Neurocomputing 133 (2014) 122–140

Contents lists available at ScienceDirect

Neurocomputing

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

Selection of genes mediating certain cancers, using a neuro-fuzzy approach

Anupam Ghosh [a], Bibhas Chandra Dhara [b], Rajat K. De [c,*]

[a] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata 700152, India
[b] Department of Information Technology, Jadavpur University, Kolkata, India
[c] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 9 May 2013
Received in revised form
24 September 2013
Accepted 4 November 2013
Communicated by L. Kurgan
Available online 23 January 2014

Keywords:
Artificial neural networks
Fuzzy membership functions
GO attributes

ABSTRACT

In this article, we propose a methodology for selecting genes that may have a role in mediating a disease in general and certain cancers in particular. The methodology, first of all, groups an entire set of genes. Then the important group is determined using two neuro-fuzzy models. Finally, individual genes from the most important group are evaluated in terms of their importance in mediating a cancer, and important genes are selected. A method for multiplying existing data is also proposed to create a data rich environment in which neuro-fuzzy models are effective. The effectiveness of the proposed methodology is demonstrated using five microarray gene expression data sets dealing with human lung, colon, sarcoma, breast and leukemia. Moreover, we have made an extensive comparative analysis with 22 existing methods using biochemical pathways, *p*-value, *t*-test, *F*-test, sensitivity, expression profile plots, *pi*-GSEA, Fisher-score, KOGS, SPEC, *W*-test and *BWS*, for identifying biologically and statistically relevant gene sets. It has been found that the proposed methodology has been able to select genes that are more biologically significant in mediating certain cancers than those obtained by the others.

1. Introduction

In the present context, the selection of disease mediating genes is based on their expression patterns in the normal as well as in the diseased sample. A genome wide expression pattern for a particular tissue is generated through microarray experiments. These experiments involve uncertainty in the preparation of microarray chip as well as sample collection and hybridization experiments. Moreover, a microarray experiment provides an indication of average expression values of genes but not their expression values. Thus, uncertainty is involved in this kind of data.

Gene selection refers to the task of selecting some important genes that best explain experimental variations [1]. It is much cheaper to focus on a small number of important genes that have different expression patterns in diseased samples. Therefore, using effective gene selection methods, a small list of highly important genes can be isolated from the whole genome, which have a direct/indirect role in causing diseases [2,3]. We call these important genes disease mediating genes. From a gene expression point of view, disease mediating genes refer to those that have changed their behavior from normal conditions in which symptoms of the disease under consideration are not present. Then, these genes can be explored to identify the cause of the disease and thereby may aid rational drug design.

Fuzzy set theory enables one to deal with uncertainties arising from deficiencies like inexactness, vagueness, and uncertainty in information in an efficient manner. Thus, fuzzy set theory forms a tool for handling these uncertainties. Artificial neural networks (ANN), having the capability of fault tolerance, adaptivity, generalization, and scope for massive parallelism, are widely used in dealing with learning and optimization tasks. Moreover, artificial neural networks (ANNs) are used to solve problems which are intractable in nature. The neuro-fuzzy models, a hybridization of the concepts of fuzzy sets and artificial neural networks, can tackle such intractability and uncertainty, in an efficient way [4]. Thus, we have used neuro-fuzzy models here.

The present article is an attempt in this regard and provides a new methodology involving neuro-fuzzy models for identifying genes mediating a disease in general and certain cancers in particular. The methodology involves grouping of genes using correlation coefficient, followed by selecting the most important group using the neuro-fuzzy models. We call these models neuro-fuzzy Model-1 (NFM-1) and neuro-fuzzy Model-2 (NFM-2). Then the most important genes from the selected most important group are identified using neuro-fuzzy models again. It is to be

* Corresponding author.
E-mail addresses: anupam.ghosh@rediffmail.com (A. Ghosh), bibhas@it.jusl.ac.in (B. Chandra Dhara), rajat@isical.ac.in (R.K. De).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.11.023

mentioned here that these neuro-fuzzy models have been developed in [5–7] for the purpose of feature selection. The methodology is applicable in a data rich environment, i.e., if the number of samples is quite large compared to the dimension of each sample. However, in the present problem, the number of microarray measurements (samples) is quite low compared to the number of genes (dimension). In order to overcome this problem, we have proposed a way of generating more data from the given microarray gene expression measurements.

The effectiveness of the proposed methodology, along with its superior performance over several other methods, has been demonstrated using five microarray gene expression data sets dealing with cancers related to human lung, colon, breast, sarcoma and leukemia. An initial set of results using lung expression data has been published in [8]. The existing methods, with which the results have been compared, are Bayesian regularization (BR) model [9,10], significance analysis of microarray (SAM) [11], signal-to-noise ratio (SNR) [12], neighborhood analysis (NA) [13], support vector machine (SVM) [14,1], Gaussian mixture model (GMM) [15], hidden Markov model (HMM) [16], constructive approach for feature selection (CAFS) [17], entropy based penalized logistic regression (Entropy-PLR) [18], minimum sum of square of the correlation (MSC) and maximum value of square of the correlation (MMC) with Naive Bayes classifier (NBC) and nearest mean scaled classifier (NMSC) (i.e., NBC-MSC, NBC-MMC, NMSC-MSC and NMSC-MMC) [19], leave-one-out calculation sequential forward selection (LOOCSFS) [20], gradient based leave-one-out gene selection (GLGS) [20], least square (LS) bound measure with sequential forward selection (SFS) and sequential floating forward selection (SFFS) [21], a method in the R package (VarSelRF) [22], and partial least squares (PLS) SlimPLS [23]. The performance comparison has been made using *t*-test, *F*-test and *p*-value (in terms of the number of enriched attributes or GO (gene ontology) attributes). In addition, we have used biological and statistical measurements like *pi*-GSEA [24], Fisher-score [25], KOGS [26], SPEC [27], *W*-test [28,30], *BWS* [29] for identifying the biologically and statistically relevant gene set.

2. Some existing methods

In the present study, we have proposed neuro-fuzzy models for identification of cancer mediating genes. We have made a survey on existing gene selection methods for comparative analysis. Among them, we have found some gene selection methods that are using SAM, SNR, BR, NA, SVM, GMM, HMM, CAFS, Entropy-PLR, NBC-MSC, NBC-MMC, NMSC-MSC, NMSC-MMC, LOOCSFS, GLGS, SFS-LSBOUND and SFFS-LSBOUND, VarSelRF, and SlimPLS. Significance analysis of microarray (SAM) [11] identifies genes with statistically significant changes in expression values by using a set of gene-specific *t* tests. Each gene is assigned a score on the basis of its change in the gene expression value. Genes with scores greater than a threshold value are deemed potentially significant. The percentage of such genes identified by chance is the false discovery rate (FDR). In order to estimate FDR, nonsense genes are identified by analyzing permutations of the measurements. The threshold value can be adjusted to identify smaller or larger sets of genes, and FDRs are calculated for each set. The disadvantage of SAM lies in the permutation stage where all the genes are put into one group for evaluation. This requires an expensive computation. Moreover, it probably confuses the analysis because of the noise in gene expression data.

The method based on the signal-to-noise ratio (SNR) [12] is applied to rank the correlated genes according to their discriminative power. The method starts with the evaluation of a single gene, and iteratively searches for other genes based on some statistical criteria. The genes with high SNR scores are chosen as the important ones. A limitation of this method is that many genes with very low correlation coefficient values are removed by the ranking criterion, because the correlation coefficient of genes is only measured by one gene to others. However, it is very likely that some of these abandoned genes may be useful when they are combined for measuring the correlation values. SNR measurement is affected by the size of the variables. When there are more variables, the mean and variance of the remaining variables of other classes are dependent on the data dispersion and the number of variables, which affects SNR ranking of the significant variables due to the general increase in noise in the data. If the number of variables can be reduced significantly, the method is more capable of detecting and ranking a smaller number of significant variables.

Neighborhood analysis (NA) [13] is a method for clustering multivariate data into distinct classes based on a given distance metric over the data. Functionally, it serves the same purposes as the K-nearest neighbor algorithm. Golub [13] applied NA to identify predictor classes defined on the different responses to therapy. The main disadvantage of the method is that it cannot detect any significant correlation. Although this failure may be due to the limited number of genes included in the study [13], it is also possible that the "response" phenotype is too complex to be associated with a cluster of genes, and a more elaborate relationship may exist between response to therapy and gene expression.

Shevade and Keerthi [9] have developed a gene selection algorithm based on sparse logistic regression (SLogReg) and provide a simple but efficient training procedure. The degree of sparsity is determined by the value of a regularization parameter, which must be carefully tuned in order to get an optimal performance. This normally involves a model selection stage, based on a computationally intensive search for the minimization of the cross-validation error. In [10] a simple Bayesian approach has been incorporated to eliminate this regularization parameter.

SVM [14] is a machine learning methodology which separates two classes by maximizing the margin between them. A novel type of regularization in support vector machines (SVMs) is used to identify important genes for cancer classification [14]. The standard SVM and Lasso (L1) SVM are often considered using quadratic programming and linear programming methods. An iterative algorithm is used to solve the smoothly clipped absolute deviation (SCAD) SVM efficiently. It is reported that the SCAD-SVM selects a smaller and a more stable (with smaller standard errors) number of genes than the L1-SVM in almost all the cases [14]. Recursive feature elimination (RFE) SVM is another algorithm of gene selection using the weight magnitude as the ranking criterion [1]. The SVM-RFE method ranks all the genes according to some scoring function, and eliminates one or more genes with the lowest scores. This process is repeated until the highest classification accuracy is achieved.

A Gaussian mixture model (GMM) is based on a parametric probability density function that is represented as a weighted sum of Gaussian component densities [15,16]. Since GMM has been used for parameter selection, we have considered it for our comparison. In our study, we have implemented it on microarray gene expression data for gene selection. Like GMM, we have implemented HMM on microarray gene expression data for identification of genes. Generalized HMMs provide an intuitive framework for representing genes with their various functional features, and efficient algorithms can be built to use such models to recognize genes [32].

A constructive approach for feature selection (CAFS) [17] is based on the concept of the wrapper approach and sequential search strategy. As a learning model, CAFS employs a typical three layered feed-forward neural network for selecting genes. In another

investigation, an entropy based gene selection method has been developed where penalized logistic regression (Entropy-PLR) is used to estimate the probability of genes [18]. A lagging prediction peephole optimization (LPPO) algorithm is used to choose an optimal gene set [19]. It is also to be noted that the minimum sum of square of the correlation (MSC) and maximum value of square of the correlation (MMC) are also combined with Naive Bayes classifier (NBC) and nearest mean scaled classifier (NMSC) for gene selection (i.e., NBC-MSC, NBC-MMC, NMSC-MSC and NMSC-MMC) [19]. In an investigation, a gene selection method, called leave-one-out calculation sequential forward selection (LOOCSFS) algorithm, has been implemented by combining the LOOC measure with the sequential forward selection scheme [20]. Further, a novel gene selection algorithm, the gradient based leave-one-out gene selection (GLGS) algorithm, has been developed in [20]. In another study, least square (LS) bound measure has been used to evaluate the criterion for gene selection [21]. The LS bound measure can be considered as a hybrid of filter and wrapper methods [21]. The LS bound measure has been combined with search algorithms, like the sequential forward selection (SFS), for the purpose of gene selection. The SFS algorithm is a simple greedy heuristic search algorithm (i.e., SFS-LSBOUND) [21]. For better performance, other complex search algorithms, such as sequential floating forward selection (SFFS), have been used but at the cost of increasing the computational complexity (i.e., SFFS-LSBOUND) [21].

A new method in the R package (VarSelRF), which uses the notion of random forest, has been developed for gene selection in classification problems that use random forest [22]. The main advantage of this method is that it returns very small sets of genes that retain a high predictive accuracy, and is competitive with existing methods of gene selection [22]. In another study, a novel gene selection technique has been developed based on the partial least squares (PLS) algorithm, and is called SlimPLS [23]. PLS aims at obtaining a low dimensional approximation of a matrix that is 'as close as possible' to a given vector.

3. Proposed methodology

Here we describe the proposed methodology for gene selection. Since the number of genes (number of dimension) is very large compared to the number of measurements (samples), we have grouped the genes based on the correlation coefficient. Then the groups are evaluated using NFM-1 [6] or NFM-2 [7,5], and the most important group is selected. Finally, important genes are selected from the most important group using the two neuro-fuzzy models again. The entire methodology is depicted in Fig. 1.

3.1. Grouping of genes

Let us consider a set $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ of n genes for each of which the first p expression values in normal samples and the next q expression values in diseased samples are given. We now compute the correlation coefficient between pairs of these genes based on their expression values in normal samples. Thus, the correlation coefficient r_{ij} between ith and jth genes is given by

$$r_{ij}=\frac{\sum_{l=1}^{p}(x_{il}-m_i)\times(x_{jl}-m_j)}{(\sum_{l=1}^{p}(x_{il}-m_i)^2)^{1/2}\times(\sum_{l=1}^{p}(x_{jl}-m_j)^2)^{1/2}} \quad (1)$$

here m_i and m_j are the mean of expression values of ith and jth genes, respectively, over normal samples. The correlation coefficient assumes values in the interval $[-1,1]$. When $r_{ij}=-1$ $(+1)$, there is a strong negative (positive) correlation between ith and jth genes. Genes with high positive correlation values are placed into the same group. The main idea of grouping is as follows. If a gene has a strong positive correlation with another gene, then the expression patterns of these two genes are similar. In that case, we may consider one of them as a representative gene and ignore the other.

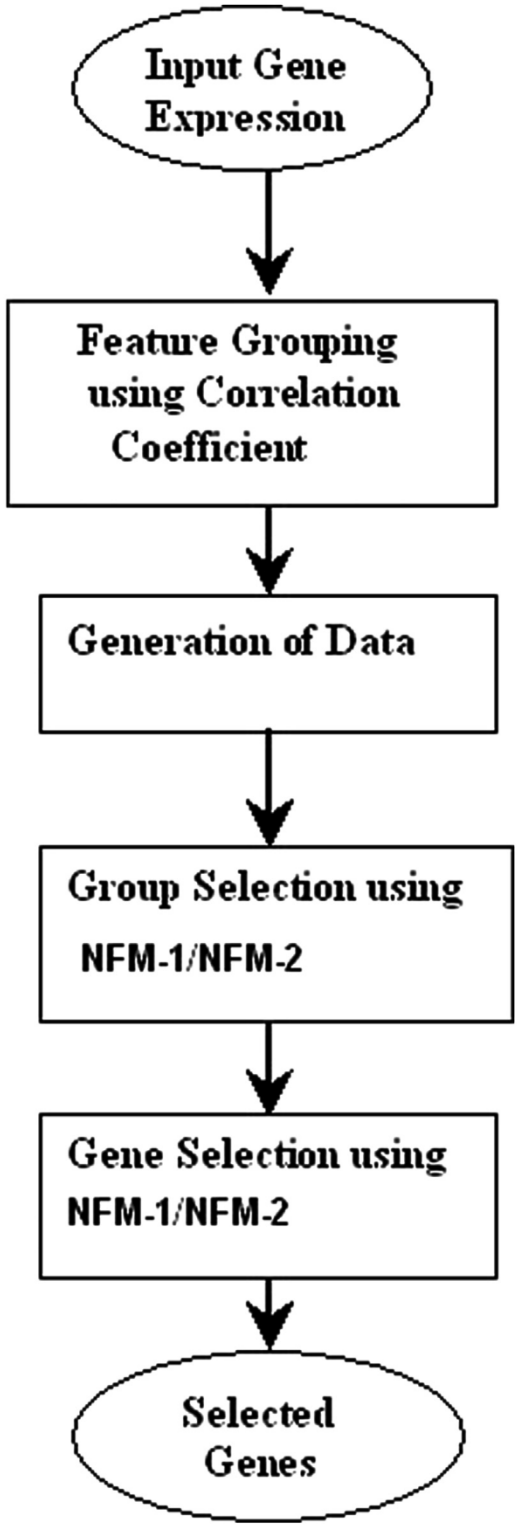

Fig. 1. Neuro-fuzzy model for gene selection.

