Significance and Functional Similarity for Identification of Disease Genes

Pradipta Maji and Ekta Shah

Abstract—One of the most significant research issues in functional genomics is insilico identification of disease related genes. In this regard, the paper presents a new gene selection algorithm, termed as SiFS, for identification of disease genes. It integrates the information obtained from interaction network of proteins and gene expression profiles. The proposed SiFS algorithm culls out a subset of genes from microarray data as disease genes by maximizing both significance and functional similarity of the selected gene subset. Based on the gene expression profiles, the significance of a gene with respect to another gene is computed using mutual information. On the other hand, a new measure of similarity is introduced to compute the functional similarity between two genes. Information derived from the protein-protein interaction network forms the basis of the proposed SiFS algorithm. The performance of the proposed gene selection algorithm and new similarity measure, is compared with that of other related methods and similarity measures, using several cancer microarray data sets.

Index Terms—Disease gene identification, microarray data analysis, feature selection, protein-protein interaction network

1 Introduction

Genetic diseases are mainly caused due to mutations in genes. New mutations or changes in the DNA, inherited genetic conditions, or some non-genetic causes in people generally lead to various forms of cancer, Alzheimer's disease, down syndrome, etc. As the name suggests, gene dysfunction is the major cause for such genetic diseases. Therefore, these genes are better known as disease genes [1]. Early detection of genetic diseases can help in increasing the survival chances of a patient, and may also aid in improving prognosis. It may also be helpful to upgrade the existing and develop advanced diagnostic tools.

Recent advancement in the field of biotechnology has been producing huge amount of data such as yeast two-hybrid system, protein complex, and gene expression profiles, which are being used in various studies to understand the function of disease genes [2]. Gene expression data is being widely used for the purpose of disease gene identification. Extensive study of the expression levels in particular cell types may give a perception about the propensity of a disease. For a particular set of genes, if a steady pattern in expression levels is observed between sick and control groups, it is very likely that the gene set would play a pathogenic role [1]. Different feature selection algorithms have been used to identify disease genes from microarray gene expression data [3], [4], [5], [6], [7], [8]. Most notable among them are *minimum redundancy-maximum relevance* (mRMR) criterion based gene selection algorithms [3], [9], [10]. The mRMR criterion, introduced in [3], selects a set of genes from microarray data by maximizing the relevance of the selected genes and minimizing the redundancy among them. While both relevance and redundancy of the selected genes are calculated using mutual information in [3], f-information is used in [9] for computing these two measures.

Recently, the theory of rough sets has gained popularity in selecting genes from microarray data [4], [5], [10], [11]. Meng et al. [5] used rough sets and neighborhood system to select genes for plant stress response, while the concept of rough hypercuboid approach has been used in [11] for gene selection from microarray data. In [10], a gene selection approach has been introduced, integrating the merits of fuzzy-rough sets and the mRMR criterion. On the other hand, the *maximum relevance-maximum significance* (MRMS) criterion has been introduced in [4] for gene selection, where the theory of rough sets is used to compute both relevance and significance measures. The MRMS criterion [4] selects a set of genes from microarray data by maximizing both relevance and significance of the selected genes.

The phenomenon of inter-protein interaction can be attributed to the fact that the *de novo* genes, which are correlated with the same disorder, tend to share the same functional features [12], [13]. Proteins produced by these genes, therefore, have a likelihood to interact with each other. Another symbolic characteristic of a disease associated gene is that its protein product is strongly linked to other disease-gene proteins. With these facts as building blocks, the protein-protein interaction (PPI) networks have been studied intensely to identify potential disease genes [14], [15], [16], [17]. Microarray data as well as the PPI network data have independently been used for identification of disease genes. However, the chance of identifying novel disease genes from such an analysis is quite scanty. In this regard, some integrated approaches have been developed that consider the gene expression data and PPI networks together for the task of disease gene identification [18], [19], [20].

• *The authors are with the Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India. E-mail: {pmaji, ekta_r}@isical.ac.in.*

Manuscript received 3 Dec. 2015; revised 21 June 2016; accepted 25 July 2016. Date of publication 3 Aug. 2016; date of current version 6 Dec. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2016.2598163

The target of identifying disease associated genes found a new vent with integrated approaches, where microarray technology is being used along with the PPI networks. These approaches presume that the proteins produced by disease genes tend to be associated with other differentially expressed genes in the PPI networks. A recent approach to prioritize cancer-associated genes by integrating PPI network and gene expression data was put forward by Wu et al. [18]. Another approach was proposed by Zhao et al. [21] to select disease genes. A similar approach was presented by Kong et al. [20], where a network constrained regularization method for linear regression analysis was used to select a set of significant genes. Jia et al. [19] put forward an algorithm, assembling genome wide association studies data into the network of protein interactions, for identifying genes associated with complex diseases. Li and Li [22] put forward another method that can identify disease associated genes using genomic and phenotypic data sets. Li et al. [23] designed another method for the construction of a gene-gene regulatory network and adopted a random walk with restart method to mine disease genes.

In [24] and [25], two similar approaches have been proposed for the selection of disease genes. A subset of genes is first selected from the gene expression data. The subset so created is further used for identification of genes lying midway on the path between two candidate genes, making use of the PPI network. While Li et al. [24] used the mRMR criterion [3] for gene selection, Paul and Maji [25] used the MRMS criterion [4]. However, none of the works reported above considered both gene expression and PPI network together at the time of disease gene selection. Recently, the RelSim algorithm [26], based on *maximum relevance-maximum functional similarity* (MRMFS) criterion, has been proposed to select disease genes from both microarray and PPI data. In this approach, the genes are selected by maximizing both relevance and functional similarity, where functional similarity between two genes are computed based on weighted PPI network. A similar work can be found in [27], where a PPI network with semantic similarity weights, generated using gene ontology terms, is used to compute functional similarity of microRNAs.

In this background, this paper introduces a new algorithm, termed as SiFS, to single out the disease associated genes. It aims to create the set of disease genes by maximizing both Significance and Functional Similarity among the members of the selected set. The proposed method strategically integrates the protein interaction network and expression profiles of genes. Mutual information is employed to compute the significance of the gene under consideration with respect to the already-selected genes, while functional similarity between two genes is calculated based on the PPI data. A new measure is introduced to quantify the functional similarity between a pair of genes. The performance of the proposed SiFS algorithm and the new similarity measure is demonstrated on several cancer data sets, along with a comparison to other related methods and similarity measures. An important observation is that the SiFS algorithm has been shown to be efficient in selecting significant and functionally similar genes from microarray data, and the gene subset so identified is shown to be well associated with corresponding disease. Thorough experimental study on several data sets substantiates the fact that the gene set selected by the proposed method contains more disease causing genes than those identified by the existing methods. The study also validates the proposed criterion used in the presented algorithm for disease gene identification. All the results indicate that the proposed method is pretty reassuring and may become a useful means for selecting disease genes.

2 PPIN Based Functional Similarity

Generally, genes having association with identical disorders, do have a tendency to share similar functional characteristics. The protein products of such genes also bear a strong likelihood of interacting with each other [12]. Thus, a major trait of a disease gene is its tendency of associating to other disease gene proteins. It has been observed that common biological functions are shared by proteins lying in close proximity within the network [28], [29] and such interactive neighbors are more likely to have identical biological function than non-interactive ones [30], [31]. The reason is that the query proteins and its interactive neighbors, generally, are involved in the same pathway or perform a specific function by forming a protein complex. Consequently, a measure is desired that can compute the inter-gene similarity efficiently.

Graph data structures are commonly used to represent PPI networks, where the proteins form the nodes of the graph and the interaction between the proteins forms the edge. The edges in these graphs are weighted, which depend mostly on experimental and predicted interaction information. Let the set of interacting neighbors or descendants of a candidate gene $\mathcal{A}_i$ be represented as $\mathcal{N}_i$ and $\omega_{ij} \in [0,1]$ be the weight value for an edge linking gene $\mathcal{A}_j \in \mathcal{N}_i$ to candidate gene $\mathcal{A}_i$. The information of PPI network can be used to obtain the set of descendants $\mathcal{N}_i$ of gene $\mathcal{A}_i$ and corresponding weight value ω_{ij}. Let the set of common successors to both genes $\mathcal{A}_i$ and $\mathcal{A}_k$ be denoted by $\mathcal{N}_{ik}$, that is, $\mathcal{N}_{ik} = \mathcal{N}_i \cap \mathcal{N}_k$. The functional similarity between two genes $\mathcal{A}_i$ and $\mathcal{A}_k$, having sets of descendant genes $\mathcal{N}_i$ and $\mathcal{N}_k$, respectively, is defined as follows:

$$\mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) = \frac{\sum\limits_{\mathcal{A}_j \in \mathcal{N}_{ik}} \min\{\omega_{ij}, \omega_{kj}\}}{\sqrt{\sum\limits_{\mathcal{A}_j \in \mathcal{N}_i} \omega_{ij} * \sum\limits_{\mathcal{A}_j \in \mathcal{N}_k} \omega_{kj}}}. \quad (1)$$

A careful analysis of the proposed measure illustrates the fact that if both the neighbors and corresponding edge weights of a pair of genes are same, then the functional similarity between them is maximum. On the contrary, a pair of genes, having no common neighbors, is functionally dissimilar. On the basis of these facts, some properties of the proposed similarity measure can be stated as follows:

1) $0 \leq \mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) \leq 1$.
2) $\mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) = 1$ iff two sets $\mathcal{N}_i$ and $\mathcal{N}_k$ contain an identical set of successors, that is, $\mathcal{N}_{ik} = \mathcal{N}_i = \mathcal{N}_k$, and weight value $\omega_{ij} = \omega_{kj}, \forall \mathcal{A}_j \in \mathcal{N}_{ik}$.
3) $\mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) = 0$ if and only if $\mathcal{N}_{ik} = \emptyset$.
4) $\mathcal{S}(\mathcal{A}_i, \mathcal{A}_k) = \mathcal{S}(\mathcal{A}_k, \mathcal{A}_i)$ (symmetric).

