphocytic leukemia, multiple myeloma and normal individuals, *Leukemia* **21** (2007), 541–549.
[47] K.Y. Detwiller, N.T. Fernando, N.H. Segal, S.W. Ryeom, P.A. D'Amore and S.S. Yoon, Analysis of hypoxia-related gene expression in sarcomas and effect of hypoxia on RNA interference of vascular endothelial cell growth factor A, *Cancer Res* **65** (2005), 5881–5889.
[48] G.F. Berriz, O.D. King, B. Bryant, C. Sander and F.P. Roth, Characterizing gene sets with funcassociate, *Data Mining and Knowledge Discovery* **19** (2003) 2502Ü–2504.