Similarly, for diseased samples, the correlation coefficient r'_{ij} between ith and jth genes is given by

$$r'_{ij}=\frac{\sum_{l'=1}^{q}(x'_{il'}-m'_i)\times(x_{jl'}-m'_j)}{(\sum_{l'=1}^{q}(x'_{il'}-m'_i)^2)^{1/2}\times(\sum_{l'=1}^{q}(x'_{jl'}-m'_j)^2)^{1/2}} \quad (2)$$

The groups of genes are identified in such a way that the genes in the same groups are strongly positively correlated. In order to do this, r_{ij} (Eq. (1)) is computed for each pair of genes. If $r_{ij} \geq 0.75$, then these genes are placed in the same group. Here we have used correlation coefficient to narrow down the search space by finding genes of a similar behavior in terms of similar expression patterns. This helps us to identify the set of responsible genes mediating certain cancers. The choice of 0.75 as a threshold value has been done through extensive experimentation for which the distances among the cluster centers have become maximum. In this way, the first group of genes is obtained.

In the same way, the second group is obtained from the remaining genes. This step is continued till all the genes are placed in one of the groups. It is to be mentioned here that some singleton groups may also be formed by this process. Thus, we have a few groups containing various genes. It is to be mentioned here that one may choose other high values (≤ 1) instead of 0.75 as the threshold. Initially, we want to form the group of genes with a similar behavior using correlation coefficient value. The correlation value close to $+1$ between two genes indicates that the behavior of the genes is almost similar in nature, i.e., they belong to the same group. On the other hand, if their correlation value is close to -1 then their behavior is opposite in nature, i.e., negatively correlated. This is why we did not consider the negative value of correlation among the genes.

It is to be specified here that genes are grouped based on the correlation coefficient. Thus, each group consists of a set of genes that have similar expression patterns. The genes with similar expression patterns should be similarly differentially regulated and are expected to be involved in a specific biochemical pathway.

We use neuro-fuzzy models (NFM-1 or NFM-2) [5–7], in the next steps, for selecting the most important group followed by the most important genes. Since the number of measurements (samples) is quite low, we need to generate more data. This will be helpful to create a data rich environment where artificial neural networks are more effective. We now describe the method for generating data. This is followed by the description of the neuro-fuzzy models [5–7].

3.2. Generation of data

After grouping, let us assume that we have a set of K non-overlapping groups, viz., $\{G_1, G_2, \ldots, G_k, \ldots, G_K\}$ such that $|G_k| = n_k, \forall k$. Let us also assume that a member of G_k is represented by $\mathbf{g}_k = [\mathbf{g}_{k1}, \mathbf{g}_{k2}, \ldots, \mathbf{g}_{kl}, \ldots, \mathbf{g}_{kp}]^T$ such that $\mathbf{g}_k = \mathbf{x}_j$, for some value of j. Then we choose one gene from each group and form a vector $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k, \ldots, \mathbf{v}_K]^T$, where $v_k = g_{kl}$, lth sample value. That is, the components of vector $\mathbf{v}$ are the lth normal sample values of K genes that are drawn from each group G_k. Similarly, other such vectors are formed by the other normal sample values and we have a total of p such vectors for each draw of K genes, one from each group. We thus create a set S of all such vectors $\mathbf{v}$ from normal samples so that the number of such vectors in S is

$$s = |S| = p \times \prod_{k=1}^{K} n_k \tag{3}$$

Similarly, another set S' of vectors $\mathbf{v}'$ is created from the diseased samples such that

$$s' = |S'| = q \times \prod_{k'=1}^{K'} n'_{k'} \tag{4}$$

where K' is the number of groups of genes and $n'_{k'}$ is the number of genes in k' th group. Now we have two sets S and S' of vectors $\mathbf{v}$ and $\mathbf{v}'$, respectively.

3.3. Neuro fuzzy models [5–7]

In the next steps, we use two neuro-fuzzy algorithms that have been developed in [5–7]. Although they developed neuro-fuzzy models for feature selection, we have implemented the methods for the purpose of gene selection. Here, we have termed these modes as neuro-fuzzy model-1 (NFM-1) and neuro-fuzzy model-2 (NFM-2).

3.3.1. Model NFM-1

Neuro-fuzzy model-1 (NFM-1) deals with formulation of an evaluation index followed by its minimization. On minimization, the weights of links in NFM-1 provide ranking of the groups of genes or individual genes, based on their importance in mediating a cancer.

Evaluation index: Let us consider a microarray gene expression data set comprising expression values of n genes in both normal (class C_1) and diseased (class C_2) conditions. That is, the class C_1 contains an expression pattern of n genes for normal samples and C_2 includes that for diseased (test) samples. Here we consider the classes C_1 and C_2 as two fuzzy sets, and the corresponding membership functions μ_{C_1} and μ_{C_2} have been considered as a triangular type [33]. Let $\mathbf{x}_i$ be the gene expression pattern of ith sample. That is, $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots x_{ij}, \ldots, x_{in}]^T$, where x_{ij} is the expression value of jth gene in the ith sample. The fuzzy evaluation index for a subset containing a few of these n genes is defined as [6]

$$E_{NFM-1} = \sum_{k=1}^{2} \sum_{\mathbf{x} \in C_k} \frac{s_k(\mathbf{x})}{\sum_{k' \neq k} s_{kk'}(\mathbf{x})} \times \alpha_k \tag{5}$$

where

$$s_k(\mathbf{x}) = \mu_{C_k}(\mathbf{x}) \times (1 - \mu_{C_k}(\mathbf{x})) \tag{6}$$

and

$$s_{12}(\mathbf{x}) = \tfrac{1}{2}[\mu_{C_1}(\mathbf{x}) \times (1 - \mu_{C_2}(\mathbf{x}))] + [\mu_{C_2}(\mathbf{x}) \times (1 - \mu_{C_1}(\mathbf{x}))] \tag{7}$$

with $\mu_{C_1}(\mathbf{x})$ and $\mu_{C_2}(\mathbf{x})$ being the membership values of the expression pattern $\mathbf{x}$ in classes C_1 and C_2, respectively. Here α_k $(= |C_k|)$ is the normalizing constant for class C_k which takes care of the effect of relative sizes of the classes. The membership function $\mu_C(\mathbf{x}_i)$ is defined as

$$\begin{aligned} \mu_C(\mathbf{x}_i) &= \left[\sum_{i,j} w_j^2 \left(\frac{x_{ij} - f_{min_j}}{f_{c_j} - f_{min_j}} \right)^2 \right]^{1/2} \quad \text{where } f_{min_j} \leq x_{ij} < f_{c_j} \\ &= \left[\sum_{i,j} w_j^2 \left(\frac{f_{max_j} - x_{i,j}}{f_{max_j} - f_{c_j}} \right)^2 \right]^{1/2} \quad \text{if } f_{c_j} \leq x_{ij} < f_{max_j} \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{8}$$

here C is C_1 for computing membership function corresponding to the normal class (C_1) and C_2 for that of the diseased class (C_2). The terms f_{max} and f_{min} are the maximum and minimum expression values, and f_c denotes the mean value of the expression pattern for corresponding classes. The weighting coefficient w_j denotes the importance of jth group (gene) in describing a class and separation between two classes. Eqs. (5)–(8) are such that the membership $\mu_C(\mathbf{x})$ of a representative gene of a group (or a gene in the most important group) gene x to class C is 1 if it is located at the mean of C, and 0.5 if it is at the boundary (i.e., at an ambiguous region) for a symmetric class structure. In practice, the class structure may not be symmetric. In that case, the membership values of some patterns at the boundary of the class will be greater than 0.5. Also, some expression patterns of other classes may have membership values greater than 0.5 for the class under consideration. For handling this undesirable situation, the membership function corresponding to a class needs to be transformed so that it can

model the real life class structures (normal or diseased) appropriately. For this purpose, we have incorporated a weighting factor (w) corresponding to a gene, so that the transformed membership functions model the class structures appropriately.

Artificial neural network model: E_{NFM-1} is now minimized with respect to w_is using NFM-1. For details of the architecture and learning algorithm of the network, one may refer to [6]. In the case of selecting the most important group, we consider $\mathbf{v}$ and $\mathbf{v}'$ (as described in Section 3.2) as the input patterns to the network. Thus, the number of nodes for selecting the most important group is K. Once the most important group is selected based on w-values, we present a gene expression pattern in this group to the network. In this case, the number of input nodes is n' ($n' \ll n$), where n' genes are included in the said group. In the schematic diagram (Fig. 2) of the NFM-1, the black circles represent the auxiliary nodes, and white circles represent the input and output nodes. Small triangles attached to the output nodes represent the modulatory connections from the respective auxiliary nodes. For further details one may refer to [6].

Algorithm-NFM-1. The steps involved in the training phase of NFM-1 are as follows:

- *Step* I: Set the weights of the feedback links from the auxiliary node corresponding to the class label, i.e., normal and disease, to $+1$ or -1.
- *Step* II: Initialize the weights of the feedforward links, with small random values from input nodes to output nodes.
- *Step* III: For each input pattern, do the following steps until convergence, i.e., until the change in the evaluation index becomes less than certain predefined threshold
  - *Step* III.1: Present the gene expression vector to the input layer of the network.
  - *Step* III.2: Activate only one auxiliary node at a time. The input nodes, in turn, send the resultant activations to the output nodes after activating an auxiliary node which sends the feedback to the input layer. The activation of the output node provides the membership value of the input pattern to the corresponding class. Hence, the membership values of the input pattern corresponding to all the classes are computed by sequentially activating the auxiliary nodes one at a time.

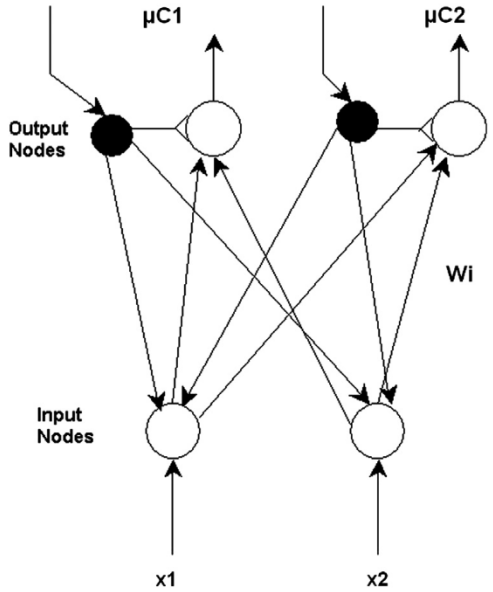

Fig. 2. Neuro-fuzzy model-1 (NFM-1) [6]. Here μ_{C_1} and μ_{C_2} indicate membership values in the normal and diseased states, respectively.

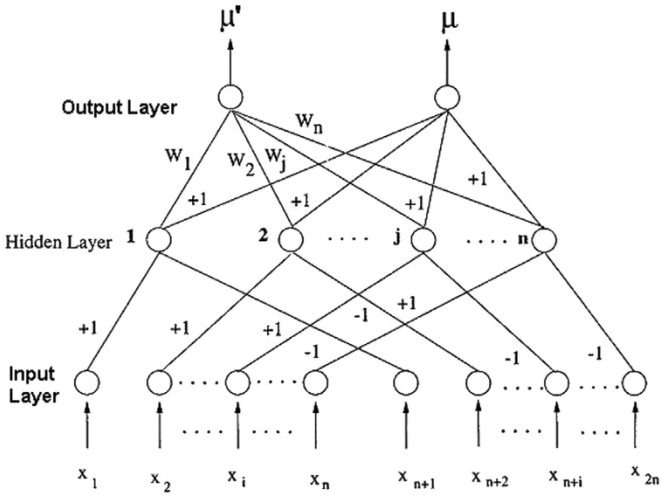

Fig. 3. Neuro-fuzzy model-2 (NFM-2) [7,5]. Here μ and μ' indicate membership values in the normal and diseased states, respectively.

  - *Step* III.3: Compute the evaluation index E_{NFM-1} using Eq. (5).
  - *Step* III.4: Compute the desired change in weight (w) values of the feedforward links to be made using the updating rule.
  - *Step* III.5: Compute total change in weight values for each input, over the entire set of patterns. Update weight values with the average change.
- *Step* IV: After convergence, the evaluation index attains a local minimum. In that case, the weight values of the feedforward links indicate the order of importance of the groups (or individual genes in the most important group).

3.3.2. Model NFM-2

Like NFM-1, the system NFM-2 involves formulation of a fuzzy evaluation index followed by minimization of the evaluation index in the framework of NFM-2. On minimization, the weights of the model indicate the order of groups and individual genes in the most important group. A set of weighting coefficients is used to denote the degree of importance of the individual genes in characterizing a disease and to provide flexibility in modeling. Although NFM-2 works under unsupervised learning, we have used given class labels (i.e., normal and diseased types) of the data. NFM-2 has just been considered as a tool for selecting the most important group and thereby the most important genes. Minimization of the evaluation index through unsupervised learning of the network determines the optimum weighting coefficients providing an ordering of the importance of genes individually. For details of the architecture and learning algorithm of the network, one may refer to Fig. 3 [7,5].