In this regard, it can be shown that if the weight value $\omega_{ij} \in \{0, 1\}$, then the proposed measure of similarity would reduce to

$$\tilde{\mathcal{S}}(\mathcal{A}_i, \mathcal{A}_k) = \frac{|\mathcal{N}_i \cap \mathcal{N}_k|}{\sqrt{|\mathcal{N}_i| * |\mathcal{N}_k|}}; \tag{2}$$

which is Cosine coefficient between $\mathcal{A}_i$ and $\mathcal{A}_k$ [32].

3 SiFS: Proposed Algorithm

Recent advancement in the field of biotechnology has produced large quantities of gene expression data, which is being extensively used in various studies to acquire an understanding of the processes governing the activities of disease-linked genes. A detailed study of gene expression profile data has demonstrated the fact that a gene, which shows uniform expression patterns in sick and control groups, has a higher probability to be a disease gene and plays a pathogenic role. The expression levels of such genes can be studied to learn about the traits of the disease. Conversely, genes having an association with the same disease have an affinity towards sharing common functional features and interacting with other disease-gene proteins.

In this regard, the paper introduces a new gene selection algorithm, termed as SiFS, coalescing carefully the benefits of PPI and gene expression data, to identify pleiotropic genes involved in the cell-level physiological mechanisms of the disease. The algorithm selects a subset $\mathbb{S}$ of disease genes from the set $\mathbb{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_m\}$ of m genes of a given microarray data, by escalating both significance and functional similarity of genes present in $\mathbb{S}$. Let $\mathbb{D}$ be the class label. Define $\sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j)$ as the significance of the gene $\mathcal{A}_j$ with respect to the set $\{\mathcal{A}_i, \mathcal{A}_j\}$, while $\mathcal{S}(\mathcal{A}_i, \mathcal{A}_j)$ as the functional similarity between two genes $\mathcal{A}_i$ and $\mathcal{A}_j$.

Definition 1. *The significance of a gene $\mathcal{A}_j$ with respect to the set $\{\mathcal{A}_i, \mathcal{A}_j\}$ can be defined as follows [33]:*

$$\sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j) = \gamma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}) - \gamma_{\mathcal{A}_i}(\mathbb{D}), \tag{3}$$

where $\gamma_{\mathcal{A}_i}(\mathbb{D})$ represents the relevance of the gene $\mathcal{A}_i$ with respect to the class label $\mathbb{D}$. So, the significance of a gene $\mathcal{A}_j$ is the observed difference in dependency when the gene $\mathcal{A}_j$ is omitted from the set $\{\mathcal{A}_i, \mathcal{A}_j\}$. A higher dependency change for a gene $\mathcal{A}_j$ being omitted represents a higher significance of the gene. A significance value of 0 renders the gene $\mathcal{A}_j$ dispensable.

So, the total significance of all the selected genes is

$$\mathcal{J}_{\text{signf}} = \sum_{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S}} \{\sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_i) + \sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j)\}, \tag{4}$$

while the total functional similarity among the selected genes is

$$\mathcal{J}_{\text{similarity}} = \sum_{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S}} \mathcal{S}(\mathcal{A}_i, \mathcal{A}_j). \tag{5}$$

Therefore, the task of creating a subset $\mathbb{S}$ of significant and functionally similar genes from the whole set $\mathbb{C}$ of m genes is analogous to maximize both $\mathcal{J}_{\text{signf}}$ and $\mathcal{J}_{\text{similarity}}$, that is, to maximize the objective function

$$\mathcal{J} = \alpha \mathcal{J}_{\text{signf}} + (1 - \alpha)\mathcal{J}_{\text{similarity}}, \tag{6}$$

where α is a weight parameter. To solve the aforementioned maximization problem, following greedy approach is used in the present study:

1) Initialize $\mathbb{C} \leftarrow \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_m\}, \mathbb{S} \leftarrow \emptyset$.
2) Compute relevance $\gamma_{\mathcal{A}_i}(\mathbb{D})$ for all genes $\mathcal{A}_i \in \mathbb{C}$.
3) Select the gene $\mathcal{A}_i$ with the highest relevance value, $\gamma_{\mathcal{A}_i}(\mathbb{D})$, as the most relevant gene. In effect, $\mathbb{S} = \mathbb{S} \cup \mathcal{A}_i$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_i$.
4) Repeat the following two steps until the desired cardinality for the selected gene set is obtained.
5) Compute the significance and similarity between every gene in the set $\mathbb{C}$ with respect to the already-selected genes of $\mathbb{S}$ using (3) and (1), respectively, and remove genes from $\mathbb{C}$ having either zero significance or zero similarity values with respect to any one of the selected genes of $\mathbb{S}$.
6) From the remaining genes of $\mathbb{C}$, select gene $\mathcal{A}_j$ that maximizes the following condition:

$$\frac{1}{|\mathbb{S}|} \sum_{\mathcal{A}_i \in \mathbb{S}} \{\alpha \sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j) + (1 - \alpha)\mathcal{S}(\mathcal{A}_i, \mathcal{A}_j)\}. \tag{7}$$

   As a result of that, $\mathbb{S} = \mathbb{S} \cup \mathcal{A}_j$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_j$.
7) Stop.

In the current study, the proposed similarity measure is used for computing the functional similarity between two genes, while mutual information is used to calculate the significance of a gene. Since the proposed SiFS algorithm aims to maximize both significance and functional similarity, the gene evaluation criterion used in the proposed SiFS method is termed as *maximum significance-maximum functional similarity* criterion.

3.1 Computation of Significance

The microarray data is used to compute the significance of a candidate gene $\mathcal{A}_j$ with respect to an already-selected gene $\mathcal{A}_i$ using (3). The significance of the gene $\mathcal{A}_j$, based on mutual information, is defined as follows:

$$\sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j) = \mathcal{I}(\{\mathcal{A}_i, \mathcal{A}_j\}, \mathbb{D}) - \mathcal{I}(\mathcal{A}_i, \mathbb{D}), \tag{8}$$

where $\mathbb{D}$ is the sample category or class label and $\mathcal{I}(\mathcal{X}, \mathcal{Y})$ represents the mutual information between two random variables $\mathcal{X}$ and $\mathcal{Y}$, which is given by

$$\mathcal{I}(\mathcal{X}, \mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}), \tag{9}$$

where $H(\mathcal{X})$ and $H(\mathcal{X}|\mathcal{Y})$ represent the entropy of the random variable $\mathcal{X}$ and conditional entropy of $\mathcal{X}$ given random variable $\mathcal{Y}$. Combining (8) and (9), we get

$$\sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j) = H(\mathbb{D}|\mathcal{A}_i) - H(\mathbb{D}|\{\mathcal{A}_i, \mathcal{A}_j\}). \tag{10}$$

So, the significance of a gene $\mathcal{A}_j$ with respect to another gene $\mathcal{A}_i$ is the change in conditional entropy of class label $\mathbb{D}$ given gene $\mathcal{A}_i$ and the set $\{\mathcal{A}_i, \mathcal{A}_j\}$. In microarray data, the expression values of genes are continuous, while class label

$\mathbb{D}$ is represented by discrete symbol. Let us assume that $\mathbb{D}$ (discrete random variable) has C alphabets (number of classes) and $c \in C$. The conditional entropy $H(\mathbb{D}|\mathcal{A}_i)$ of $\mathbb{D}$ given gene $\mathcal{A}_i$ (random variable) is defined as follows:

$$H(\mathbb{D}|\mathcal{A}_i) = -\int_{\mathcal{A}_i} \mathrm{p}(w_i) \sum_{c \in C} \mathrm{p}(c \,|\, w_i) \log \mathrm{p}(c \,|\, w_i)\, dw_i, \tag{11}$$

where $\mathrm{p}(w_i)$ is the true probability density function of the gene $\mathcal{A}_i$ and $\mathrm{p}(c \,|\, w_i)$ represents the conditional probability density function of $\mathbb{D}$ given gene $\mathcal{A}_i$. By using Bayesian rule, the conditional probability $\mathrm{p}(c \,|\, w_i)$ can be calculated as follows:

$$\mathrm{p}(c|w_i) = \frac{\mathrm{p}(w_i|c)\mathrm{p}(c)}{\mathrm{p}(w_i)}, \tag{12}$$

where the probability density function $\mathrm{p}(c)$ is given by

$$\mathrm{p}(c) = \Pr\{\mathbb{D} = c\}, c \in C. \tag{13}$$

Similarly, the conditional entropy $H(\mathbb{D}|\{\mathcal{A}_i, \mathcal{A}_j\})$ of $\mathbb{D}$ given the gene set $\{\mathcal{A}_i, \mathcal{A}_j\}$ is as follows:

$$\begin{aligned} H(\mathbb{D}|\{\mathcal{A}_i, \mathcal{A}_j\}) = -\int_{\mathcal{A}_i}\int_{\mathcal{A}_j} \mathrm{p}(w_i, w_j) \\ \times \sum_{c \in C} \mathrm{p}(c|\{w_i, w_j\}) \log \mathrm{p}(c|\{w_i, w_j\})\, dw_i\, dw_j, \end{aligned} \tag{14}$$

where $\mathrm{p}(w_i, w_j)$ is the true joint probability density function of two genes $\mathcal{A}_i$ and $\mathcal{A}_j$, and the conditional probability $\mathrm{p}(c|\{w_i, w_j\})$ can be computed as follows:

$$\mathrm{p}(c|\{w_i, w_j\}) = \frac{\mathrm{p}(\{w_i, w_j\}|c)\mathrm{p}(c)}{\mathrm{p}(w_i, w_j)}. \tag{15}$$