Evaluation index: Let μ_{ij} be the degree of similarity between the expression patterns of ith normal and jth diseased samples in the space of original expression pattern, and μ'_{ij} be that corresponding to some transformed space involving weighting coefficients. That is, μ may be interpreted as the membership values of expression patterns of a pair of genes belonging to the fuzzy set "similar". The evaluation index for a set of diseased samples is defined as [7,5]

$$E_{NFM-2} = \frac{2}{s(s-1)} \sum_i \sum_{j \neq i} \frac{1}{2} [\mu'_{ij}(1-\mu_{ij}) + \mu_{ij}(1-\mu'_{ij})] \tag{9}$$

The term s is the number of presentation of expression patterns. The membership function μ is defined as [7,5]

$$\mu_{ij} = 1 - \frac{\| \mathbf{x}_i - \mathbf{x}_j \|}{\beta \times \| \mathbf{x}_{max} - \mathbf{x}_{min} \|} \quad \text{where } u \leq u'$$
$$= 0 \quad \text{otherwise} \tag{10}$$

The terms $u = \|\mathbf{x}_i - \mathbf{x}_j\|$ and $u' = \beta \times \|\mathbf{x}_{max} - \mathbf{x}_{min}\|$. The terms $\mathbf{x}_{max}$ and $\mathbf{x}_{min}$ are the maximum and the minimum expression pattern in the data set. $\beta \in (0, 1]$ defines the degree of flattening of the membership function. As the value of β increases, numbers of non-zero elements for μ and μ' also increase.

The membership function μ' is defined as [7,5]

$$\mu'_{ij} = 1 - \frac{(\sum_{l=1}^{n}(w_l^2) \times (x_{il} - x_{jl})^2)^{1/2}}{\beta \times \|\mathbf{x}_{max} - \mathbf{x}_{min}\|} \quad \text{where } z \le z'$$
$$= 0 \quad \text{otherwise} \tag{11}$$

The terms $z = (\sum_{l=1}^{n}(w_l^2) \times (x_{il} - x_{jl})^2)^{1/2}$ and $z' = \beta \times (\|\mathbf{x}_{max} - \mathbf{x}_{min}\|)$. Here, $w_l \in [0, 1]$ represents the weighting coefficient corresponding to lth gene. It has the following characteristics: (1) for $\mu_{ij} < 0.5$, E_{NFM-2} decreases as $\mu'_{ij} \to 0$. For $\mu_{ij} > 0.5$, as $\mu'_{ij} \to 1$, E_{NFM-2} decreases. In both the cases, the contribution of the pair of expression patterns to the evaluation index E_{NFM-2} becomes minimum ($=0$) when $\mu_{ij} = \mu'_{ij} = 0$ or 1; (2) for $\mu_{ij} < 0.5$, as $\mu'_{ij} \to 1$, E_{NFM-2} increases. For $\mu_{ij} > 0.5$ as $\mu'_{ij} \to 0$, E_{NFM-2} increases. In both the cases, the contribution of the pair of expression patterns to E_{NFM-2} becomes maximum ($=0.5$) when $\mu_{ij} = 0$ and $\mu'_{ij} = 1$, or $\mu_{ij} = 1$ and $\mu'_{ij} = 0$; and (3) if $\mu_{ij} = 0.5$, the contribution to E_{NFM-2} becomes constant ($= 0.25$), i.e., independent of μ'_{ij}.

Therefore, the evaluation index decreases as the degree of similarity between a normal expression pattern and a diseased expression pattern of a gene tends to be either zero (when $\mu < 0.5$) or one (when $\mu > 0.5$), and becomes minimum for $\mu_{ij} = \mu'_{ij} = 0$ or 1. Therefore, our objective is to select those genes for which the evaluation index becomes minimum, thereby optimizing the decision on the similarity between a pair of such gene expression patterns. Characteristic (2) implies that E_{NFM-2} increases when similar (dissimilar) gene expression patterns (one in the normal condition and the other in the diseased condition) become dissimilar (similar) in the weighted space. That is, the process will automatically protect any occurrence of such a situation through minimizing E_{NFM-2}. Similarly, when $\mu_{ij} = 0.5$, i.e., decision regarding this similarity is the most ambiguous, the contribution of the pattern pair to E_{NFM-2} does not have any impact on the minimization process.

Artificial neural network model: As in the case of NFM-1, E_{NFM-2} is now minimized with respect to w_ls using an artificial neural network model. For details of the architecture and learning algorithm of the network, one may refer to [7,5]. For the purpose of selecting the most important group, we consider μ and μ' to generate Eq. (11). Thus, the number of nodes for selecting the most important group is $2K$. Once the most important group is selected based on w-values, we present the gene expression pattern in this group to the network. The neuro-fuzzy model-1 network (Fig. 3) consists of an input, a hidden, and an output layer. In this case, the number of input nodes is $2n'$ ($n' \ll n$), where n' genes are included in the said group. The hidden layer consists of n' number of nodes. The output layer consists of two nodes: one of them computes μ and the other μ'. A hidden node is connected only to ith and $(i+n')$ th input nodes via weights $+1$ and -1, respectively.

Algorithm-NFM-2. The steps involved in the training phase under unsupervised leaning are as follows:

- *Step* I: Initialize the weight values of the feedforward links with small random numbers from input nodes to output nodes.
- *Step* II: For each input pattern, do the following steps until convergence, i.e., until the change in evolution index becomes less than certain predefined threshold
  - *Step* II.1: Present the gene expression vector to the input layer of the network.
  - *Step* II.2: A hidden node is connected only to ith and $(i+n)$ th input nodes via weights $+1$ and -1, respectively. The output node, computing membership values μ', which are connected to a jth hidden node via weight, whereas that computing the membership values in the original space is connected to all the hidden nodes via weights $+1$ each.
  - *Step* II.3: Compute evaluation index E_{NFM-2} using Eq. (9).
  - *Step* II.4: Compute the change in weights of the feedforward links.
  - *Step* II.5: Compute the total change in weight values over the entire set of patterns. Update weight values with the average change.
- *Step* III: After convergence, evaluation index attains a local minimum. In that case, the weight values of the feedforward links indicate the order of importance of the genes groups (or individual genes in the most important group).

3.4. Selection of the most important group

For NFM-1, we consider that normal and diseased samples form two classes, viz., *normal* (C_1) and *diseased* (C_2). We take the number of input nodes as K, and the other nodes along with the architecture of the system are decided automatically [6]. In the case of NFM-2, the number of input nodes is $2K$, and the other nodes along with its architecture are decided automatically [7,5]. The first K nodes receive the vectors $\mathbf{v}$ as their inputs and second K nodes receive $\mathbf{v}'$. Thus, the number of such presentations is $s \times s'$. After learning in both the systems, we get weight values ($\mathbf{w}$) representing the importance of each group. Thus, the most important group is selected for which the weight value ($\mathbf{w}$) is the largest.

The term "important group" means that the genes in the group being involved in a particular biochemical pathway (alteration of which leading to development of a cancer) have changed their expression patterns in diseased samples significantly. The groups are then ranked using neuro-fuzzy models based on the extent of change in the expression patterns of genes in a group, in diseased samples. The top ranked group is called "the most important group". Thus, the most important group is expected to contain the genes responsible for the development of a cancer.

3.5. Selection of important genes from the most important group

Once the most important group is selected, only the genes in this group are considered. If the number of genes in the most important group is n' ($n' \ll n$), the numbers of input nodes in NFM-1 and NFM-2 are n' and $2n'$, respectively. The remaining parts of the architecture of both the systems are determined automatically [5–7]. As in the case of selection of groups, the number of classes for NFM-1 is 2. For NFM-2, the first n' input nodes receive expression values of genes (in the most important group) of normal samples and the next n' nodes receive that of diseased samples. Thus, the number of presentations in NFM-2 is $p \times q$. After learning, we get weight values corresponding to each gene representing its importance. Then we may consider selecting a few important genes based on their w-values.

The genes in the most important group are then ranked by the neuro-fuzzy models based on the extent of deviations of their expression patterns in the diseased samples. The top ranked genes are called "important genes".

4. Computational cost

In this section, we shall derive an estimate for the cost incurred in the computation. We have done it on an n number of genes.

Cost for calculating correlation coefficient: The entire method starts with grouping of genes using correlation coefficient. The

number of computations for calculating correlation coefficients of n genes is $\binom{n}{2} = n(n-1)/2$. Thus, the computational complexity for grouping phase will be $O(n^2)$.

Cost for data generation: In the very next phase, generation of data requires s computations where $s = p \times \prod_{k=1}^{K} n_k$. It can be represented as $p \times [(n/K) \times (n/K) \times \cdots \times (n/K)]$ or $(n/K)^K$ where we assume that n genes are distributed equal to K groups such that $n \gg K$. Hence, the computational cost for data generation phase will be $O(n^K)$. For group selection, an s number of patterns are taken as an input to NFM-1 and NFM-2.

Cost for computations in NFM -1 and NFM-2: After the selection of a group, the genes within the group, i.e., n/K genes are taken as an input to NFM-1 and NFM-2. If the number of iterations required for convergence is t, then the number of computations requires $O((n/K)*t)$ or $O(t*n)$. Thus, total computational cost becomes $O(n^2)+O(n^K)+O(t*n)=O(n^K)$.

5. Description of data sets

5.1. Human lung expression data

Human lung gene expression data is obtained by microarray experiments of Affymetrix Corporation data for Ann Arbor tumors and normal lung samples [34]. 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 can be found in [34]. The database web link is http://ncbi.nlm.nih.gov/projects/geo/.

5.2. Human colon expression data

This data set [35] consists of 18 tumor and 18 normal samples. Samples were obtained from colon adenocarcinoma specimens snap-frozen in liquid nitrogen within 20 min of removal/collection from patients. From some of these patients, paired normal colon tissue was also obtained. The microarrays were hybridized using an Affymetrix Hum600 array by a standard protocol. Two thousand highest intensity genes were selected and published on the web at http://microarray.princeton.edu/oncology/. From this subset, seven diagnostic genes were selected which give 100% correct classification. In this data set, samples of colon adenocarcinoma and paired normal tissue from the same patient were obtained from the Cooperative Human Tissue Network. The tissue was snap-frozen in liquid nitrogen within 20–30 min of harvesting and stored thereafter at −80 °C. mRNA was extracted from the bulk tissue samples and hybridized to the array using standard procedures. The adenocarcinoma samples were specifically re-reviewed by a pathologist where the samples were obtained using paraffin-embedded tissue that was adjacent or in close proximity to the frozen sample from which the mRNA was extracted. The publicly available data set consists of 18 adenocarcinoma and 18 normal samples. The set consists of 6600 genes and expressed sequence tags (ESTs).

5.3. Human breast cell expression data

In this data set, there are 22,645 genes of breast cancer cell expression profiles (HG-U133B) [36]. The data set consists of array based gene expression profiling of breast cancer cell lines HCC 1954 and MDA-MB-436 in reference to mammary epithelial cells (data set ID: GDS 823). In this data set there are 6 samples; two samples are for normal breast epithelium control replicate human mammary epithelial cells, and the remaining four samples for Breast Cancer cells. The database web link is http://ncbi.nlm.nih.gov/projects/geo/.

5.4. Human soft tissue sarcoma expression data

This data set consists of expression profiling of soft tissue sarcoma samples of *Homo sapiens*. Hypoxic regions often develop in tumors as they increase in size. Results provide insight into the expression of hypoxia-related genes in sarcomas under oligonucleotide technology. In this data set, there are 22283 genes with 15 normal samples and 39 diseased samples [37]. Among these 39 diseased samples, 7 fibrosarcoma samples, 2 GIST (gastrointestinal stromal) samples, 6 Leiomyosarcoma samples, 4 dedifferentiated liposarcoma samples, 3 pleomorphic liposarcoma samples, 9 MFH (malignant fibrous histiocytoma) samples, 4 Round cell sarcoma samples and 4 Synovial sarcoma samples are present in this data set (data set ID: GDS 1209). The database web link is http://ncbi.nlm.nih.gov/projects/geo/.

5.5. Human lymphocytes and plasma cell expression data

The title of the data set is Waldenstrom's macroglobulinemia (B lymphocytes and plasma cells) [38]. It has been used for analysis of B

Table 1
Selection of groups and genes by the neuro-fuzzy models for different data sets. *Note*: Both NFM-1 and NFM-2 have selected the same group as the best group for all the data sets. The third column indicates the set of common genes that are identified after applying NFM-1 and NFM-2 on the most important group.

Data sets	Selected group	No. of selected genes from selected group	Groups	No. of genes in each group
Lung	1	22	1	1659
			2	1247
			3	1290
			4	741
			5	666
			6	1526
Colon	2	21	1	846
			2	1156
			3	1041
			4	1293
			5	763
			6	653
			7	903
			8	809
Leukemia	4	19	1	2814
			2	2987
			3	2770
			4	2567
			5	2896
			6	2778
			7	2888
			8	2583
Sarcoma	7	18	1	2557
			2	2371
			3	2744
			4	2379
			5	2289
			6	2638
			7	2191
			8	2640
			9	2456
Breast	5	17	1	2178
			2	1956
			3	2207
			4	1882
			5	2014
			6	1971
			7	2303
			8	2264
			9	2369
			10	1854
			11	1647

lymphocytes (BL) and plasma cells (PC) from patients with Waldenstrom's macroglobulinemia (WM), a B-lymphoproliferative disorder (BLPD) (data set ID: GDS 2643). The entire data set consists of 22,283 genes with 56 samples. Among them, there are 13 normal samples (8 normal B lymphocytes and 5 normal plasma cells) and 43 diseased (20 Waldenstrom's macroglobulinemia, 11 chronic lymphocytic leukemia, and 12 multiple myeloma) samples. The database web link is http://ncbi.nlm.nih.gov/projects/geo/.

6. Results and discussion

In this section, the effectiveness of the proposed methodology has been demonstrated on human lung [34], colon [35], breast cell [36], soft tissue Sarcoma [37], and human lymphocytes and plasma cell [38] gene expression data. A comparative analysis with SAM, SNR, BR, NA, SVM, GMM, HMM, CAFS, Entropy-PLR, NBC-MSC, NBC-MMC, NMSC-MSC, NMSC-MMC, LOOCSFS, GLGS, SFS-LSBOUND and SFFS-LSBOUND, VarSelRF and SlimPLS has also been included.

We have applied the methodology on the aforesaid gene expression data sets for selecting some important gene mediating diseases. According to the methodology, first of all, the genes are placed into groups based on correlation values. For human lung gene expression data, we have found 6 groups, containing 1659, 1247, 1290, 741, 666 and 1526 genes (Table 1). The group containing 1659 genes has been selected as the most important group by both NFM-1 and NFM-2.

For both NFM-1 and NFM-2, we have considered two classes: one representing the data of normal samples and the other for the diseased samples. The weighting coefficient **w** has been initialized by

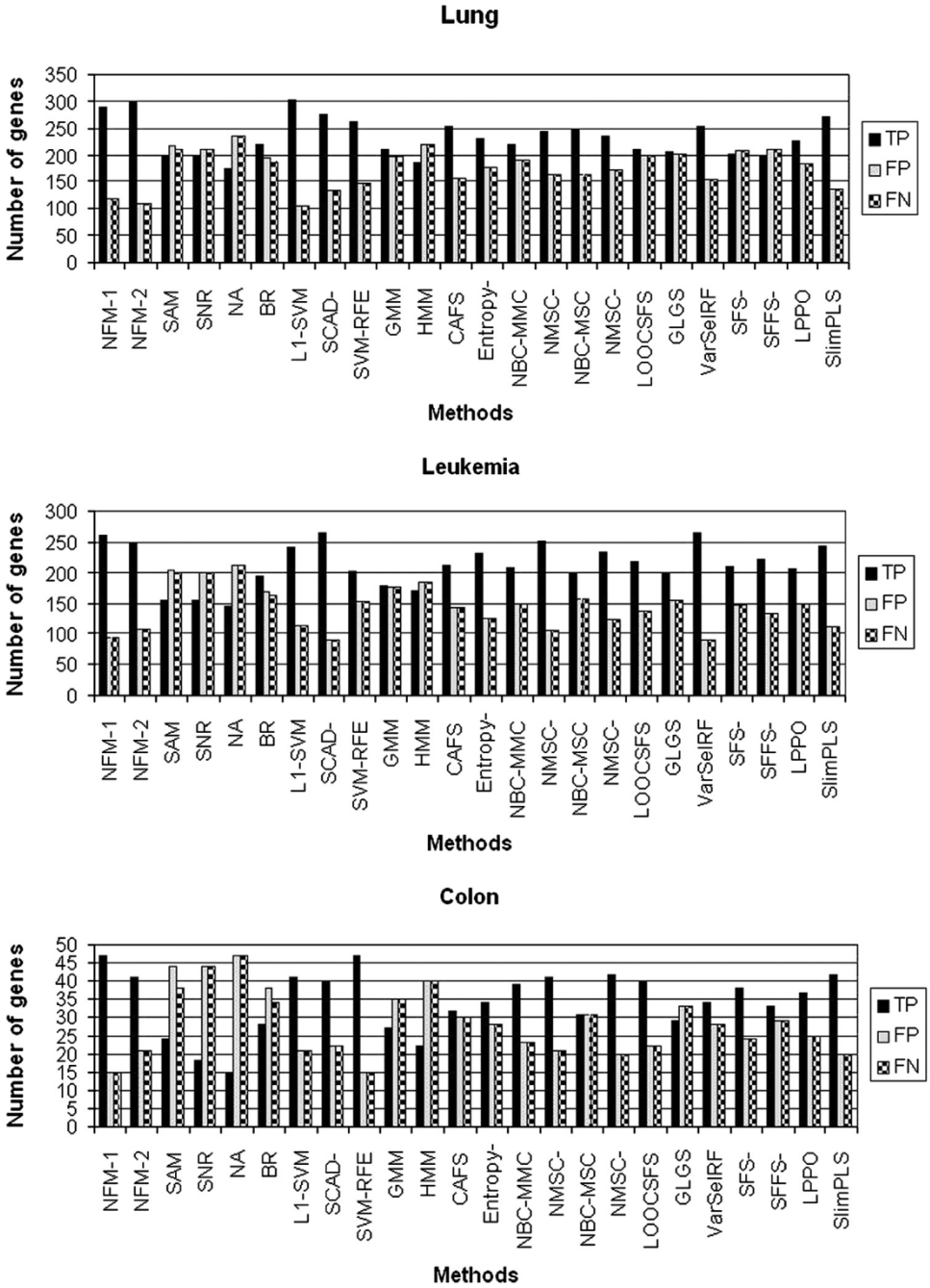

Fig. 4. Comparison among the methods in terms of biochemical pathways. Here *TP*, *FP* and *FN* indicate the number of true positive, false positive and false negative, respectively.

random numbers. Learning rate parameter η has been set to 0.1. After grouping of genes and selection of the most important group, we have found 30 and 32 genes, respectively, by NFM-1 and NFM-2. Among these genes, we have found 22 genes that are present in both the results. Finally, we have selected the 20 most important genes based on their weight values $\mathbf{w}$ (chosen from higher values of w_i's).

Similar experiments have been carried out for the other data sets too. For example, in the case of human colon gene expression data set, eight groups containing 846, 1156, 1041, 1293, 763, 653, 903 and 809 genes have been found. Likewise, eleven groups (containing 2178, 1956, 2207, 1882, 2014, 1971, 2303, 2264, 2369, 1854, and 1647 genes) for human breast cell gene expression data, nine groups (containing 2575, 2371, 2744, 2379, 2289, 2638, 2191, 2640, and 2456 genes) for human soft tissue sarcoma gene expression data, eight groups (containing 2814, 2987, 2770, 2567, 2896, 2778, 2888, and 2583 genes) for human lymphocyte and plasma cell gene expression data have been found.

On applying both NFM-1 and NFM-2, we have got the groups containing 1156 genes for human colon expression data, 2014 genes for human breast cell expression data, 2191 genes for human soft tissue sarcoma expression data, and 2567 genes for human lymphocyte and plasma cell expression data to be the best groups. Finally, we

Table 2
Comparative results on the number of enriched attributes for various sets of significant genes corresponding to different methods. We have termed our methodology as NFM-1 and NFM-2 depending on the use of NFM-1 and NFM-2 in the second and third steps.

Data set	Gene set	Number of enriched attributes obtained by										
		NFM-1	NFM-2	SAM	SNR	NA	BR	L1-SVM	SCAD-SVM	SVM-RFE	GMM	HMM
Lung	First 5	62	60	14	10	12	17	56	52	61	29	22
	First 10	75	76	20	19	15	24	62	61	67	37	27
	First 15	80	82	25	24	16	28	70	72	73	43	31
	First 20	82	95	29	33	18	32	78	75	79	47	38
Colon	First 5	49	54	27	23	25	28	51	42	39	21	19
	First 10	61	60	31	32	31	31	52	45	51	28	28
	First 15	63	63	33	35	36	31	58	56	55	28	32
	First 20	66	68	35	44	40	33	61	57	55	32	36
Sarcoma	First 5	80	84	41	49	47	56	78	82	69	55	59
	First 10	92	95	50	54	57	63	91	92	78	67	65
	First 15	100	101	61	60	68	74	101	94	83	72	79
	First 20	106	113	65	73	78	84	105	104	98	82	81
Leukemia	First 5	86	81	59	19	36	52	71	80	78	43	50
	First 10	92	89	60	26	48	59	76	87	89	51	58
	First 15	104	101	70	37	55	70	88	89	99	57	69
	First 20	111	110	81	43	68	73	92	97	107	71	81
Breast	First 5	36	33	15	6	16	12	38	31	27	12	17
	First 10	48	44	18	12	18	13	42	38	32	15	17
	First 15	56	51	22	14	21	17	47	38	35	15	18
	First 20	59	60	29	17	27	20	56	41	37	17	19

Table 3
Comparative results on the number of enriched attributes for various sets of significant genes corresponding to different methods. We have termed our methodology as NFM-1 and NFM-2 depending on the use of NFM-1 and NFM-2 in the second and third steps.

Data set	Gene set	Number of enriched attributes obtained by														
		NFM-1	NFM-2	CAFS	Entropy-PLR	NBC-MMC	NMSC-MMC	NBC-MSC	NMSC-MSC	LOOCSFSFS	GLGS	VarSelRF	SFS-LSBOUND	SFFS-LSBOUND	LPPO	SlimPLS
Lung	First 5	62	60	34	32	36	44	40	31	28	37	45	37	27	32	40
	First 10	75	76	38	33	36	48	49	51	30	39	49	50	29	38	44
	First 15	80	82	38	37	43	48	53	51	35	42	50	52	41	61	67
	First 20	82	95	45	47	51	53	57	60	43	54	60	61	55	69	70
Colon	First 5	49	54	24	30	24	19	22	19	31	22	25	15	22	33	32
	First 10	61	60	29	31	27	22	22	20	36	25	25	18	25	35	39
	First 15	63	63	31	31	29	27	27	20	36	28	30	22	29	36	45
	First 20	66	68	40	34	30	29	29	22	38	30	32	24	31	40	48
Sarcoma	First 5	80	84	44	40	45	51	50	38	37	43	60	47	56	72	67
	First 10	92	95	49	47	55	55	51	40	43	49	67	67	59	75	78
	First 15	100	101	52	52	59	59	55	57	56	50	70	80	62	75	91
	First 20	106	113	61	57	63	67	57	62	66	54	87	82	67	79	95
Leukemia	First 5	86	81	51	44	35	55	62	82	43	60	52	76	80	78	81
	First 10	92	89	56	45	50	57	71	85	61	69	55	82	84	90	89
	First 15	104	101	62	50	61	57	79	95	67	78	73	85	92	95	97
	First 20	111	110	67	59	67	61	81	101	87	78	84	89	95	96	104
Breast	First 5	36	33	19	33	26	29	29	15	12	28	23	32	16	29	37
	First 10	48	44	29	47	35	33	50	28	19	39	24	37	18	42	46
	First 15	56	51	37	62	39	45	59	29	28	45	24	49	34	46	54
	First 20	59	60	52	69	53	56	62	34	39	50	25	58	39	54	66

have got 25, 22, 21 and 24 most important genes corresponding to the aforesaid data sets through the algorithm NFM-1. Using NFM-2, these numbers are 27, 21, 21, and 23 for human colon, human breast cell, human soft tissue sarcoma, and human lymphocyte and plasma cell gene expression data sets, respectively. The numbers of genes that are obtained by both NFM-1 and NFM-2 are 21, 17, 18, and 19 corresponding to these data sets (Table 1). As before, the values of η are set to 0.1. These selected genes are then explored for their role in causing diseases through computing the number of functional enrichments that are obtained corresponding to them, and this is described in the next subsection.

6.1. Comparison and validation of results

In this section, we compare the results obtained by various methods including NFM-1 and NFM-2. This comparison has been done using biochemical pathways, *p*-value, *t*-test, *F*-test and

Table 4
Comparative results on the number of enriched attributes for various sets of least significant genes corresponding to different methods. We have termed our methodology as NFM-1 and NFM-2 depending on the use of NFM-1 and NFM-2 in the second and third steps.

Data set	Gene set	Number of enriched attributes obtained by										
		NFM-1	NFM-2	SAM	SNR	NA	BR	L1-SVM	SCAD-SVM	SVM-RFE	GMM	HMM
Lung	Last 5	0	0	1	1	1	2	0	0	0	1	2
	Last 10	0	0	3	1	2	2	0	0	0	2	2
	Last 15	0	0	4	3	2	3	0	0	0	2	2
	Last 20	0	0	5	3	4	3	0	0	1	2	2
Colon	Last 5	0	0	1	2	2	4	0	0	1	4	3
	Last 10	0	0	2	4	2	5	1	0	1	4	3
	Last 15	0	0	3	5	3	7	1	1	1	5	5
	Last 20	1	1	4	7	4	8	1	1	3	6	7
Sarcoma	Last 5	0	0	4	6	10	8	0	1	0	2	2
	Last 10	0	0	9	10	12	10	1	2	1	4	4
	Last 15	0	0	11	10	12	13	2	2	2	4	5
	Last 20	1	1	14	13	18	17	2	2	2	4	7
Leukemia	Last 5	0	0	7	9	11	10	0	0	0	0	0
	Last 10	0	0	9	10	12	10	0	0	0	2	3
	Last 15	0	0	10	11	12	12	0	2	1	3	4
	Last 20	0	0	10	13	13	15	1	2	1	4	6
Breast	Last 5	0	0	6	3	6	7	0	0	0	1	1
	Last 10	0	0	8	4	6	7	0	0	3	2	2
	Last 15	0	1	11	4	7	10	1	0	5	4	4
	Last 20	0	2	12	8	9	13	2	1	5	4	5

Table 5
Comparative results on the number of enriched attributes for various sets of least significant genes corresponding to different methods. We have termed our methodology as NFM-1 and NFM-2 depending on the use of NFM-1 and NFM-2 in the second and third steps.

Data set	Gene set	Number of enriched attributes obtained by														
		NFM-1	NFM-2	CAFS	Entropy-PLR	NBC-MMC	NMSC-MMC	NBC-MSC	NMSC-MSC	LOOCSFSFS	GLGS	Var Sel RF	SFS LS BOUND	SFFS LS BOUND	LPPO	Slim PLS
Lung	Last 5	0	0	2	1	0	0	4	2	0	4	0	1	0	0	0
	Last 10	0	0	2	3	3	0	4	2	3	6	1	1	0	1	1
	Last 15	0	0	2	3	6	0	5	2	3	8	1	2	2	3	1
	Last 20	0	0	4	3	6	2	6	2	6	10	2	2	3	4	1
Colon	Last 5	0	0	3	2	2	2	0	3	0	0	3	2	2	0	1
	Last 10	0	0	4	3	5	4	2	4	2	1	5	2	4	0	2
	Last 15	0	0	5	4	7	7	6	5	2	1	5	5	4	2	2
	Last 20	1	1	5	5	8	9	9	8	3	3	6	7	5	2	3
Sarcoma	Last 5	0	0	0	0	0	1	2	0	2	6	1	0	1	2	2
	Last 10	0	0	1	0	2	2	3	1	2	7	2	2	2	3	3
	Last 15	0	0	3	1	3	3	3	2	4	9	3	3	2	3	4
	Last 20	1	1	4	1	6	4	4	4	7	9	4	4	2	5	4
Leukemia	Last 5	0	0	0	1	2	1	0	0	1	1	0	1	0	3	0
	Last 10	0	0	0	1	2	2	3	0	4	1	0	2	2	4	0
	Last 15	0	0	1	2	3	2	4	2	4	1	2	2	3	4	2
	Last 20	0	0	1	2	4	2	5	3	4	3	3	5	7	4	2
Breast	Last 5	0	0	1	0	0	3	1	2	0	2	2	1	0	0	0
	Last 10	0	0	2	1	0	4	2	2	1	3	2	1	0	0	0
	Last 15	0	1	2	1	3	5	2	3	2	4	2	4	2	0	0
	Last 20	0	2	2	1	3	5	3	5	2	5	2	6	2	2	1

sensitivity. We have also tried to validate some of our results using some earlier investigations. In addition, we have implemented various types of biological and statistical parameters like *pi*-GSEA, Fisher-score, KOGS, SPEC, *W*-test, and BWS for performance comparisons and validation.

6.1.1. Using biochemical pathways

Here we consider various biochemical pathways that are involved in different cancers considered here. We have found these pathways from NCBI databases (http://ncbi.nlm.nih.gov/projects/geo/) for lung, leukemia and colon cancers only. Thus, we have been restricted to compare the results for these cancers only. From the biochemical pathways involved in a particular cancer, we consider the genes/proteins involved in them.