The Parzen window density estimator is used to approximate the true probability density of the genes having continuous expression values [34]. It involves the superposition of a normalized window function centered on a set of random samples. Given a set of n d-dimensional samples of a variable $\mathcal{X}$, the approximate density function $\hat{\mathrm{p}}(x)$ has the following form [35]:

$$\hat{\mathrm{p}}(x) = \frac{1}{n}\sum_{k=1}^{n} \delta(x - x_k, h), \tag{16}$$

where $\delta(\cdot)$ is the Parzen window function, x_k is the kth sample, and h is the window width parameter. The window function is required to be a finite-valued non-negative density function [34], where

$$\int_{-\infty}^{\infty} \delta(z, h)\, dz = 1. \tag{17}$$

So, the window function $\delta(\cdot)$ is chosen as the Gaussian window, which is given by

$$\delta(z, h) = \frac{1}{(2\pi)^{d/2} h^d |\Sigma|^{1/2}} \exp\left(-\frac{z^T \Sigma^{-1} z}{2h^2}\right), \tag{18}$$

where $z = (x - x_k)$, d is the dimension of the sample x, and Σ is the covariance matrix of a d-dimensional vector of random variable $\mathcal{Z}$. When $d = 1$, (18) returns the estimated marginal density; when $d = 2$, (18) can be used to estimate the density $\mathrm{p}(x, y)$ of bivariate variable $(\mathcal{X}, \mathcal{Y})$, which is actually the joint density of $\mathcal{X}$ and $\mathcal{Y}$.

Hence, the approximate density function for the gene $\mathcal{A}_i$ with continuous expression values is as follows:

$$\hat{\mathrm{p}}(w_i) = \frac{1}{n\sqrt{2\pi h^2 \Sigma_i}} \sum_{k=1}^{n} \exp\left\{-\frac{(w_i - w_{ik})^2}{2h^2\Sigma_i}\right\}, \tag{19}$$

where $w_{ik} \in \Re$ is the measured expression level of gene $\mathcal{A}_i$ in the kth sample and Σ_i represents the variance of continuous expression values of gene $\mathcal{A}_i$. Similarly, the approximate joint density function between two genes $\mathcal{A}_i$ and $\mathcal{A}_j$ is given by

$$\begin{aligned} \hat{\mathrm{p}}(w_i, w_j) = \frac{1}{2\pi n h^2 \sqrt{\Sigma_i \Sigma_j}} \\ \times \sum_{k=1}^{n} \exp\left\{-\frac{1}{2h^2}\left\{\frac{(w_i - w_{ik})^2}{\Sigma_i} + \frac{(w_j - w_{jk})^2}{\Sigma_j}\right\}\right\}. \end{aligned} \tag{20}$$

In microarray data, since the sample category is discrete valued, the approximate conditional probability density function $\hat{\mathrm{p}}(w_i|c)$ of the gene $\mathcal{A}_i$ given the class label $\mathbb{D}$ is as follows:

$$\hat{\mathrm{p}}(w_i|c) = \frac{1}{n_c\sqrt{2\pi h_c^2 \Sigma_c}} \sum_{w_{ik} \in c} \exp\left\{-\frac{(w_i - w_{ik})^2}{2h_c^2\Sigma_c}\right\}, \tag{21}$$

where n_c is the number of samples belonging to class c, h_c is the class specific window width, and Σ_c represents the variance of expression values of gene $\mathcal{A}_i$ that belong to class c. Similarly, the approximate conditional density function $\hat{\mathrm{p}}(\{w_i, w_j\}|c)$ of the gene set $\{\mathcal{A}_i, \mathcal{A}_j\}$ given the sample category $\mathbb{D}$ is as follows:

$$\begin{aligned} \hat{\mathrm{p}}(\{w_i, w_j\}|c) = \frac{1}{2\pi n_c h_c^2 \sqrt{\Sigma_{ic} \Sigma_{jc}}} \\ \times \sum_{w_{ik}, w_{jk} \in c} \exp\left\{-\frac{1}{2h_c^2}\left\{\frac{(w_i - w_{ik})^2}{\Sigma_{ic}} + \frac{(w_j - w_{jk})^2}{\Sigma_{jc}}\right\}\right\}. \end{aligned} \tag{22}$$

Also, the approximate density function $\hat{\mathrm{p}}(c)$ is computed as n_c/n. The value of h is determined as [36]:

$$h = \left\{\frac{4}{d+2}\right\}^{\frac{1}{d+4}} \times n^{-\frac{1}{d+4}}, \tag{23}$$

where $d = 1$ and 2 for the estimation of marginal density and joint density, respectively. Hence, to compute the significance of a gene using mutual information, (19), (20), (21), and (22) are used to approximate the required marginal and joint distributions.

3.2 Computation of Functional Similarity

The proposed similarity measure, based on the information of PPI network, is used for computing functional similarity between two genes. The weighted PPI network is retrieved from Search Tool for the Retrieval of Interacting Genes (STRING) [37], a large online database resource providing both experimental and predicted protein interactions, along

with a confidence score. The nodes in the graph correspond to proteins, while edge denotes the association between two proteins. Experimental repositories, computational prediction methods, etc., have been used to derive direct and indirect interactions. The confidence in any interaction represents the probability with which a pair of nodes may be present together in a metabolic process. The confidence in the interaction between two proteins is used as the weightage of each interaction in the current study.

3.3 Optimum Value of Weight Parameter

The weight parameter α in (7) regulates the relative importance of significance and functional similarity of the candidate gene with respect to the already-selected genes. If $\alpha = 1$, only the significance of the candidate gene is considered for each gene selection. The value of $\alpha < 1$ is crucial in order to obtain good results. If the functional similarity between two genes is not taken into account, selecting the genes with the highest significance may tend to produce a set of functionally dissimilar genes that may leave out useful complementary information. On the other hand, if $\alpha = 0$, the genes are selected based on their similarity values only without considering the significance of each gene. In effect, the selected gene set may contain a number of insignificant genes. Hence, the value of parameter α should be in between zero and one in order to obtain good results, that is, $0 < \alpha < 1$.

To find out the optimum value of α for a given data set, the class separability index $\mathcal{S}$ [35] is used. The $\mathcal{S}$ index of a data set is defined as $\mathcal{S} = \mathrm{trace}(V_B^{-1}V_W)$, where V_W is the within-class scatter matrix and V_B is the between-class scatter matrix, defined as follows:

$$V_W = \sum_{c \in C} \mathrm{p}(c) E\{(X - \mu_c)(X - \mu_c)^T | c\} = \sum_{c \in C} \mathrm{p}(c)\Sigma_c;$$
$$V_B = \sum_{c \in C} \mathrm{p}(c)(\mu_c - \bar{\mu})(\mu_c - \bar{\mu})^T;\ \bar{\mu} = \sum_{c \in C} \mathrm{p}(c)\mu_c,$$

where C is the number of classes, $\mathrm{p}(c)$ is *a priori* probability that a pattern belongs to class c, X is a feature vector, $\bar{\mu}$ is the sample mean vector for the entire data points, μ_c and Σ_c represent sample mean and covariance matrix of class c, respectively, and $E\{\cdot\}$ is the expectation operator. A lower value of $\mathcal{S}$ ensures that classes are well separated by their scatter means. For each data set, the value of α is varied from 0.0 to 1.0, and both $\mathcal{S}$ index and change in $\mathcal{S}$ index, that is $\Delta\mathcal{S}(\alpha)$, are computed, where

$$\Delta\mathcal{S}(\alpha) = |\,\mathcal{S}(\alpha) - \mathcal{S}(\alpha - 1)\,|. \tag{24}$$

The optimum value of α, that is $\alpha^\star$, is obtained using the following relation:

$$\alpha^\star = \arg\max_{\alpha}\{\mathcal{S}^{-1}(\alpha) \times \Delta\mathcal{S}(\alpha)\}. \tag{25}$$

3.4 Computational Complexity

The proposed SiFS algorithm has low computational complexity with respect to the number of genes present in microarray data. The computation of relevance of m genes is carried out in step 2 of the proposed algorithm, which has a time complexity of $\mathcal{O}(m)$. The selection of most relevant gene from the set of m genes, which is carried out in step 3, has also a complexity of $\mathcal{O}(m)$. There is only one loop in step 4 of the proposed algorithm, which needs to be executed $(d-1)$ times, where d is the desired number of genes to be selected. The computation of significance of a candidate gene with respect to the already-selected gene set takes only a constant amount of time. If $\tilde{m}$ represents the cardinality of the already-selected gene set, then total complexity to compute the significance of $(m - \tilde{m})$ genes, which is carried out in step 5 of the algorithm is $\mathcal{O}(m - \tilde{m})$. The computation of functional similarity of a candidate gene with respect to a gene in the already-selected set of genes takes $\mathcal{O}(n_0^2)$ time, where n_0 is the average number of neighbors to a protein in the PPI network. The total complexity to compute the functional similarity of $(m - \tilde{m})$ candidate genes, which is carried out in step 5, is $\mathcal{O}((m - \tilde{m})n_0^2)$. The selection of a gene from $(m - \tilde{m})$ candidate genes by maximizing both significance and functional similarity, which is carried out in step 6, has also a complexity $\mathcal{O}(m - \tilde{m})$. In effect, the selection of a set of d significant and functionally similar genes from the whole set of m genes using the proposed SiFS algorithm has an overall computational complexity of $\mathcal{O}(m) + \mathcal{O}(m) + \mathcal{O}(d(m - \tilde{m})n_0^2) = \mathcal{O}(mn_0^2)$ since $d, \tilde{m} << m$.