For lung cancer, we have found non-small cell lung cancer and small cell cancer pathways. A set of 409 genes is involved in these two pathways. We have compared this set of genes with those obtained by 24 methods. Here we have identified 291 and 298 genes that are common in database information and the results of NFM-1 and NFM-2, respectively. We have called these genes *true positive* (*TP*) genes. Thus, we have 118 and 111 genes that are in the set of top ranked 409 genes, respectively, obtained by NFM-1 and NFM-2, but not involved in the pathways. These 118 and 111 genes are considered as *false positive* (*FP*). Similarly, the number of *false negative* genes is 118 and 111 for NFM-1 and NFM-2, respectively. Likewise, we have computed the number of true positive, false positive and false negative genes for the other methods. From Fig. 4, it is clear that both NFM-1 and NFM-2 have been able to identify more true positive genes, but less false positive and false negative genes compared to all the other methods, along with L1-SVM, SCAD-SVM, and SlimPLS.

For human colon expression data, we have found 62 genes that are present in a colon cancer related pathway (i.e., colorectal cancer pathway). From Fig. 4, it is clearly observed that NFM-1 performs better than NFM-2 along with SVM-RFE and SlimPLS. NFM-2 generates a similar performance along with L1-SVM, SCAD-SVM, NMSC, LOOCSFS for colon cancer. Similarly, we have found 355 genes in leukemia related pathways like chronic myeloid and acute myeloid leukemia. As in the case of lung cancer both NFM-1 and NFM-2 along with L1-SVM, SCAD-SVM, NMSC, VarSelRF and SlimPLS have outperformed the other methods in terms of identifying more true positive genes, but less false positive and false negative genes (Fig. 4).

Table 6
Significant genes for human lung expression data set. The results are validated by *t*-test and *F*-test. Some of the results are validated by references. We have termed our methodology as NFM-1 and NFM-2 depending on the use of NFM-1 and NFM-2 in the second and third steps.

Method	Level of significance (%)	Genes (*t*-test)	Genes (*F*-test)	References
NFM-1	99.9	CALCA, PFKP, TYMS, IGFBP3, IARS, HBB, HLA-B, SFTPA2, TNF		[40–54]
	99	IGHG3, PRKACA, SORT1, SFTPA1, MEN1, IGHM	CALCA, PFKP, TYMS, IGFBP3, IARS, HBB, HLA-B, SFTPA2, TNF, IGHG3, PRKACA, SORT1, SFTPA1, MEN1, IGHM	
	95	RPLP0, SMCIL1, SFTPC, HLA-DRA, MGP, RNASE1	RPLP0, SMCIL1, SFTPC, HLA-DRA, MGP, RNASE1	
NFM-2	99.9	CALCA, PFKP, TYMS, IGFBP3, IARS, HBB, HLA-B, SFTPA2, TNF		[40–54]
	99	IGHG3, PRKACA, SORT1, MEN1, SFTPA1, IGHM	CALCA, PFKP, TYMS, IGFBP3, IARS, HBB, HLA-B, SFTPA2, TNF, IGHG3, PRKACA, SORT1, SFTPA1, MEN1, IGHM	
	95	RPLP0, SMCIL1, MGP, RNASE1, SFTPC, HLA-DRA	RPLP0, SMCIL1, SFTPC, HLA-DRA, MGP, RNASE1	
SAM	99.9	CALCA, HBB, SFTPA2, IGFBP3, TNF		[44,45,40–42,48–52]
	99	PRKACA, SORT1	CALCA, HBB, SFTPA2, IGFBP3, TNF, PRKACA, SORT1	
	95	HLA-DRA, POLB, PIGA	HLA-DRA, POLB, PIGA	
SNR	99.9	CALCA, IGFBP3, HBB, SFTPA2,		[48,52,40–42,44,45,53]
	99	IGHG3, MEN1, SORT1	CALCA, HBB, SFTPA2, IGFBP3, IGHG3, MEN1, SORT1	
	95	HLA-DRA, RPLP0, PIGA, ITGA9, POLB	HLA-DRA, POLB, PIGA, RPLP0, ITGA9	
NA	99.9	CALCA, IGFBP3, SFTPA2, TNF, HBB		[48,52,40–42,49–51,44,45,53]
	99	IGHG3, CEACAM4, MYL6, SORT1, SFTPA1	CALCA, HBB, SFTPA2, IGFBP3, TNF, IGHG3, SFTPA1, SORT1	
	95	POLB, PIGA	MYL6, PIGA, CEACAM4, POLB	
BR	99.9	CALCA, IGFBP3, HBB, SFTPA2, TNF		[48,40–42,49–52,44,45,54]
	99	IGHM, MYL6, SORT1, MEN1, SFTPA1	CALCA, HBB, SFTPA2, IGFBP3, TNF, MEN1, IGHM, SFTPA1, SORT1	
	95	SMCIL1, RPLP0	MYL6, RPLP0, SMCIL1	
SVM	99.9	PFKP, IGFBP3, IARS, HBB, SFTPA2, TNF		[40,43–45,49–54]
	99	IGHG3, MEN1, IGHM	PFKP, TYMS, IGFBP3, IARS, HBB, SFTPA2, TNF, SORT1, SFTPA1, MEN1, IGHM	
	95	HLA-DRA, MGP, RNASE1	RPLP0, SMCIL1, SFTPC, HLA-DRA, MGP, RNASE1	
GMM	99.9	PFKP, IARS, HBB, SFTPA2, TNF		[40–46,49–54]
	99	PRKACA, SORT1, MEN1,	CALCA, PFKP, TYMS, IGFBP3, IARS, TNF, IGHG3, PRKACA, SORT1, SFTPA1, MEN1, IGHM	
	95	RPLP0, SMCIL1, MGP, RNASE1, SFTPC, HLA-DRA	RPLP0, SMCIL1, SFTPC, HLA-DRA, MGP, RNASE1	
HMM	99.9	CALCA, PFKP, TYMS, HLA-B, SFTPA2, TNF		[40,43–54]
	99	SORT1, MEN1, SFTPA1, IGHM	CALCA, PFKP, TYMS, HLA-B, SFTPA2, TNF, IGHG3, SFTPA1, MEN1, IGHM	
	95	RPLP0, SMCIL1, MGP, RNASE1, SFTPC, HLA-DRA	RPLP0, SMCIL1, SFTPC, HLA-DRA, MGP, RNASE1	

6.1.2. Using p-values

In our study, the enrichment of each GO category [39] for each of the genes has been calculated by its p-value. A low p-value indicates that the genes belonging to the enriched functional categories are biologically significant. Here only functional categories with $p-\text{value} < 5.0 \times 10^{-5}$ were considered. We have made comparative study, with other methods, viz., SAM, SNR, BR, NA, SVM, GMM, HMM, CAFS, Entropy-PLR, NBC-MSC, NBC-MMC, NMSC-MSC, NMSC-MMC, LOOCSFS, GLGS, SFS-LSBOUND and SFFS-LSBOUND, VarSelRF and SlimPLS in terms of their ability to identify functionally enriched genes. Tables 2 and 3 show the number of functionally enriched attributes corresponding to these methods for different sets of genes. It has been found that for all the data sets, NFM-1, NFM-2 along with SlimPLS, SVM performed the best for all the data sets. These results show that the proposed methodology has been able to select the more important genes responsible for mediating a disease than the other methods, except SVM-RFE, L1-SVM, SCAD-SVM, VarselRF, LPPO, SlimPLS, and SFS-LSBOUND considered here.

Moreover, we made a comparative analysis for those genes that are least significant, as obtained by NFM-1 and NFM-2 as well as by the existing methods (Tables 4 and 5). From Tables 4 and 5, it is quite clear that the set of genes obtained as the least significant by NFM-1 and NFM-2 generates almost no GO attributes. For other methods, we have still found some GO attributes. This contradicts the fact that these subsets of genes should not have the lowest rank. On the other hand, the subsets of genes of somewhat higher ranks obtained by the existing methods correspond to no GO attributes. From this comparative study, we can conclude that NFM-1 and NFM-2 are more effective in identifying more functionally enriched genes than all the other methods, viz., SAM, SNR, BR, NA, SVM, GMM, HMM, CAFS, Entropy-PLR, NBC-MSC, NBC-MMC, NMSC-MSC, NMSC-MMC, LOOCSFS, GLGS, SFS-LSBOUND and SFFS-LSBOUND, VarSelRF and SlimPLS. In addition, we have shown that NFM-1 and NFM-2 are capable of finding out more number of true positive genes in terms of identifying the GO attributes and cancer attributes with respect to other existing methods (see Fig. 9). Here we have identified totally 485 GO attributes of 387 gene set and among them, we have identified totally 105 cancer related GO attributes.

6.1.3. Using t- and F-tests

In order to validate the results statistically, we have applied t-test on the genes identified by NFM-1, NFM-2, SAM, SNR, BR, NA, SVM, GMM, HMM, CAFS, Entropy-PLR, NBC-MSC, NBC-MMC, NMSC-MSC, NMSC-MMC, LOOCSFS, GLGS, SFS-LSBOUND and SFFS-LSBOUND, VarSelRF and SlimPLS on each data set. For human lung expression data, on applying NFM-1 and NFM-2, we have identified some important genes like CALCA (4.02), PFKP (5.78), TYMS (3.98), IGFBP3 (6.98), IARS (5.98), HBB (7.08), HLA-B (5.42), SFTPA2 (6.89) and TNF (4.23). The number in the bracket indicates the t-value corresponding to the gene. The t-values of these genes

Table 7
Significant genes for human lung expression data set. The results are validated by t-test and F-test. Some of the results are validated by references. We have termed our methodology as NFM-1 and NFM-2 depending on the use of NFM-1 and NFM-2 in the second and third steps.

Method	Level of significance (%)	Genes (t-test)	Genes (F-test)	References
NFM-1	99.9	CALCA, PFKP, TYMS, IGFBP3,IARS, HBB, HLA-B, SFTPA2,TNF		[40–54]
	99	IGHG3,PRKACA, SORT1,SFTPA1,MEN1, IGHM	CALCA, PFKP, TYMS, IGFBP3,IARS, HBB, HLA-B, SFTPA2,TNF, IGHG3,PRKACA, SORT1,SFTPA1,MEN1,IGHM	
	95	RPLP0,SMCIL1,SFTPC, HLA-DRA, MGP, RNASE1	RPLP0,SMCIL1,SFTPC, HLA-DRA, MGP, RNASE1	
NFM-2	99.9	CALCA, PFKP, TYMS, IGFBP3,IARS, HBB, HLA-B, SFTPA2,TNF		[40–54]
	99	IGHG3,PRKACA, SORT1,MEN1,SFTPA1, IGHM	CALCA, PFKP, TYMS, IGFBP3,IARS, HBB, HLA-B, SFTPA2,TNF, IGHG3,PRKACA, SORT1,SFTPA1,MEN1,IGHM	
	95	RPLP0,SMCIL1,MGP, RNASE1,SFTPC, HLA-DRA	RPLP0,SMCIL1,SFTPC, HLA-DRA, MGP, RNASE1	
CAFS and Entropy-PLR	99.9	CALCA, IGFBP3,TNF		[52,42,50,51]
	99	PRKACA, SORT1	CALCA, IGFBP3,TNF, PRKACA, SORT1	
	95	HLA-DRA, POLB, PIGA	HLA-DRA, POLB, PIGA	
NBC-MMC/MSC and NMSC-MMC/MSC	99.9	CALCA, IGFBP3,TNF		[48,52,49,50,53]
	99	IGHG3, CEACAM4, MYL6, SORT1, SFTPA1	CALCA, IGFBP3, TNF, IGHG3, SFTPA1, SORT1	
	95	POLB, PIGA	MYL6, PIGA, CEACAM4, POLB	
LOOCSFS and GLGS	99.9	CALCA, PFKP, TYMS, HLA-B, SFTPA2, TNF		[40,43–45,52–54]
	99	SORT1, MEN1	TYMS, HLA-B, SFTPA2, TNF, MEN1, IGHM	
	95	RPLP0, SFTPC, HLA-DRA	RPLP0, SMCIL1, MGP, RNASE1	
VarSelRF	99.9	CALCA, IGFBP3, SFTPA2, TNF, HBB		[48,52,49,50,53]
	99	IGHG3, CEACAM4, MYL6	CALCA, HBB, TNF, IGHG3, SFTPA1, SORT1	
	95	POLB, PIGA	MYL6, PIGA, CEACAM4, POLB	
SFS/SFFS-LSBOUND	99.9	CALCA, HBB, SFTPA2		[52,44,42,48,49,51]
	99	PRKACA, SORT1	CALCA, HBB, PRKACA, SORT1	
	95	HLA-DRA, POLB, PIGA	HLA-DRA, POLB, PIGA	
SlimPLS and LPPO	99.9	CALCA, PFKP, TYMS, IGFBP3, IARS, HLA-B, TNF		[40,43,46,52–54]
	99	IGHG3, PRKACA	CALCA, PFKP, TNF, IGHG3, PRKACA, IGHM	
	95	RPLP0, SMCIL1, SFTPC	HLA-DRA, MGP, RNASE1	

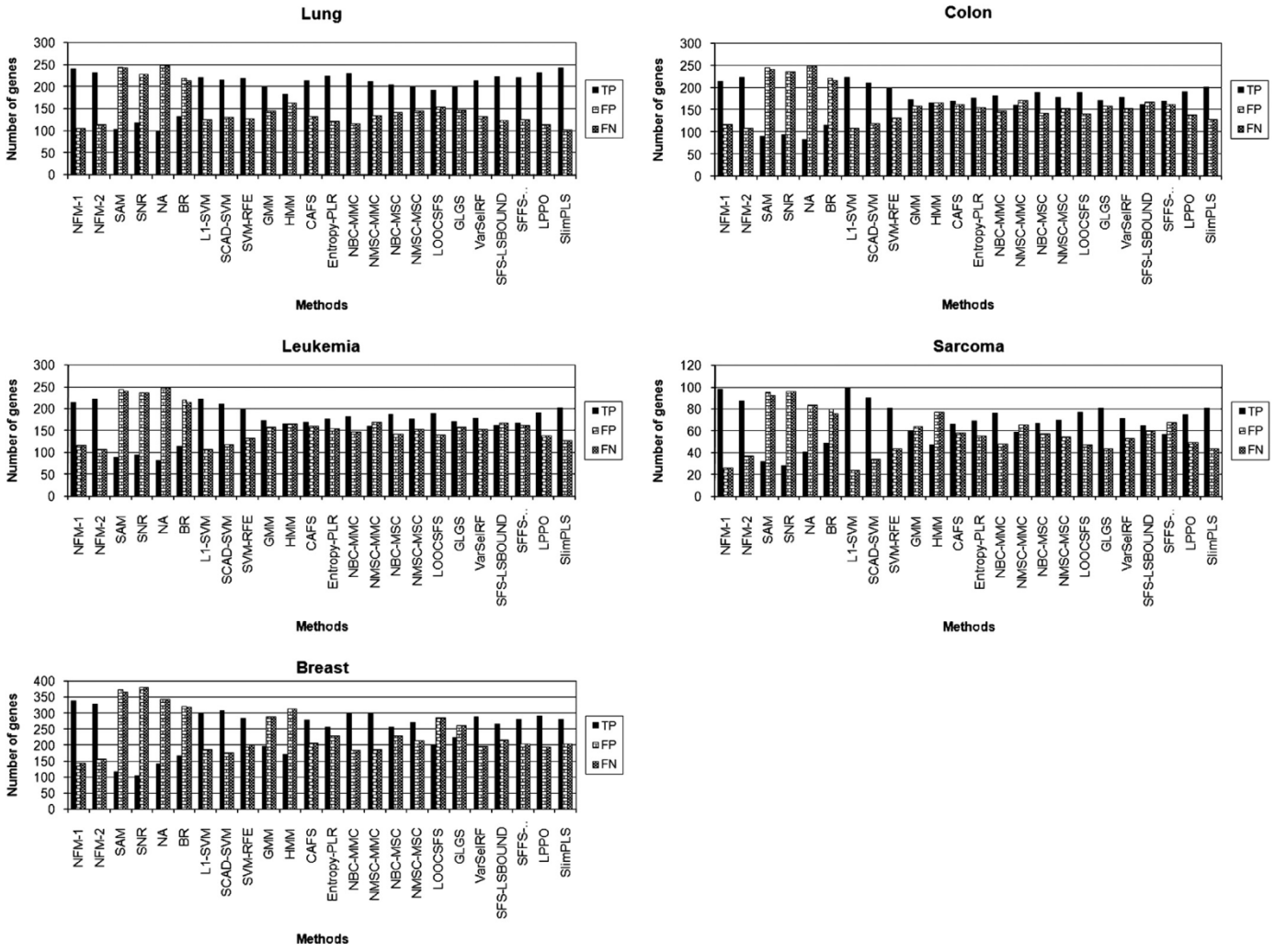

Fig. 5. Comparison among the methods using NCBI database. Here *TP*, *FP* and *FN* indicate the number of true positive, false positive and false negative, respectively.

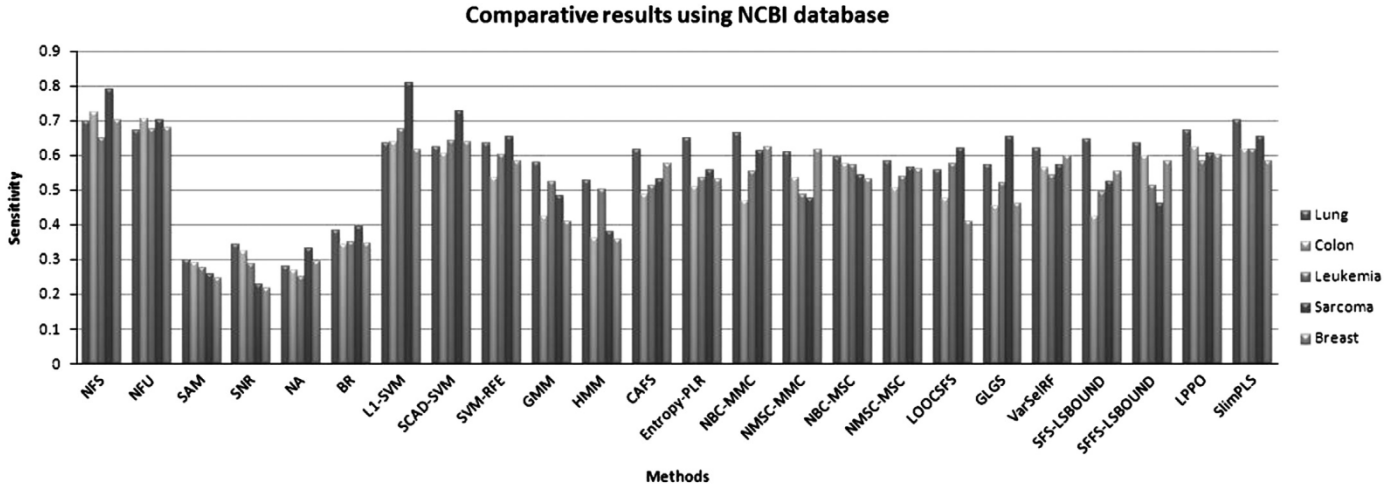

Fig. 6. Comparison among the methods using NCBI database in terms of sensitivity.

exceed the value for $P=0.001$. It indicates that these genes are highly significant (99.9% level of significance). Similarly, genes like IGHG3 (2.67), PRKACA (2.89), SORT1 (2.76), MEN1 (3.15), SFTPA1 (2.92) and IGHM (3.25) exceed the t-value for $P=0.01$. This means that these genes are significant at the level of 99%. Likewise, RPLP0 (2.12), SMCIL1 (2.07), MGP (2.31), RNASE1 (2.43), SFTPC (2.37) and HLA-DRA (2.27) genes are important at the level of 95% significance. We have performed t-test for the genes identified by other gene selection algorithms like SAM, SNR, BR, NA, SVM, GMM, HMM, CAFS, Entropy-PLR, NBC-MSC, NBC-MMC, NMSC-MSC, NMSC-MMC, LOOCSFS, GLGS, SFS-LSBOUND and SFFS-LSBOUND, VarSelRF and SlimPLS. But highly significant (99.9% significance

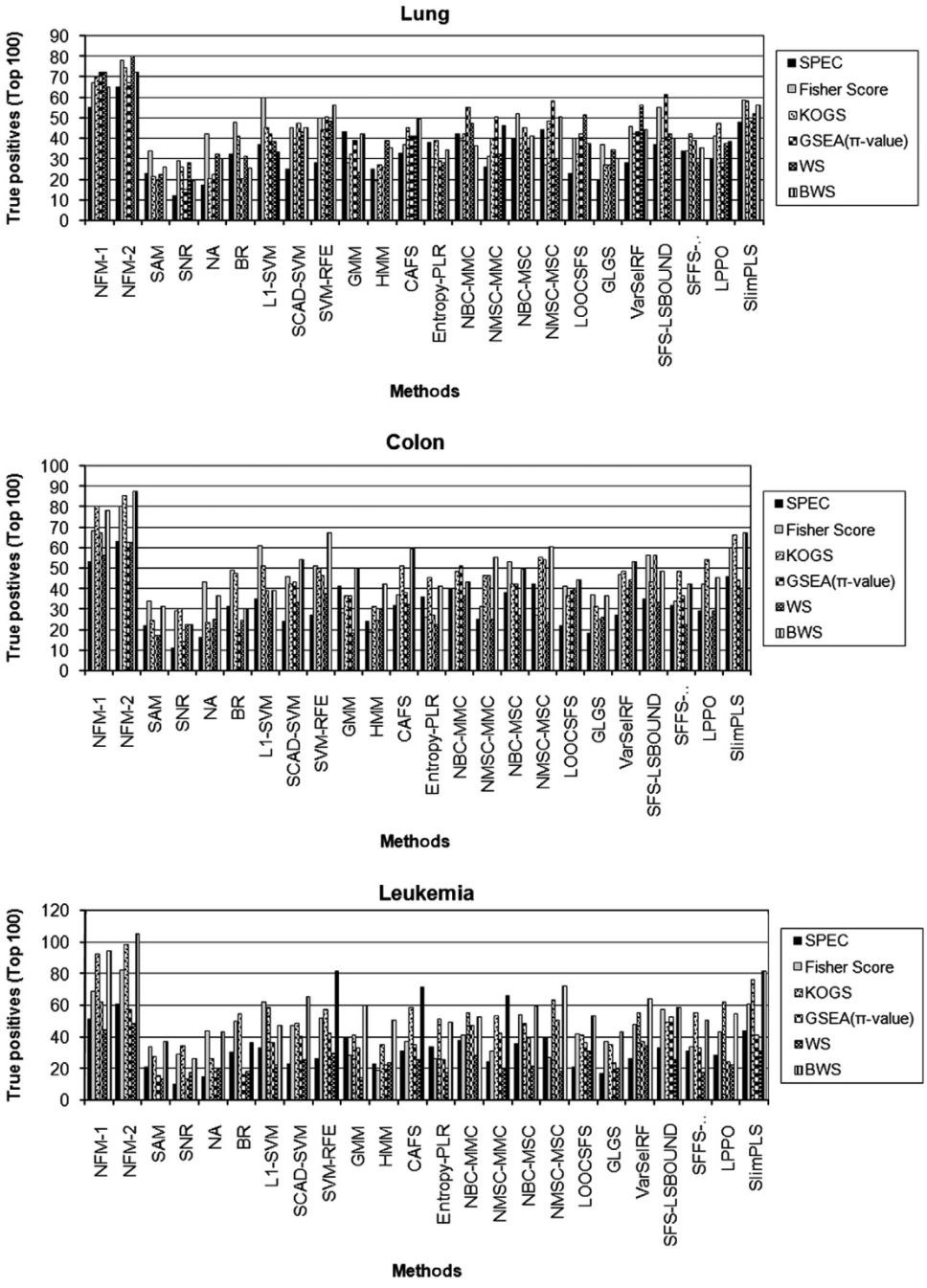

Fig. 7. Performance comparisons of different methods using various types of statistical and biological measurements.

level) genes like PFKP, TYMS, IARS and HLA-B are not present in the first 20 selected genes by these methods. This result suggests that NFM-1 and NFM-2 are able to find more true positive genes than the existing methods.

Like *t*-test, we have applied *F*-test on the genes identified by all the aforesaid methods. Applying NFM-1 and NFM-2 on human lung expression data, we have identified some important genes like CALCA (6.78), PFKP (7.65), TYMS (8.23), IGFBP3 (9.12), IARS (6.89), HBB (10.56), HLA-B (8.67), SFTPA2 (8.88), and TNF (9.45), IGHG3 (5.67), PRKACA (5.78), SORT1 (6.09), SFTPA1 (8.74), MEN1 (4.98) and IGHM (6.78). The number in the bracket indicates the $F_{0.01}-$value corresponding to the gene. It indicates that these genes are highly significant (more than 99% level of significance). Similarly, genes like RPLP0 (3.92), SMCIL1 (3.15), SFTPC (4.09), HLA-DRA (4.42), MGP (3.17) and RNASE1 (3.76) exceed the $F_{0.05}-$value. This means that these genes are significant at the level of 95%. We have performed *F*-test for the genes identified by the above gene selection algorithms. It is observed that some highly significant (more than 99% significance level) genes are not present in the first 20 selected genes by these methods. This result suggests that NFM-1 and NFM-2 are able to find more true positive genes than the existing methods. To validate our results we have made the comparative analysis in finding out the significant genes by each method for each data set. Due to the restriction of the manuscript size we have only mentioned the results on lung expression data set. The results are shown in Tables 6 and 7. It is to be noted that NFM-1 and NFM-2 are able to select more statistically significant genes (based on both *t*- and *F*-tests) compared to the existing methods.

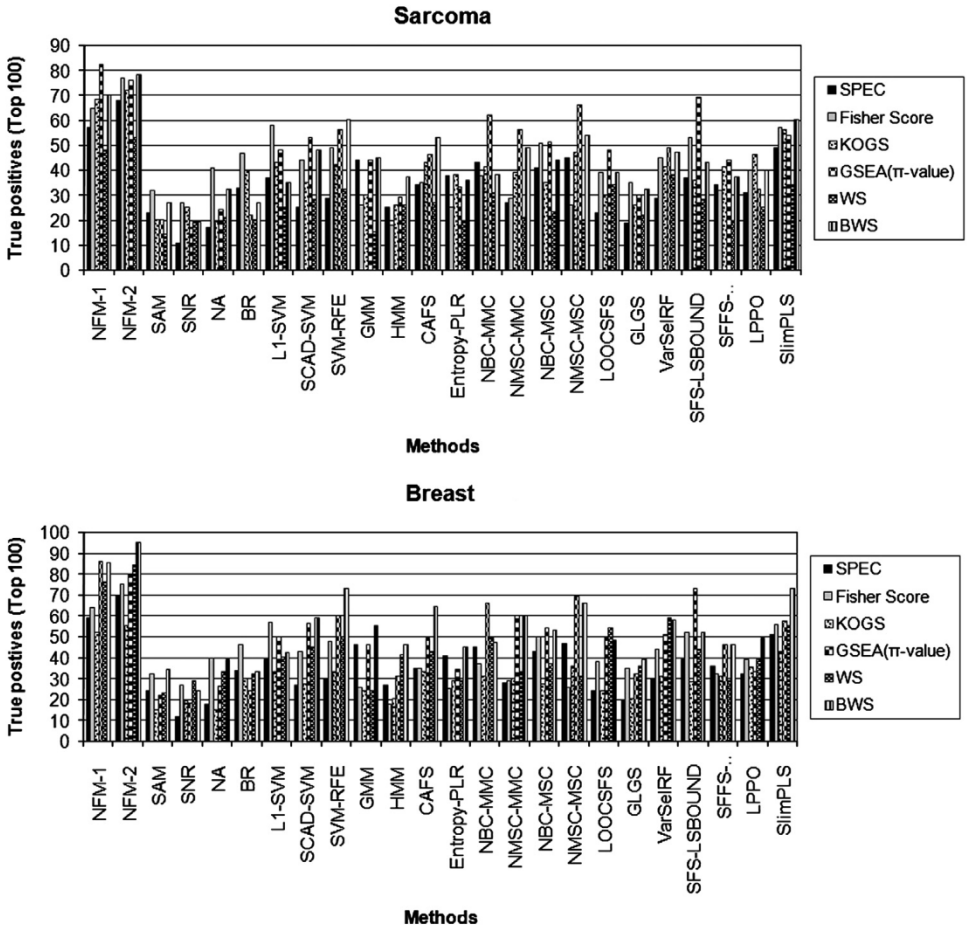

Fig. 8. Performance comparisons of different methods using various types of statistical and biological measurements.

6.1.4. Using NCBI database

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. From each gene list, we can consider the first c genes if a method results in c genes. The database results in 346, 210, 329, 483 and 124 genes for lung cancer, colon cancer, leukemia, breast cancer and sarcoma, respectively. For lung expression data (7129 genes), we have identified 346 genes each using NFM-1 and NFM-2. We have compared this set of genes with 346 genes from NCBI. Here we have identified 241 and 232 genes for NFM-1 and NFM-2, respectively, that are common in both the sets. We call these genes *true positive* (*TP*) genes. Thus, 105 ($=346-241$) and 114 ($=346-232$) genes for NFM-1 and NFM-2, respectively, are not in the list of genes obtained from NCBI. We denote these genes as *false positives* (*FP*). Likewise, 105 ($=346-241$) and 114 ($=346-232$) genes that are in the NCBI list are not in the set of genes obtained by NFM-1 and NFM-2 are called *false negative* (FN) genes. Similarly, we have compared our results with other methods, viz., SAM, SNR, BR, NA, SVM, GMM, HMM, CAFS, Entropy-PLR, NBC-MSC, NBC-MMC, NMSC-MSC, NMSC-MMC, LOOCSFS, GLGS, SFS-LSBOUND and SFFS-LSBOUND, VarSelRF and SlimPLS on lung expression data set as well as other four data sets. From Fig. 5, it is clear that SlimPLS produces better result than NFM-1 and NFM-2 for lung expression data. NFM-1 and NFM-2 have produced the same true positives along with SVM, CAFS, Entropy-PLR, NBC-MSC, VarselRF, SFS-LSBOUND, SFFS-LSBOUND, and LPPO for lung expression data set. However, NFM-1 and NFM-2 have performed the best along with SVM and SlimPLS for colon and leukemia data sets. Similarly, NFM-1 and NFM-2 have produced a similar performance along with SVM, GLGS and SlimPLS for sarcoma expression data. Lastly, NFM-1 and NFM-2 have performed the best along with NMC, Slimpls for breast expression data. Thus, NFM-1 and NFM-2 have produced the best results in terms of true positives, false positive and false negative, compared to the other existing methods along with SlimPLS, SVM for all the five data sets.