4 Experimental Results and Discussion

This section presents the performance of the proposed SiFS algorithm, which is based on the MSMFS criterion. The performance of the proposed MSMFS criterion is compared with that of several other criteria, namely, *maximum relevance* (MR), *maximum significance* (MS) [4], *maximum functional similarity* (MFS), mRMR [3], MRMS [4], and MRMFS [26]. The proposed similarity measure is also compared with other similarity measures. Finally, the performance of the proposed SiFS algorithm is compared with that of some existing methods for disease gene selection, namely, t-test, MR+PPIN [1], mRMR+PPIN [24], MRMS+PPIN [25], RelSim [26], and CLAIM [38]. To evaluate the performance of different methods, breast cancer, leukemia, and three colon cancer microarray data sets, namely, GSE25070, GSE10950, and GSE11223, are used. The values of $\alpha^\star$, obtained using (25), are 0.5, 0.4, 0.7, 0.4, and 0.2 for GSE25070, GSE10950, GSE11223, breast cancer, and leukemia data, respectively. The software version of SiFS algorithm is available at www.isical.ac.in/~bibl/results/sifs/sifs.html.

4.1 Importance of MSMFS Criterion

The proposed SiFS method is based on the MSMFS criterion. In order to establish the importance of the MSMFS criterion over other criteria, namely, MR, MS, MRMS, mRMR, MFS, and MRMFS, extensive experiments are carried out and the comparative results are reported in Tables 1, 2, and 3. While the gene selection using MR, MS, MRMS, and mRMR criteria is based on gene expression data alone and the MFS criterion uses only PPI network data, the MRMFS and MSMFS criteria are based on both gene expression and PPI network data. The results corresponding to t-test are also reported in these tables for comparative analysis as t-test based gene selection algorithm uses the MR criterion.

The performance of various gene selection criteria is first evaluated for three colon cancer data sets, namely, GSE25070,

TABLE 1
Degree of Overlapping with Known Gene Lists and P-Value for Genes Identified by Different Criteria

Data Sets	Criteria	LIST 1:	Overlap and P-Value	LIST 2:	Overlap and P-Value	LIST 3:	Overlap and P-Value	LIST 2-3	LIST 1-2-3
GSE25070	*t*-test	3	7.71E-01	32	2.68E-29	6	3.50E-05	33	36
	MR	6	2.13E-01	24	5.50E-19	9	1.09E-08	25	30
	MS	8	4.78E-02	31	6.21E-28	11	2.66E-11	32	38
	mRMR	5	3.74E-01	22	1.22E-16	10	5.70E-10	23	27
	MRMS	6	2.13E-01	25	3.39E-20	10	5.70E-10	26	31
	MFS	9	1.90E-02	5	6.00E-02	6	3.50E-05	10	19
	MRMFS	15	1.04E-05	29	2.89E-25	13	4.23E-14	34	45
	MSMFS	12	6.59E-04	31	6.21E-28	17	3.62E-20	36	45
GSE10950	*t*-test	4	5.74E-01	1	8.81E-01	1	4.60E-01	2	6
	MR	5	3.74E-01	2	6.25E-01	1	4.60E-01	3	8
	MS	12	6.59E-04	4	1.60E-01	3	2.36E-02	5	17
	mRMR	8	4.78E-02	2	6.25E-01	0	1.00E+00	2	10
	MRMS	6	2.13E-01	2	6.25E-01	0	1.00E+00	2	8
	MFS	19	1.45E-08	5	6.00E-02	5	3.77E-04	9	28
	MRMFS	15	1.04E-05	4	1.60E-01	7	2.74E-06	9	24
	MSMFS	19	1.45E-08	6	1.90E-02	8	1.85E-07	11	30
GSE11223	*t*-test	4	5.74E-01	8	1.24E-03	1	4.60E-01	9	13
	MR	8	4.78E-02	5	6.00E-02	3	2.36E-02	7	15
	MS	15	1.04E-05	4	1.60E-01	4	3.34E-03	6	21
	mRMR	4	5.74E-01	5	6.00E-02	3	2.36E-02	7	11
	MRMS	9	1.90E-02	5	6.00E-02	3	2.36E-02	7	16
	MFS	15	1.04E-05	6	1.90E-02	7	2.74E-06	10	25
	MRMFS	19	1.45E-08	8	1.24E-03	8	1.85E-07	11	30
	MSMFS	21	3.65E-10	7	5.19E-03	7	2.74E-06	10	31

GSE10950, and GSE11223, on the basis of the overlap level with three gene lists, namely, LIST 1, LIST 2, and LIST 3. The LIST 1 is a gene set of 742 genes those are related to different forms of cancer. This list is a compilation of genes obtained from the Cancer Gene Census of the Sanger Centre, Atlas of Genetics and Cytogenetic in Oncology [39], and Human Protein Reference Database [40]. Both LIST 2 and LIST 3 are sets of genes known to be associated with colon cancer. While LIST 2 is a collection of 438 genes prepared by Sabates-Bellver et al. [41], LIST 3 comprises of 134 genes, prepared by Nagaraj and Reverter [42]. LIST 2 and LIST 3 are merged to form a single list, named LIST 2-3, while all three lists are combined to form LIST 1-2-3.

Table 1 compares the performance of different criteria with respect to the degree of overlapping with three cancer gene lists, namely, LIST 1, LIST 2, and LIST 3. For three colorectal cancer data sets, 100 top-ranked genes are selected using different gene selection criteria for further analysis. From the results reported in Table 1, it can be seen that the degree of overlapping of the MSMFS criterion is better than that of the MRMFS criterion in six cases out of total nine cases for colorectal cancer gene lists such as LIST 2, LIST 3, and LIST 2-3. Also, the proposed MSMFS criterion attains maximum degree of overlapping in 10 cases, out of total 15 cases, irrespective of the microarray data, gene lists, and criteria used. Table 1 also represents the statistical significance test of the gene sets selected by different criteria with respect to the genes of LIST 1, LIST 2, and LIST 3. Using the Fisher's exact test, statistical analysis of the overlapped genes is done. Results reported in Table 1 confirm that the MSMFS criterion provides statistically significant results for all the cases and is better for disease gene selection as compared to other existing criteria.

Tables 2 and 3 compare the performance of different criteria with respect to gene ontology (GO), KEGG pathways, and disease ontology (DO). The biological significance for the generated set of genes is analyzed using the ClueGO v1.8 [43]. It combines the KEGG/BioCarta pathways with the GO terms, and creates a GO/pathway term network. The ClueGO computes enrichment score for the GO terms and pathways identified, based on hypergeometric distribution. The significance of any GO term/pathway to a group of genes is represented using the corrected p-value. A lower p-value is used to denote a higher significance of annotated term. While Table 2 compares the performance of different criteria with respect to gene products in terms of their associated biological processes, along with respective p-values, Table 3 presents KEGG pathway enrichment and DO based analysis of the obtained gene sets for different criteria on five data sets. The DO aims to provide an open source ontology for the integration of biomedical data that is associated with human disease. The DOSE package in R is used to perform the DO based analysis [44].