In order to validate our results further, we have computed *Sensitivity* on the gene expression data sets. First, we have calculated the number of true positives corresponding to each method for each data set. *Sensitivity* is computed using the following equations:

$$Sensitivity = \frac{TP}{(TP+FN)} \tag{12}$$

As a result, *Sensitivity* of NFM-1 and NFM-2 is much higher than that of the other existing methods. Fig. 6, it is clearly observed that NFM-1 and NFM-2 have performed the best along with L1-SVM, SCAD-SVM, SVM-RFE, LPPO and SlimPLS for lung expression data set. Similarly, NFM-1 and NFM-2 have produced a similar performance along with LPPO and SlimPLS for colon expression data set. However, NFM-1 and NFM-2 have performed the best along with L1-SVM, SCAD-SVM, SVM-RFE and SlimPLS. Likewise, NFM-1 and NFM-2 have produced a similar performance along with SCAD-SVM, SVM-RFE, GLGS and SlimPLS for sarcoma expression data set. It is to be noted that L1-SVM has produced the best result for sarcoma data set. Finally, NFM-1 and NFM-2 have performed the best along with NMC, Slimpls for breast expression data set. From Fig. 6, we can conclude that NFM-1 and NFM-2 have performed the best in terms of sensitivity

compared to SAM, SNR, NA, BR, SVM, GMM, HMM, CAFS, Entropy-PLR, NBC-MSC, NBC-MMC, NMSC-MSC, NMSC-MMC, LOOCSFS, GLGS, SFS-LSBOUND and SFFS-LSBOUND, VarSelRF and SlimPLS for all the five data sets.

6.1.5. Validation based on some earlier investigations

Applying NFM-1 and NFM-2 on human lung expression data, we have found some important genes like CALCA [40–42], TYMS [43], IGFBP3 [44,45], HLA-B [46,47], and HBB [48] that have a quite significant number of enriched attributes, and these genes have changed their expression level from normal to tumor samples. Genes like TNF [49–51], IGHG3 [53], SFTPA1 [54,52], and SFTPA2 [52] have changed their expression levels from normal lung samples to tumor samples quite significantly. Some earlier investigations also support this fact. Genes like PFKP and IARS have a quite significant number of enriched attributes, but there is no information in the literature to our knowledge about these genes. Applying existing methods on this data set, we have found a set of important genes (CALCA, IGFBP3, HBB, and SFTPA2) that are also present in the results of NFM-1 and NFM-2. Thus, we may conclude that genes like CALCA, IGFBP3, HBB, and SFTPA2 have a significant role in the development of lung adenocarcinoma. Due to the restriction of the manuscript size, we have provided only some of the investigations on lung adenocarcinoma (see Tables 6 and 7). Thus, the methodology developed in this article is able to select biologically more significant genes than the others. Similar findings have been obtained for the other data sets too.

Some of these genes like CALCA [40–42] HBB [48], IGFBP3 [44,45], TYMS [43], SFTPA1 [54,52], SFTPA2 [52], TNF [49–51], IGHG3 [53], and HLA-B [46,47] were already found to be responsible for lung adenocarcinoma by some earlier investigations. It is to be mentioned that the proposed methodology has found two genes PFKP and IARS whose number of functional attributes is quite significant but there is no information in the literature to our knowledge about them. This result suggests that the aforesaid genes may have impact on these diseases. These genes may be considered for further investigation.

6.1.6. Validation through expression profile plots

Here we consider some genes that are among the most significant top genes of our results. The expression values of these genes have changed significantly from normal samples to diseased samples. Applying the proposed methodology on human lung expression data, we report that genes like IGFBP3, PFKP, IARS, and TYMS, among the top 10 most important genes, have been over expressed in tumor samples. On the other hand, the expression value of gene HBB has reduced quite significantly in tumor samples. This gene is identified as an under expressed gene. In order to restrict the size for the manuscript, we have provided only the expression profile plots of some important genes in lung adenocarcinoma (Fig. 10). In the case of human colon expression profile, the genes like calcitonin (CALCA), colon carcinoma kinase-4 (CCK4), isoleucyl-tRNA synthetase (IARS), thymidylate syntase (TYMS), hemoglobin beta chain (HBB), tumor necrosis factor receptor (TNF), and insulin-like growth factor binding protein 6 (IGFBP6) have changed their expression values from normal colon samples to tumor ones. Among these genes, HBB, TNF, and IGFBP6 are down regulated. On the other hand, CALCA, CCK4, IARS, and TYMS have been identified as up regulated genes.

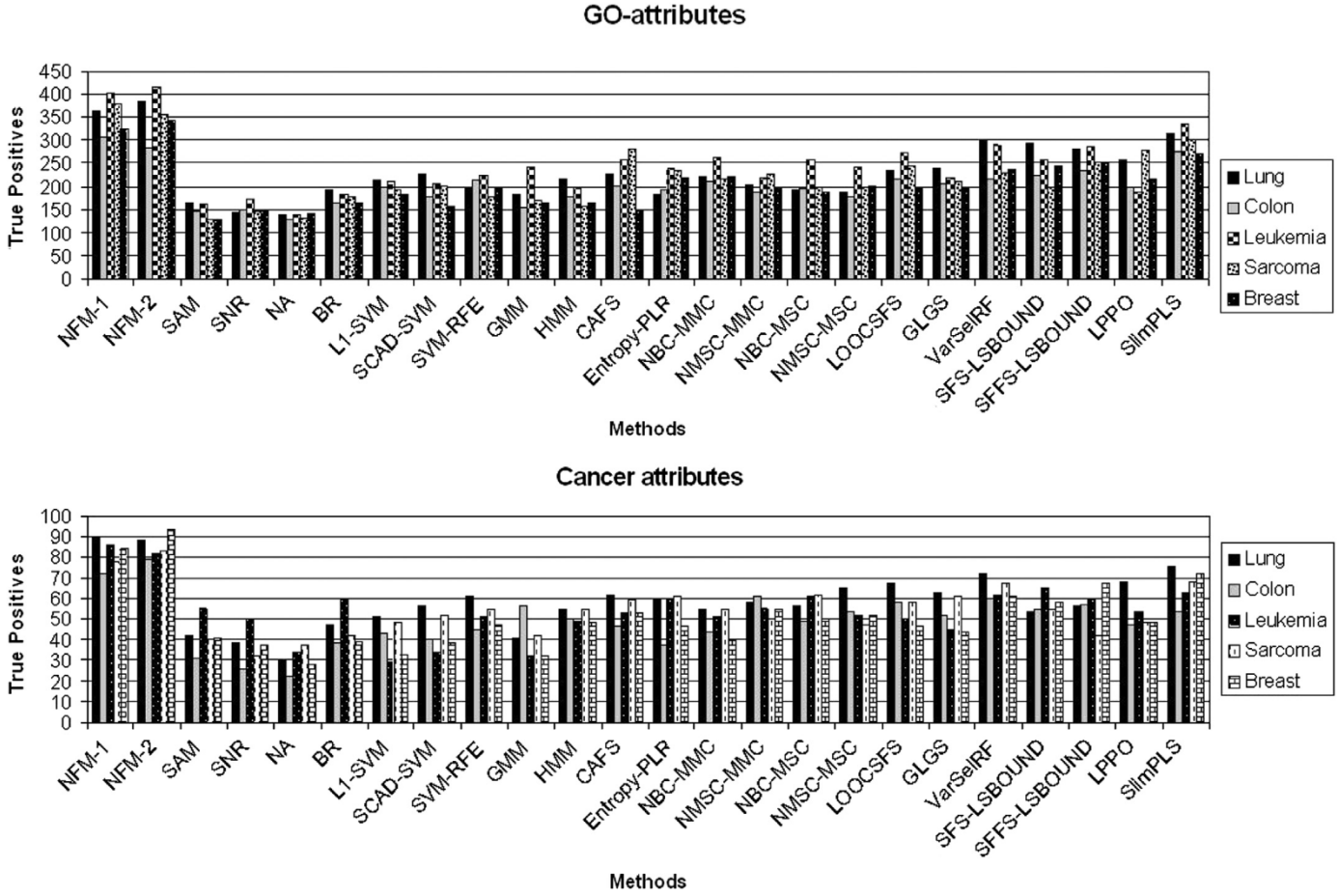

Fig. 9. Performance comparisons of NFM-1 and NFM-2 with other existing methods in terms of identification of GO attributes and Cancer attributes.

For human breast cell expression profile, we have observed that genes like BCAN, GDI2, ERBB2, and NARS (among the ten most important genes obtained by NFM-1 and NFM-2) change their expression levels quite significantly from normal breast mammary epithelial cell samples to breast cancer ones. The expression value of the gene BCAN has been increased in breast cancer cell lines whereas the expression values of genes GDI2 and NARS have been decreased drastically in breast cancer cell lines.

Similarly, for human soft tissue sarcoma expression data, genes like BRCA1, TYMS, IARS, and HBB have changed their expression values from normal tissue to sarcoma tissue. The expression value of the gene HBB drastically decreases in diseased sarcoma samples, whereas that of BRCA1, TYMS, and IARS increase in diseased samples. For human lymphocytes and plasma cell expression data, we report that expression values of the genes like BAX, CALCA, ATP6V0B, and NARS have changed significantly from normal B lymphocytes and plasma cells to macroglobulinemia, chronic lymphocytic leukemia, and multiple myeloma samples. It is to be mentioned that genes like BAX, NARS, and ATP6V0B have been over expressed in diseased samples, whereas the gene CALCA is under expressed.

6.1.7. *Validation through some recent statistical analysis*

To enrich our study, we have implemented various types of biological and statistical parameters like *pi*-GSEA, Fisher-score, KOGS, SPEC, *W*-test, and *BWS* for performance comparisons. The aforesaid methods have mostly been used and implemented in some recent investigations. This study will tell us the trueness of the gene set identified by NFM-1 and NFM-2 from biological and statistical points of view. For the comparison, we have taken top ranked 100 genes and then found out the true positives for each methods for all the data sets.

In gene set enrichment analysis (GSEA), a score that can combine fold change and *p*-value together is needed for better gene ranking [24]. In a gene functional study of cancer profiles, we are trying to validate the result using π-value based GESA. In human lung expression, leukemia and breast cancer data sets, it is clearly observed that NFM-1 and NFM-2 have performed the best with respect to all the other 22 methods to identify the top ranked biological and statistical relevant genes. For human colon expression data, NFM-1 and NFM-2 have performed best along with SlimPLS with respect to other existing methods. Likewise, NFM-1 and NFM-2 have performed best along with SFS-LSBOUND with respect to other existing methods for human sarcoma data set (Figs. 7 and 8).

We have implemented a generalized Fisher score [25] to validate the results of gene selection algorithms. NFM-1 and NFM-2 are the best to identify the top ranked biologically and statistically relevant genes under Fisher Score.

SPEC (spectral feature selection) is a general framework for feature selection [27]. It can generate a range of spectral feature selection algorithms for both unsupervised and supervised learning. NFM-1 and NFM-2 are the best to identify the top ranked biologically and statistically relevant genes under SPEC for all the data sets (Figs. 7 and 8).

Knowledge-oriented gene selection (KOGS) has been developed for systematically integrating different types of knowledge to

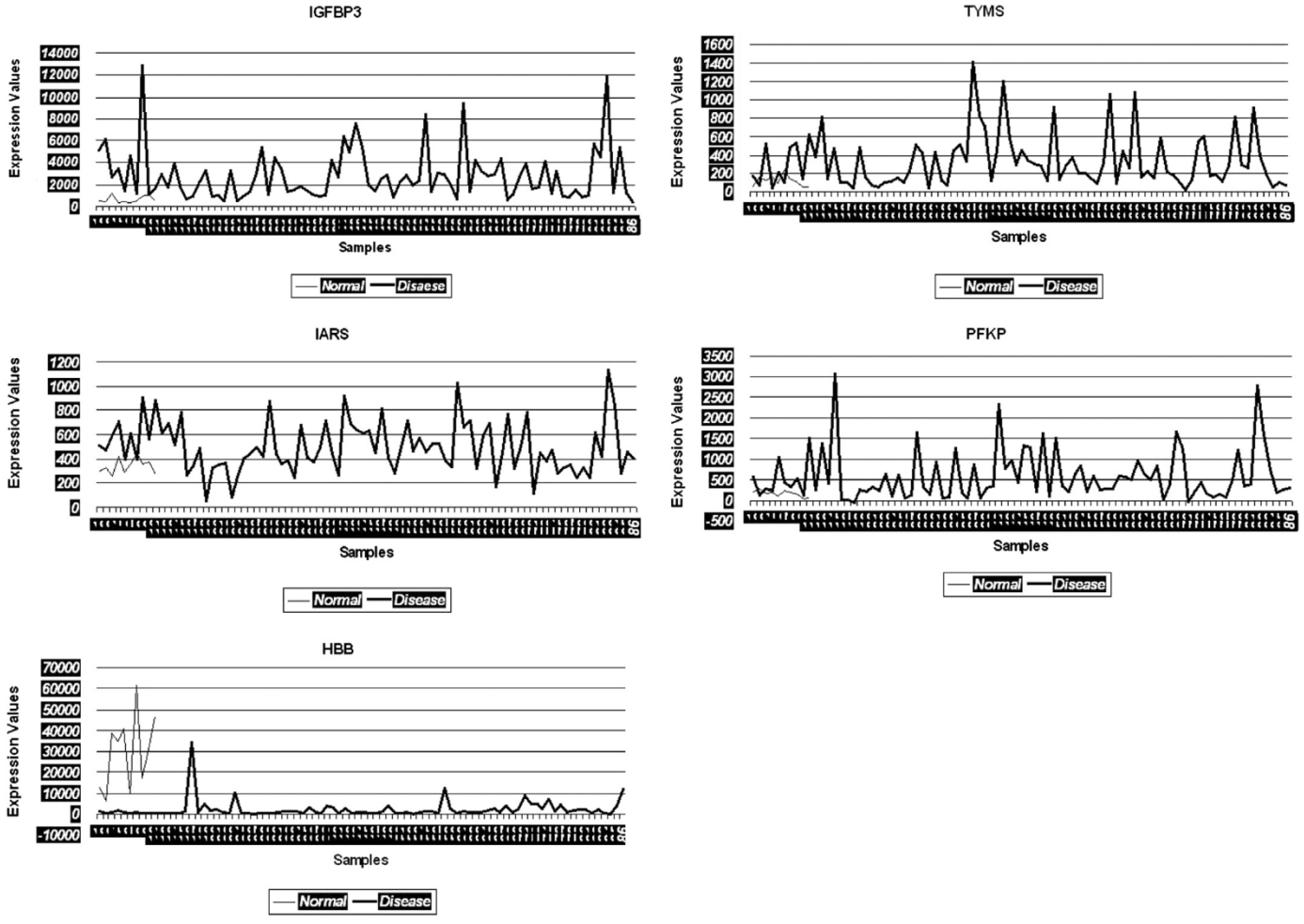

Fig. 10. Expression profiles of some over-expressed (IGFBP3, PFKP, IARS, and TYMS) and under-expressed (HBB) genes in normal samples represented by light line and tumor samples represented by bold line of human lung expression data with 10 normal and 86 disease samples.

achieve the ranking of biologically enriched genes [26]. KOGS integrates different types of knowledge such as the KEGG pathway repository and Gene Ontology database that could provide more information about genes and samples. The approach converts different types of external knowledge to its internal knowledge, which can be used to rank genes. Upon obtaining the ranking lists, it aggregates them via a probabilistic model and generates a final list. Under KOGS, NFM-1 and NFM-2 are the best with respect to other methods for all the gene expression data sets (Figs. 7 and 8).

The Wilcoxon rank sum non-parametric test (*W* test) is applicable when data coming from two independent samples can be converted to ordinal ranks [28,30]. Despite the fact that this conversion causes some loss of information, this method has the strength that does not make any assumption about the probability distribution from which the samples are taken. Under *W*-test, NFM-1 and NFM-2 are the best with respect to other methods for all the gene expression data sets to identify the biologically and statistically relevant genes (Figs. 7 and 8).

Baumgartner–Wei–Schindler non-parametric test (BWS test) makes the same assumptions about the samples as *W* test does [29]. However, it has probed to be less conservative and to yield better results. Neuhauser and Senske successfully applied this test for the detection of differentially expressed genes [30,31]. Under *BWS*-test, it is also observed that NFM-1 and NFM-2 are the best with respect to other methods for all the gene expression data sets to identify the biologically and statistically relevant genes (Figs. 7 and 8).

Thus, from Figs. 7 and 8, it is clearly observed that NFM-1 and NFM-2 have been generated more true positives with respect to all the existing methods for all the tests. Thus, we can say that NFM-1 and NFM-2 are able to identify the more biologically relevant and statically significant cancer mediating gene set from a gene expression data set.

7. Conclusions

In this article, we have provided a methodology, based on neuro-fuzzy models [5–7], for selection of genes whose over/under expression may cause diseases in general and various types of cancers in particular. The methodology, first of all, finds various groups of genes based on correlation values. This is followed by determining the most important group using two neuro-fuzzy systems, NFM-1 and NFM-2. The genes in this group have been evaluated further using NFM-1 and NFM-2. This results in important genes that may have a role in mediating the development of a particular cancer. The effectiveness of the methodology has been demonstrated on various gene expression data sets related various cancers where each gene is treated as a feature. The most important genes obtained by the methodology have also been verified using their *p*-values [39]. The superior performance of the methodology compared to some existing ones has been shown. The results have been verified using biochemical pathways, *t*-test, *F*-test, *sensitivity* and some existing results, expression profile plots and some biological and statistical measurements like *pi*-GSEA, Fisher-score, KOGS, SPEC, *W*-test, and *BWS*. It has been found that the methodology provided here has performed the best in identifying important genes in mediating various cancers than the other existing methods considered here.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.neucom.2013.11.023.

References

[1] I. Guyon, J. Weston, S. Barnhill, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.
[2] Z. Wang, V. Palade, Y. Xu, Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis, in: International Symposium on Evolving Fuzzy Systems, 2006.
[3] Z. Wang, V. Palade, Building interpretable fuzzy models for high dimensional data analysis in cancer diagnosis, BMC Genomics 12 (2011) S2–S5.
[4] S.K. Pal, S. Mitra, Neuro-fuzzy Pattern Recognition: Methods in Soft Computing, Wiley, New York, 1999.
[5] S.K. Pal, R.K. De, J. Basak, Unsupervised feature evaluation: a neuro-fuzzy approach, IEEE Trans. Neural Netw. 11 (2000) 366–376.
[6] R.K. De, J. Basak, S.K. Pal, Neuro-fuzzy feature evaluation with theoretical analysis, Neural Netw. 12 (1999) 1429–1455.
[7] J. Basak, R.K. De, S.K. Pal, Unsupervised feature selection using neuro-fuzzy approach, Pattern Recognit. Lett. 19 (1998) 997–1006.
[8] R.K. De, A. Ghosh, Neuro-fuzzy methodology for selecting genes mediating lung cancer, in: Proceedings of the 4th International Conference on Pattern Recognition and Machine Intelligence (PReMI-11), Moscow, June 27–July 1, 2011, pp. 388–393.
[9] S.K. Shevade, S.S. Keerthi, A simple and efficient algorithm for gene selection using sparse logistic regression, Bioinformatics 19 (2003) 2246–2253.
[10] G.C. Cawley, N.L.C. Talbot, Gene selection in cancer classification using sparse logistic regression with Bayesian regularization, Bioinformatics 22 (2006) 2348–2355.
[11] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 5116–5121.
[12] L. Goh, Q. Song, N. Kasabov, A novel feature selection method to improve classification of gene expression data, in: Asia Pacific Bioinformatics Conference, Dunedin, New Zealand, vol. 29, 2004, pp. 161–166.
[13] 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, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[14] H.H. Zhang, J. Ahn, X. Lin, C. Park, Gene selection using support vector machines with non-convex penalty, Bioinformatics 22 (2006) 88–95.
[15] K.Y. Yeung, C. Fraley, A.E. Raftery, W.L. Ruzzo, Model-based clustering and data transformations for gene expression data, J. Mol. Biol. 17 (2001) 977–987.
[16] L.R. Rabiner, B.H. Juang, An introduction to hidden Markov models, IEEE ASSP Mag. 3 (1986) 4–16.
[17] M. Kabir, M. Islam, M. Murase, A new wrapper feature selection approach using neural network, Neurocomputing 73 (2010) 3273–3283.
[18] H. Mahmoodian, M.H. Marhaban, R.A. Rahim, R. Rosli, I. Saripan, New entropy based method for gene selection, IETE J. Res. 55 (2009) 162–168.
[19] Q. Liu, A.H. Sung, Z. Chen, J. Liu, L. Chen, M. Qiao, Z. Wang, X. Huang, Y. Deng, Gene selection and classification for cancer microarray data based on machine learning and similarity measures, BMC Genomics 12 (2011) S1.
[20] E.K. Tang, K.N. Suganthan, X. Yao, Gene selection algorithms for microarray data based on least squares support vector machine, BMC Bioinform. 7 (2006).
[21] X. Zhou, K.Z. Mao, LS Bound based gene selection for DNA microarray data, Bioinformatics 21 (2005) 1559–1564.
[22] R. Díaz-Uriarte, S.A. Andres, Gene selection and classification of microarray data using random forest, BMC Bioinform. (2006), http://dx.doi.org/10.1186/1471-2105-7-3 (online).
[23] M. Gutkin, R. Shamir, G. Dror, SlimPLS: a method for feature selection in gene expression-based disease classification, PLoS ONE 4 (2009) e6416.
[24] Y. Xiao, T. Hsiao, U. Suresh, H.H. Chen, X. Wu, S.E. Wolf, Y. Chen, A novel significance score for gene selection and ranking, Bioinformatics, 2012, http://dx.doi.org/10.1093/bioinformatics/btr671 (published online).
[25] R.O. Duda, P.E. Hart, P.G. Stork, Pattern Classification, 2 ed., John Wiley and Sons, New York, 2001.
[26] Z. Zhaoy, J. Wangz, S. Sharmay, N. Agarwaly, H. Liuy, Y. Changz, An integrative approach to identifying biologically relevant genes (RECOMB 2009), in: Annual International Conference on Research in Computational Molecular Biology, 2009.
[27] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, in: International Conference on Machine Learning (ICML), 2007.
[28] D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, 2nd ed., Chapman & Hall, Boca Raton, 2000.
[29] W. Baumgartner, P. Wei, H. Schindler, A nonparametric test for the general two-sample problem, Biometrics 54 (1998) 1129–1135.
[30] M. Neuhauser, An exact two-sample test based on the Baumgartner–Wei–Schindler statistic and a modification of the Lepage's test, Commun. Stat. Theory Methods 29 (2000) 67–78.
[31] M. Neuhauser, R. Senske, The Baumgartner–Wei–Schindler test for the detection of differentially expressed genes in replicated microarray experiments, Bioinformatics 20 (2004) 3553–3564.
[32] A. Krogh, M. Brown, I.S. Mian, K. Sjolander, D. Haussler, Hidden Markov models in computational biology: applications to protein modeling, J. Mol. Biol. 235 (1994) 1501–1531.
[33] R.K. De, A. Ghosh, Interval based fuzzy systems for identification of important genes from microarray gene expression data: application to carcinogenic development, Int. J. Biomed. Inform. 42 (2009) 1022–1028.

[34] G.D. Beer, et al., Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat. Med. 8 (2002) 816–823.
[35] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, 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. U. S. A. 96 (1999) 6745–6750.
[36] B.H. Mecham, G.T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D. Z. Wetmore, T.J. Mariani, I.S. Kohane, Z. Szallasi, 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.
[37] K.Y. Detwiller, N.T. Fernando, N.H. Segal, S.W. Ryeom, P.A. D'Amore, 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.
[38] N.C. Gutierrez, E.M. Ocio, J. delas Rivas, P. Maiso, M. Delgado, E. Ferminan, M. J. Arcos, M.L. Sanchez, J.M. Hernandez, J.F.S. Miguel, Gene expression profiling of B lymphocytes and plasma cells from Waldenstrom's macroglobulinemia: comparison with expression patterns of the same cell counterparts from chronic lymphocytic leukemia, multiple myeloma and normal individuals, Leukemia 21 (2007) 541–549.
[39] D.W. Kim, K.H. Lee, D. Lee, Detecting clusters of different geometrical shapes in microarray gene expression data, Bioinformatics 21 (2005) 1927–1934.
[40] S. Amatschek, U. Koenig, H. Auer, P. Steinlein, M. Pacher, A. Gruenfelder, G. Dekan, S. Vogl, E. Kubista, K.H. Heider, C. Stratowa, M. Schreiber, W. Sommergruber, Tissue-wide expression profiling using cDNA subtraction and microarrays to identify tumor-specific genes, Cancer Res. 64 (2004) 556–844.
[41] A.M. Marchevsky, J. Tsou, I. Laird-Offring, Classification of individual lung cancer cell lines based on dna methylation markers: use of linear discriminant analysis and artificial neural networks, J. Mol. Diagn. 6 (2004) 28–36.
[42] A.K. Virmani, J.A. Tsou, K.D. Siegmund, L.Y. Shen, T.I. Long, P.W. Laird, A.F. Gazdar, I. A. Laird-Offringa, Hierarchical clustering of lung cancer cell lines using DNA methylation markers, Cancer Epidemiol. Biomark. Prev. 11 (2002) 291–297.
[43] Q. Shi, Z. Zhang, A.S. Neumann, G. Li, M.R. Spitz, Q. Wei, Case-control analysis of thymidylate synthase polymorphisms and risk of lung cancer, Carcinogenesis 26 (2005) 649–656.
[44] Y.S. Chang, L. Wang, D. Liu, L. Mao, W.K. Hong, F.R. Khuri, H.Y. 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.
[45] 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.
[46] V.D. Mottironi, S.M. Banks, B. Pollara, U.H. Rudofsky, HLA and survival in lung cancer, Clin. Immunol. Immunopathol. 45 (1987) 55–62.
[47] T. So, M. Takenoyama, M. Mizukami, Y. Ichiki, M. Sugaya, T. Hanagiri, K. Sugio, K. Yasumoto, Haplotype loss of HLA class I antigen as an escape mechanism from immune attack in lung cancer, Cancer Res. 65 (2005) 5945–5952.
[48] J.F. Morere, Role of epoetin in the management of anaemia in patients with lung cancer, Lung Cancer 46 (2004) 149–156.
[49] 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.
[50] 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.
[51] 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.
[52] 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.
[53] M. Remmelink, T. Mijatovic, A. Gustin, A. Mathieu, K. Rombaut, R. Kiss, I. Salmon, C. Decaestecker, Identification by means of cDNA microarray analyses of gene expression modifications in squamous non-small cell lung cancers as compared to normal bronchial epithelial tissue, Int. J. Oncol. 26 (2005) 247–258.
[54] 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.

Anupam Ghosh is an Assistant Professor of the Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India. He completed his Master of Science in Computer Science, and Master of Technology in Computer Science and Engineering in the years 2002 and 2004, respectively, from University of Calcutta, Kolkata, India. He received his Ph. D. (Engineering) degree in 2013 from Jadavpur University, Kolkata, India. His research interest includes bioinformatics, computational biology, systems biology, pattern recognition and soft computing.

Bibhas Chandra Dhara received Bachelor in Science (Honours in Mathematics) degree and Bachelor of Technology in Computer Science and Engineering from University of Calcutta, Kolkata, India in 1997 and 2000, respectively. He earned Master of Technology and Ph.D. both in Computer Science from Indian Statistical Institute, Kolkata, India, in 2002 and 2008, respectively. Currently, he is working as an Assistant Professor in the Department of Information Technology, Jadavpur University, Kolkata, India. His research area and interest include image processing, video processing, pattern recognition and audio processing.

Rajat K. De is a Professor of the Indian Statistical Institute, Kolkata, India. He completed his Bachelor of Technology in Computer Science & Engineering, and Master of Computer Science and Engineering in the years 1991 and 1993, from University of Calcutta, Kolkata, India, and Jadavpur University, Kolkata, India, respectively. He obtained his Ph.D. degree from the Indian Statistical Institute, Kolkata, India, in the year 2000. He was a Distinguished Postdoctoral Fellow at the Whitaker Biomedical Engineering Institute, the Johns Hopkins University, USA, during 2002–2003. He has published about 70 research articles in international journals, conference proceedings and in edited books to his credit. His research interest includes bioinformatics, computational biology, systems biology, pattern recognition and soft computing.