All the results reported in Table 2 establish the fact that the set of genes selected by the MSMFS criterion annotates to a term more significantly as compared to other criteria. It can be seen that the MSMFS criterion gives the lowest p-value, annotating to the biological process *"response to lipid"* for GSE25070 data, representing the fact that the gene set obtained using the MSMFS criterion plays a role in regulating the activity initiated due to the presence of lipids. Several studies exist that mark the importance of lipid metabolism in cancer cell growth, proliferation, differentiation and motility. Increased level of lipids in liver and muscles increases the intracellular diacylglycerol and ceramide content, which further impairs insulin signaling, leading to increased insulin

TABLE 2
Gene Ontology Based Analysis for Genes Identified by Different Criteria

Data Sets	Criteria	Biological Processes: Term and P-Value	
GSE25070	*t*-test	one-carbon metabolic process	5.42E-07
	MR	one-carbon metabolic process	5.69E-06
	MS	negative regulation of locomotion	2.00E-05
	mRMR	collagen catabolic process	7.51E-03
	MRMS	one-carbon metabolic process	1.62E-04
	MFS	response to lipopolysaccharide	1.10E-11
	MRMFS	TNF signaling pathway	1.19E-06
	MSMFS	response to lipid	2.00E-23
GSE10950	*t*-test	positive regulation of Notch signaling pathway	4.06E-03
	MR	regulation of systemic arterial blood pressure	2.75E-03
	MS	cellular extravasation	7.61E-05
	mRMR	female sex differentiation	1.58E-03
	MRMS	regulation of systemic arterial blood pressure	7.17E-03
	MFS	positive regulation of cell proliferation	1.16E-23
	MRMFS	response to cytokine	5.36E-24
	MSMFS	response to cytokine	3.54E-36
GSE11223	*t*-test	chromosome separation	7.06E-03
	MR	intrinsic apoptotic signaling pathway in response to oxidative stress	6.26E-04
	MS	intrinsic apoptotic signaling pathway in response to oxidative stress	5.58E-04
	mRMR	pyrimidine-containing compound catabolic process	1.40E-03
	MRMS	spleen development	3.43E-03
	MFS	immune response	1.14E-54
	MRMFS	response to cytokine	4.57E-31
	MSMFS	response to cytokine	2.83E-32
Breast	*t*-test	intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress	7.29E-05
	MR	response to amino acid	8.99E-07
	MS	response to vitamin	2.94E-07
	mRMR	response to amino acid	4.25E-08
	MRMS	response to amino acid	1.37E-05
	MFS	positive regulation of immune system process	2.37E-35
	MRMFS	regulation of immune response	1.47E-33
	MSMFS	positive regulation of immune system process	6.84E-33
Leukemia	*t*-test	myeloid dendritic cell activation	3.02E-05
	MR	myeloid leukocyte activation	1.88E-05
	MS	myeloid leukocyte activation	3.00E-05
	mRMR	chaperone-mediated autophagy	6.41E-05
	MRMS	myeloid leukocyte activation	1.79E-05
	MFS	response to molecule of bacterial origin	1.12E-20
	MRMFS	leukocyte activation	9.65E-20
	MSMFS	leukocyte activation	5.93E-22

secretion and increased IGF1 availability. IGF1 and insulin both promote growth of tumor cells and provide tumor cells with an ability to protect themselves from apoptosis [45]. Lipids are considered as functionally active organelles in colon cancer cells and are involved in PGE_2 synthesis. For GSE10950 and GSE11223 data sets, both MRMFS and MSMFS criteria annotate the term *"response to cytokine"* significantly, but the p-value is lower for the MSMFS criterion. The set of genes, annotating the term *"response to cytokine"*, is rich in interferons, interleukins, and growth factors. They are known to regulate the activity of other cells. Several studies exist that demonstrate the role played by cytokines in tumor development, angiogenesis, metastasis, etc. Cytokines possess the ability to induce cell transformation and malignancy [46], [47].

Table 3 compares the performance of different criteria using the KEGG pathway enrichment analysis of the obtained gene sets. Analyzing Table 3, it can be seen that the term *"TNF signaling pathway"*, annotated by the MSMFS criterion for GSE25070, has the lowest p-value. The MRMFS criterion also annotates to the same term but with a higher p-value. The term *"TNF signaling pathway"* is very closely associated to colorectal cancer. Tumor necrosis factor (TNF), being a multifunctional cytokine, plays a role in growth, proliferation, invasion and metastasis of cancer cells. For GSE10950, both MRMFS and MSMFS criteria annotate the term *"cytokine-cytokine receptor interaction"*. However, the proposed MSMFS criterion achieves it with significantly lower p-value than the MRMFS criterion. The *"cytokine-cytokine receptor interaction"* is known to induce a cascade of signaling pathways, for example, the interaction of TNFα and IL1β with their respective receptors activates the *"NF-kappa B signaling pathway"*, while IL-6 with its receptor gp130 activates STAT3, a major oncogenic transcription factor. Several other such interactions exist, which demonstrate the importance of cytokine interaction in colorectal cancer [46], [47]. The *"Jak-STAT signaling pathway"*, annotated by

TABLE 3
Comparative Performance Analysis of Different Criteria Using KEGG Pathway and Disease Ontology

Data Sets	Criteria	KEGG Pathway: Term and P-Value		Disease Ontology: Term and P-Value	
GSE25070	*t*-test	Retinol metabolism	4.82E-04	stomach cancer	4.12E-05
	MR	Nitrogen metabolism	2.09E-04	autosomal genetic disease	3.65E-11
	MS	Pentose and glucuronate interconversions	2.55E-04	gastrointestinal system disease	2.62E-04
	mRMR	*	*	stomach carcinoma	4.03E-05
	MRMS	Fatty acid degradation	2.12E-03	gastrointestinal system disease	4.77E-07
	MFS	Signaling pathways regulating pluripotency of stem cells	7.08E-06	colorectal cancer	2.57E-02
	MRMFS	TNF signaling pathway	1.19E-06	colorectal cancer	4.56E-02
	MSMFS	TNF signaling pathway	1.57E-11	colorectal cancer	8.81E-08
GSE10950	*t*-test	Viral myocarditis	3.55E-03	*	*
	MR	*	*	*	*
	MS	Nucleotide excision repair	1.12E-02	*	*
	mRMR	*	*	*	*
	MRMS	*	*	*	*
	MFS	Hepatitis B	2.02E-13	colorectal cancer	2.40E-12
	MRMFS	Cytokine-cytokine receptor interaction	1.10E-14	colorectal cancer	1.51E-08
	MSMFS	Cytokine-cytokine receptor interaction	7.88E-23	colorectal cancer	3.36E-11
GSE11223	*t*-test	*	*	*	*
	MR	Sphingolipid signaling pathway	4.54E-03	*	*
	MS	Prostate cancer	5.15E-04	colorectal cancer	5.70E-08
	mRMR	*	*	*	*
	MRMS	Long-term depression	2.31E-03	connective tissue cancer	4.11E-02
	MFS	Cytokine-cytokine receptor interaction	9.20E-29	colorectal cancer	5.70E-08
	MRMFS	Cytokine-cytokine receptor interaction	5.74E-19	colorectal cancer	2.41E-09
	MSMFS	Jak-STAT signaling pathway	1.93E-19	colorectal cancer	6.40E-12
Breast	*t*-test	Estrogen signaling pathway	2.60E-02	breast cancer	9.14E-04
	MR	Leishmaniasis	7.22E-06	breast cancer	8.04E-03
	MS	Leishmaniasis	1.68E-04	breast cancer	1.94E-02
	mRMR	Leishmaniasis	1.66E-04	breast cancer	1.08E-03
	MRMS	Leishmaniasis	8.46E-06	breast cancer	3.29E-02
	MFS	Pathways in cancer	2.89E-23	breast cancer	4.75E-21
	MRMFS	Pathways in cancer	1.69E-21	breast cancer	7.74E-16
	MSMFS	Pathways in cancer	1.06E-21	breast cancer	3.61E-22
Leukemia	*t*-test	Epstein-Barr virus infection	1.24E-03	leukemia	4.43E-03
	MR	Lysosome	7.60E-03	leukemia	4.87E-02
	MS	Epstein-Barr virus infection	6.98E-04	leukemia	5.30E-04
	mRMR	Lysosome	8.44E-03	leukemia	5.48E-03
	MRMS	Lysosome	1.00E-02	prostate cancer	4.54E-02
	MFS	Drug metabolism	2.89E-10	leukemia	1.87E-19
	MRMFS	Pathways in cancer	3.81E-11	leukemia	4.87E-20
	MSMFS	NF-kappa B signaling pathway	3.19E-12	leukemia	3.90E-25

the proposed criterion for GSE11223, is known to be intensely involved in the growth and progression of colorectal cancer. It is known to be associated with cell growth, survival, invasion and motility through regulation of genes like BCL2, P16, VEGF, MMPs, etc. [48]. Hence, the pathways, annotated significantly by the selected set of genes for proposed MSMFS criterion, contribute in some form to the growth, proliferation, motility or survival of colorectal cancer cells. For leukemia, the proposed criterion annotates the term *"NF-kappa B signaling pathway"*, which plays a vital role in cell survival signaling [49]. It is also known to promote the Wnt/β-catenin activity and development of cancer, and thus, is known as a key agent for inflammation mediated cancer [50].

Table 3 also compares the performance of different criteria on the basis of DO. All the results reported in Table 3 show that the gene sets obtained using the MS, MRMFS, and MSMFS criteria annotate to DO terms, which are significantly related to corresponding diseases, but the MSMFS criterion gives significantly lower p-values in most of the cases. From all the results reported in Tables 1, 2, and 3, it can be seen that the MRMFS and MSMFS criteria outperform other criteria. Also, the MS criterion performs better than the MR criterion as the former considers joint distribution between an already-selected gene and a candidate gene, rather than the individual distribution of candidate gene considered in the latter. On the other hand, the MRMS criterion performs better than the mRMR criterion in most of the cases, as the redundancy measure of the mRMR criterion does not take into account the supervised information of class labels. A significantly better performance is obtained for the proposed MSMFS criterion due to the fact that it considers both gene expression and PPI network data while selecting a gene as a candidate disease gene. Also, both significance and functional similarity measures, used in the MSMFS criterion, consider the association of an already-selected gene with a candidate gene.

TABLE 4
Degree of Overlapping with Known Gene Lists and P-Value for Genes Identified by Different Similarity Measures

Data Sets	Measures	LIST 1:	Overlap and P-Value	LIST 2:	Overlap and P-Value	LIST 3:	Overlap and P-Value	LIST 2-3	LIST 1-2-3
	Jaccard	6	2.13E-01	31	6.21E-28	15	4.63E-17	34	38
	Simpson	10	6.37E-03	5	5.79E-02	6	3.30E-05	10	20
GSE25070	Geometric	8	4.78E-02	29	2.89E-25	11	2.66E-11	30	36
	Dice	12	6.59E-04	34	4.31E-32	17	3.62E-20	38	46
	Cosine	12	6.59E-04	31	6.21E-28	17	3.62E-20	36	45
	Jaccard	17	4.44E-07	4	1.60E-01	2	1.25E-01	5	22
	Simpson	19	1.45E-08	4	1.60E-01	3	2.36E-02	6	24
GSE10950	Geometric	4	5.74E-01	1	8.81E-01	0	1.00E+00	1	5
	Dice	19	1.45E-08	6	1.90E-02	8	1.85E-07	11	30
	Cosine	21	3.65E-10	3	3.52E-01	3	2.36E-02	5	26
	Jaccard	19	1.45E-08	6	1.90E-02	6	3.50E-05	8	27
	Simpson	19	1.45E-08	5	6.00E-02	3	2.36E-02	7	26
GSE11223	Geometric	18	8.28E-08	4	1.60E-01	4	3.34E-03	6	24
	Dice	21	3.65E-10	7	5.19E-03	8	1.85E-07	10	31
	Cosine	21	3.65E-10	7	5.19E-03	7	2.74E-06	10	31

4.2 Performance of Different Similarity Measures

This section presents the comparative performance analysis of different similarity measures, namely, Cosine, Dice, Geometric, Simpson, and Jaccard, for disease gene selection. The weighted versions of last four similarity measures are defined similar to that of Cosine from corresponding non-weighted measures available in [32]. Tables 4, 5, and 6 compare the performance of different similarity measures with respect to degree of overlapping with known cancer-related gene lists, prediction accuracy and gene ontology, and disease ontology along with KEGG pathway analysis, respectively.

All the results reported in Table 4 show that the performance of both Dice and Cosine coefficients is similar for GSE25070 and GSE11223. However, for GSE10950, the Dice coefficient performs better than Cosine with respect to both LIST-2 and LIST-3. Next, Table 5 compares the performance of different similarity measures with respect to both area under the curve (AUC) and area under the precision-recall (AUPR) curve. To compute both AUC and AUPR, k-nearest neighbor rule [35] is used. The value of k is taken as square root of the number of samples in training set. To compute both AUC and AUPR, leave-one-out cross-validation is performed on each gene expression data set. All the results reported in Table 5 establish the fact that although the performance of both Dice and Cosine is same for GSE10950 and leukemia data, Cosine attains overall good AUC and AUPR values, irrespective of the data sets used.

Table 5 also compares the performance of different measures with respect to gene ontology. From the results reported here, it can be seen that Cosine coefficient annotates relevant term with significantly lower p-value, irrespective of the data sets. For example, the term, annotated by Cosine for GSE25070 data, has a strong relation to colon cancer, as explained in Section 4.1. Though the term annotated by Dice coefficient has also a significant relation to colorectal cancer, but the lowest p-value is obtained using Cosine coefficient. The set of genes obtained using Geometric for GSE10950 annotates significantly to a particular term. However, no evidence is found to support the association of this term to colorectal cancer. Table 6 compares the performance of all the measures using the KEGG pathway enrichment analysis. Here, it can be seen that Jaccard gives the lowest p-value for the term *"Chemokine signaling pathway"* for GSE25070, which bears a close association to colon cancer [51]. Both Dice and Cosine annotate to the term *"TNF Signaling Pathway"*, which also has a very significant association to colon cancer, but the latter gives a lower p-value. For leukemia data, both Simpson and Cosine annotate to *"NF-kappa B signaling pathway"*, which is significantly associated with leukemia, but the p-value of Cosine is lower than that of Simpson. The disease ontology based analysis, reported in Table 6, shows that the gene sets selected using Cosine annotate disease specific terms with significantly lower p-values in most of the cases. Analyzing all the results reported in Tables 4, 5, and 6, it can be inferred that Cosine coefficient is better among all the measures for computing functional similarity. Even though degree of overlapping is higher for Dice coefficient, the gene set selected using Cosine coefficient is biologically more meaningful and has a higher association to corresponding disease.

4.3 Performance of Different Integrated Methods

Finally, the performance of the proposed SiFS algorithm is compared with that of MR+PPIN [1], mRMR+PPIN [24], MRMS+PPIN [25], RelSim [26], and CLAIM [38], and corresponding results are reported in Tables 7, 8, and 9. Table 7 compares the performance of the proposed SiFS algorithm with that of existing integrated methods with respect to the degree of overlapping with three known gene lists for three colon cancer data sets. From the results reported in Table 7, it can be seen that the proposed method performs significantly better than the existing methods in 14 cases out of total 15 cases. Also, the computed p-value for the proposed SiFS algorithm is significantly lower as compared to other existing methods. Only for GSE25070 data, the MRMS+PPIN attains highest degree of overlapping with LIST-1.

Table 8 compares the performance of different integrated methods with respect to both AUC and AUPR. All the results reported in Table 8 establish the fact that the SiFS algorithm provides better performance than existing methods in most

TABLE 5
Prediction and Gene Ontology Based Analysis for Genes Identified by Different Similarity Measures

Data Sets	Measures	AUC	AUPR	Biological Processes: Term and P-Value	
GSE25070	Jaccard	0.991	0.953	regulation of locomotion	2.35E-18
	Simpson	0.991	0.953	response to lipopolysaccharide	1.36E-09
	Geometric	0.991	0.953	negative regulation of growth	2.24E-05
	Dice	0.991	0.953	response to lipopolysaccharide	3.38E-15
	Cosine	0.999	0.960	response to lipid	2.00E-23
GSE10950	Jaccard	1.000	0.958	Wnt signaling pathway	1.88E-16
	Simpson	1.000	0.958	actin filament-based process	1.41E-21
	Geometric	1.000	0.958	SRP-dependent cotranslational protein targeting to membrane	8.74E-45
	Dice	1.000	0.958	epithelium development	7.59E-30
	Cosine	1.000	0.958	response to cytokine	3.54E-36
GSE11223	Jaccard	0.988	0.977	intrinsic apoptotic signaling pathway	6.19E-09
	Simpson	0.988	0.976	response to cytokine	1.59E-40
	Geometric	0.987	0.977	negative regulation of G1/S transition of mitotic cell cycle	1.70E-05
	Dice	0.987	0.974	response to cytokine	1.46E-33
	Cosine	0.988	0.976	response to cytokine	2.83E-32
Breast	Jaccard	0.770	0.769	mRNA splicing, via spliceosome	7.63E-25
	Simpson	0.883	0.855	positive regulation of immune system process	1.41E-40
	Geometric	0.867	0.844	mRNA splicing, via spliceosome	8.28E-21
	Dice	0.980	0.943	response to lipid	7.47E-34
	Cosine	0.967	0.931	positive regulation of immune system process	6.84E-33
Leukemia	Jaccard	0.984	0.971	leukocyte activation	8.88E-20
	Simpson	0.908	0.933	inflammatory response	7.98E-25
	Geometric	0.975	0.966	cell activation	7.36E-24
	Dice	0.964	0.961	response to xenobiotic stimulus	6.50E-18
	Cosine	0.964	0.961	leukocyte activation	5.93E-22

of the cases, with respect to both AUC and AUPR. Table 8 also compares the performance of different methods using biological processes of gene ontology. All the results reported in Table 8 confirm that the performance of the proposed algorithm is significantly better than that of all other algorithms. The proposed SiFS achieves lowest p-values in four cases out of total five cases, with significant gene ontology terms related to corresponding diseases. For

TABLE 6
Comparative Performance Analysis of Different Measures Using KEGG Pathway and Disease Ontology

Data Sets	Measures	KEGG Pathway: Term and P-Value		Disease Ontology: Term and P-Value	
GSE25070	Jaccard	Chemokine signaling pathway	4.62E-13	gastrointestinal system disease	1.05E-11
	Simpson	Signaling pathways regulating pluripotency of stem cells	7.08E-06	colorectal cancer	2.44E-02
	Geometric	Pentose and glucuronate interconversions	2.31E-03	colorectal cancer	3.33E-02
	Dice	TNF signaling pathway	2.00E-09	colorectal cancer	2.42E-05
	Cosine	TNF signaling pathway	1.57E-11	colorectal cancer	8.81E-08
GSE10950	Jaccard	Wnt signaling pathway	3.55E-13	colorectal cancer	1.15E-03
	Simpson	Regulation of actin cytoskeleton	9.27E-17	colorectal cancer	3.10E-05
	Geometric	Ribosome	2.34E-27	Diamond-Blackfan anemia	5.23E-03
	Dice	Pathways in cancer	1.26E-25	colorectal cancer	6.87E-06
	Cosine	Cytokine-cytokine receptor interaction	7.88E-23	colorectal cancer	3.36E-11
GSE11223	Jaccard	FoxO signaling pathway	3.74E-08	colorectal cancer	2.15E-04
	Simpson	Cytokine-cytokine receptor interaction	6.80E-29	colorectal cancer	5.04E-11
	Geometric	Pathways in cancer	1.23E-05	colorectal cancer	1.19E-02
	Dice	Cytokine-cytokine receptor interaction	5.64E-20	colorectal cancer	4.75E-11
	Cosine	Jak-STAT signaling pathway	1.93E-19	colorectal cancer	6.40E-12
Breast	Jaccard	Epstein-Barr virus infection	1.61E-07	breast cancer	2.91E-03
	Simpson	Toll-like receptor signaling pathway	1.12E-17	breast cancer	3.77E-14
	Geometric	Epstein-Barr virus infection	2.34E-04	breast cancer	3.99E-03
	Dice	Pathways in cancer	8.16E-23	breast cancer	5.38E-23
	Cosine	Pathways in cancer	1.06E-21	breast cancer	3.61E-22
Leukemia	Jaccard	Pathways in cancer	8.57E-12	leukemia	3.69E-23
	Simpson	NF-kappa B signaling pathway	7.81E-10	leukemia	1.17E-25
	Geometric	TNF signaling pathway	1.18E-09	leukemia	1.56E-15
	Dice	Prostate cancer	1.29E-12	leukemia	6.40E-19
	Cosine	NF-kappa B signaling pathway	3.19E-12	leukemia	3.90E-25

TABLE 7
Degree of Overlapping with Known Gene Lists and P-Value for Genes Identified by Different Integrated Methods

Data Sets	Methods	LIST 1:	Overlap and P-Value	LIST 2:	Overlap and P-Value	LIST 3:	Overlap and P-Value	LIST 2-3	LIST 1-2-3
GSE25070	MR+PPIN	9	2.84E-05	7	2.10E-05	5	5.00E-06	10	18
	mRMR+PPIN	8	1.91E-04	4	1.06E-02	3	2.02E-03	5	13
	MRMS+PPIN	**21**	2.00E-10	15	1.76E-09	11	1.90E-11	19	38
	CLAIM	4	5.74E-01	10	4.89E-05	2	1.25E-01	11	15
	RelSim	14	4.48E-05	28	5.78E-24	12	1.11E-12	32	42
	SiFS	12	6.59E-04	**31**	6.21E-28	**17**	3.62E-20	**36**	**45**
GSE10950	MR+PPIN	2	1.91E-01	2	6.54E-02	1	1.16E-01	2	4
	mRMR+PPIN	3	2.92E-01	3	7.60E-02	1	2.51E-01	3	6
	MRMS+PPIN	5	6.90E-02	2	3.22E-01	1	2.87E-01	2	7
	CLAIM	5	3.74E-01	3	3.52E-01	1	4.60E-01	4	9
	RelSim	13	1.79E-04	3	3.52E-01	1	4.60E-01	3	16
	SiFS	**19**	1.45E-08	**6**	1.90E-02	**8**	1.85E-07	**11**	**30**
GSE11223	MR+PPIN	5	3.74E-01	4	1.60E-01	4	3.34E-03	7	12
	mRMR+PPIN	6	1.07E-01	1	8.22E-01	2	8.82E-02	3	9
	MRMS+PPIN	4	5.74E-01	2	6.25E-01	2	1.25E-01	4	8
	CLAIM	7	1.07E-01	6	1.90E-02	3	2.36E-02	7	14
	RelSim	20	2.37E-09	6	1.90E-02	7	2.74E-06	9	29
	SiFS	**21**	3.65E-10	**7**	5.19E-03	**7**	2.74E-06	**10**	**31**

GSE25070 data, the SiFS annotates the term *"response to lipid"* with significantly lower p-value, which bears a strong relation to colon cancer, as explained in Section 4.1. The RelSim annotates to the term *"TNF signaling pathway"*, which is also related to colon cancer, as explained in Section 4.1, but the corresponding p-value is higher. The MR+PPIN annotates to the term *"intracellular steroid hormone receptor signaling pathway"* significantly, which also plays an important role in colorectal cancer progression. However, no significant association to colorectal cancer was found for the terms annotated by mRMR+PPIN and CLAIM. The term annotated by the MRMS+PPIN is associated to colon cancer in a manner

TABLE 8
Prediction and Gene Ontology Based Analysis for Genes Identified by Different Integrated Methods

Data Sets	Methods	AUC	AUPR	Biological Processes: Term and P-Value	
GSE25070	MR+PPIN	0.982	0.944	intracellular steroid hormone receptor signaling pathway	6.78E-13
	mRMR+PPIN	0.905	0.879	DNA-templated transcription, initiation	1.32E-12
	MRMS+PPIN	0.991	0.953	cellular response to lipid	2.68E-13
	CLAIM	0.996	0.957	regulation of cell junction assembly	1.15E-03
	RelSim	0.991	0.953	TNF signaling pathway	7.57E-05
	SiFS	0.999	0.960	response to lipid	2.00E-23
GSE10950	MR+PPIN	0.993	0.952	*	*
	mRMR+PPIN	0.997	0.955	membrane protein proteolysis	4.69E-04
	MRMS+PPIN	1.000	0.958	protein polyubiquitination	7.92E-09
	CLAIM	1.000	0.958	*	*
	RelSim	1.000	0.958	regulation of viral transcription	6.99E-04
	SiFS	1.000	0.958	response to cytokine	3.54E-36
GSE11223	MR+PPIN	0.989	0.980	tyrosine phosphorylation of Stat1 protein	1.26E-04
	mRMR+PPIN	0.987	0.977	regulation of protein import	4.53E-06
	MRMS+PPIN	0.987	0.978	protein import into nucleus	5.13E-06
	CLAIM	0.975	0.949	positive regulation of lymphocyte activation	2.73E-06
	RelSim	0.987	0.976	positive regulation of immune system process	8.35E-36
	SiFS	0.988	0.976	response to cytokine	2.83E-32
Breast	MR+PPIN	0.602	0.635	response to laminar fluid shear stress	2.93E-04
	mRMR+PPIN	0.707	0.760	regulation of carbohydrate biosynthetic process	1.36E-04
	MRMS+PPIN	0.912	0.883	positive regulation of lymphocyte activation	1.32E-05
	CLAIM	0.475	0.576	interspecies interaction between organisms	3.15E-21
	RelSim	0.950	0.918	response to alcohol	3.61E-10
	SiFS	0.967	0.931	positive regulation of immune system process	6.84E-33
Leukemia	MR+PPIN	0.792	0.859	platelet degranulation	6.93E-04
	mRMR+PPIN	0.794	0.867	regulation of tissue remodeling	8.08E-05
	MRMS+PPIN	0.810	0.873	chaperone-mediated autophagy	3.23E-05
	CLAIM	0.943	0.948	stimulatory C-type lectin receptor signaling pathway	1.97E-04
	RelSim	0.994	0.975	myeloid leukocyte activation	4.35E-07
	SiFS	0.964	0.961	leukocyte activation	5.93E-22

TABLE 9
Comparative Performance Analysis of Different Integrated Methods Using KEGG Pathway and Disease Ontology

Data Sets	Methods	KEGG Pathway: Term and P-Value		Disease Ontology: Term and P-Value	
GSE25070	MR+PPIN	Thyroid hormone signaling pathway	2.74E-09	colon carcinoma	1.01E-03
	mRMR+PPIN	Thyroid hormone signaling pathway	2.08E-13	coronary artery disease	1.12E-02
	MRMS+PPIN	Thyroid hormone signaling pathway	1.47E-08	colon carcinoma	6.68E-05
	CLAIM	Arachidonic acid metabolism	7.78E-03	vascular disease	3.06E-02
	RelSim	TNF signaling pathway	7.57E-05	colon carcinoma	4.25E-02
	SiFS	TNF signaling pathway	1.57E-11	colorectal cancer	8.81E-08
GSE10950	MR+PPIN	*	*	*	*
	mRMR+PPIN	*	*	disease of metabolism	4.70E-02
	MRMS+PPIN	Ubiquitin mediated proteolysis	3.10E-05	*	*
	CLAIM	Endometrial cancer	5.45E-03	*	*
	RelSim	AMPK signaling pathway	8.39E-03	colon adenocarcinoma	4.65E-02
	SiFS	Cytokine-cytokine receptor interaction	7.88E-23	colorectal cancer	3.36E-11
GSE11223	MR+PPIN	GnRH signaling pathway	3.59E-03	musculoskeletal system cancer	1.73E-03
	mRMR+PPIN	Pancreatic cancer	3.44E-03	hematologic cancer	9.34E-04
	MRMS+PPIN	Chagas disease (American trypanosomiasis)	2.20E-04	hematologic cancer	2.29E-04
	CLAIM	B cell receptor signaling pathway	1.56E-02	organ system cancer	3.12E-02
	RelSim	Cytokine-cytokine receptor interaction	4.76E-21	colorectal cancer	5.71E-10
	SiFS	Jak-STAT signaling pathway	1.93E-19	colorectal cancer	6.40E-12
Breast	MR+PPIN	Intestinal immune network for IgA production	6.21E-03	breast cancer	9.41E-03
	mRMR+PPIN	Toxoplasmosis	2.75E-04	hereditary breast ovarian cancer	1.68E-02
	MRMS+PPIN	Intestinal immune network for IgA production	1.76E-04	breast cancer	9.41E-03
	CLAIM	Ribosome	1.79E-08	*	*
	RelSim	TNF signaling pathway	2.27E-05	breast cancer	6.97E-06
	SiFS	Pathways in cancer	1.06E-21	breast cancer	3.61E-22
Leukemia	MR+PPIN	Inositol phosphate metabolism	1.08E-01	*	*
	mRMR+PPIN	Malaria	1.04E-04	*	*
	MRMS+PPIN	B cell receptor signaling pathway	1.16E-03	leukemia	1.19E-02
	CLAIM	Epstein-Barr virus infection	2.45E-05	neurodegenerative disease	9.63E-04
	RelSim	Osteoclast differentiation	8.22E-10	leukemia	1.64E-06
	SiFS	NF-kappa B signaling pathway	3.19E-12	leukemia	3.90E-25

similar to that of the SiFS algorithm. The significantly lower p-values of the SiFS algorithm for GO, irrespective of the data sets used, establish its superiority over other algorithms.

Table 9 performs a similar comparison on the basis of KEGG enrichment analysis and disease ontology. For each data set, the set of genes selected by the proposed algorithm is annotated to a term with significantly lower p-value, which bears a strong association with the corresponding disease. For example, both RelSim and SiFS annotate to *"TNF signaling pathway"*, which is significantly associated with colon cancer as established in Section 4.1, but the p-value of the SiFS is significantly lower as compared to that of the RelSim. The term *"Thyroid hormone signaling pathway"*, annotated by the MR+PPIN, mRMR+PPIN, and MRMS+PPIN for GSE25070, also bears a strong association with colorectal cancer. Moreover, the disease ontology based analysis, reported in Table 9, establishes the fact that the proposed SiFS algorithm annotates significant disease associated terms with lowest p-values for all the cases. All the results reported in Tables 7, 8, and 9 justify the superiority of the proposed algorithm and establish the fact that the SiFS achieves significantly better performance.

4.4 Biological Interpretation of Identified Genes

In the present study, the proposed SiFS algorithm is used to select 100 top-ranked genes for GSE25070 data, 45 of which are already enlisted as disease genes in either of LIST 1, LIST 2, or LIST 3 [39], [40], [41], [42]. Among them, following 17 genes, namely, ***CCL20, CHGA, CLDN1, CLEC3B, CNTN3, CXCL12, ETV4, FOXQ1, IL1B, LCN2, MMP11, MSX1, PCK1, PYY, SERPINE1, SFRP1***, and **TNFRSF17**, are the unique disease genes identified only by the SiFS algorithm. On the other hand, among the remaining 55 genes that have been predicted to be associated with colorectal cancer, there exists a set of 32 unique genes, namely, ***ACVRL1, ANGPT2, BATF, BMP2, CARD14, CEBPB, CTLA4, DHRS9, FAP, GZMB, IFNG, IL23A, IRAK1, IRAK2, LPHN3, LRP8, LYN, MAPK10, MB, OLR1, PDCD4, RIPK2, SOX4, STAT1, STAT3, STAT5A, TLR5, TNFRSF13B, TNFRSF1A, TNFSF10, WISP1***, and **WISP2**, those have been identified only by the proposed SiFS algorithm, not by any other existing integrated methods for disease gene selection. This section presents the biological interpretation of some of these predicted disease genes.

ANGPT2. The protein product of ANGPT2 acts as an antagonist of ANGPT1. It shows an increased expression in colorectal adenocarcinoma, while ANGPT1 is down-regulated. This balance leads to blood vessel formation and rapid growth of the adenocarcinoma [52].

BMP2. In sporadic colorectal cancer, BMP2 and BMP7 at the ligand level have been found to play a growth suppressive role in tumor development. It is also known to inhibit colon cancer cell migration and invasiveness. Being a

member of the TGFβ signaling pathway, they are involved in the initiation and formation of colon cancer. Its pathway has been involved in the initiation of colon cancer among individuals with juvenile polyposis harboring BMPR1A receptor mutations [53].

FAP. Seprase is a known alias of FAP. Its presence in increased levels in the stroma, is likely to show an aggressive disease progression and has a strong potential for metastasis occurrence in human colorectal tumor patients. Inhibition of its enzymatic action is known to control tumor growth [54].

IL23A. IL-23 is a heterodimeric cytokine naturally expressed by activated dendritic cells, having a stimulatory effect on memory T cells and bears a close association to bowel inflammation and colorectal cancer. Higher IL-23 levels along with lower levels of Socs3, assisted by STAT5 in colorectal cancer tissues, leads to an increase in metastatic cases [55].

IFNG. Interferons (IFNs) are proteins that function to control processes like apoptosis, cell cycle, cytokine actions. IFN gamma (IFNG) is a cytokine that supports inflammation and modulates many immune related genes. It is known to act as a two sided sword, playing roles in suppressing as well as promoting tumorigenic activities. Its property of protecting normal cells of an imflammated site while repairing of tissues destroyed in this process, favors tumor cells by allowing them to evade destruction. Its interaction with NF-κB1 and IL6 is known to play a pivotal role in inflammation driven colon cancer [56].

SOX4. The protein encoded for this gene plays a major role in the apoptosis pathway leading to cell death and tumor invasion. Gut epithelium, showing an expression of SOX protein, positively modulates the Wnt Signaling pathway and tumor cell invasion in the gastrointestinal tract. Nuclear SOX4 is highly associated with tumor invasion and metastasis in colon cancer tissues. miR-335 is known to inhibit metastasis through down-regulation of SOX4 [57].

TLR5. It has been shown that blocking TLR5/MyD88-mediated signaling in human colon cancer mouse xenografts suppresses innate immune responses in the model, which is represented by reduced neutrophil infiltration and enhanced tumor growth. However, TLR5 activation using flagellin is known to dramatically reduce tumor growth [58].

TNFSF10. The encoded protein of this gene is a cytokine belonging to the tumor necrosis factor ligand family. It induces selective apoptosis, that is, its interaction with death receptors transmits a death signal only in the transformed tumor cells through caspese activation and does not damage the normal cells [59].

5 Conclusion

The paper presents a novel gene selection algorithm, termed as SiFS, for identification of disease related genes. The information of gene expression profiles and PPI networks is integrated judiciously to find out the disease genes. A set of genes is selected by the proposed SiFS algorithm as disease genes from microarray data and PPI network by maximizing the significance and functional similarity among the selected genes. To compute the functional similarity between two genes, a new weighted similarity measure is defined, which is based on the information of PPI network. The performance of the proposed SiFS algorithm is extensively studied using several cancer data sets, and compared with that of other related methods. The experimental results on different data sets show that the proposed algorithm identifies more disease causing genes as compared to existing disease gene selection methods. It also establishes the importance of proposed method and indicates the possibility of SiFS algorithm to be an useful technique for disease gene identification.

Acknowledgments

The authors would like to thank Dr. Sushmita Paul of the Indian Institute of Technology, Jodhpur, India, for providing helpful and valuable criticisms. This work is partially supported by the Department of Electronics and Information Technology, Government of India (PhD-MLA/4(90)/2015-16).

References

[1] P. Maji and S. Paul, *Scalable Pattern Recognition Algorithms: Applications in Computational Biology and Bioinformatics*. London, U.K.: Springer, Apr. 2014.

[2] T. Huang, et al., "Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks," *PLoS One*, vol. 5, no. 6, 2010, Art. no. e10972.

[3] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.

[4] P. Maji and S. Paul, "Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data," *Int. J. Approximate Reasoning*, vol. 52, no. 3, pp. 408–426, 2011.

[5] J. Meng, J. Zhang, and Y. Luan, "Gene selection integrated with biological knowledge for plant stress response using neighborhood system and rough set theory," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 2, pp. 433–444, Mar./Apr. 2015.

[6] B. Liao, Y. Jiang, W. Liang, W. Zhu, L. Cai, and Z. Cao, "Gene selection using locality sensitive Laplacian score," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 6, pp. 1146–1156, Nov./Dec. 2014.

[7] Y. Leung and Y. Hung, "A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 7, no. 1, pp. 108–117, Jan.–Mar. 2010.

[8] A. Sharma, S. Imoto, and S. Miyano, "A top-r feature selection algorithm for microarray gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 3, pp. 754–764, May/Jun. 2012.

[9] P. Maji, "f-information measures for efficient selection of discriminative genes from microarray data," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1063–1069, Apr. 2009.

[10] P. Maji and S. K. Pal, "Fuzzy-rough sets for information measures and selection of relevant genes from microarray data," *IEEE Trans. Syst Man Cybern. Part B: Cybern.*, vol. 40, no. 3, pp. 741–752, Jun. 2010.

[11] J.-M. Wei, S.-Q. Wang, and X.-J. Yuan, "Ensemble rough hypercuboid approach for classifying cancers," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 3, pp. 381–391, Mar. 2010.

[12] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabsi, "The human disease network," *Proc. Nat. Academy Sci. USA*, vol. 104, no. 21, pp. 8685–8690, 2007.

[13] J. Zhao, P. Jiang, and W. Zhang, "Molecular networks for the study of TCM pharmacology," *Briefings Bioinf.*, vol. 11, no. 4, pp. 417–430, 2010.

[14] J. Chen, B. Aronow, and A. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks," *BMC Bioinf.*, vol. 10, no. 1, 2009, Art. no. 73.
[15] S. Kohler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *Amer. J. Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.
[16] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinf.*, vol. 26, no. 8, pp. 1057–1063, 2010.
[17] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein-protein interactions," *J. Med. Genetics*, vol. 43, no. 8, pp. 691–698, 2006.
[18] C. Wu, J. Zhu, and X. Zhang, "Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes," *BMC Bioinf.*, vol. 13, no. 1, 2012, Art. no. 182.
[19] P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao, "dmGWAS: Dense module searching for genome-wide association studies in protein-protein interaction networks," *Bioinf.*, vol. 27, no. 1, pp. 95–102, 2011.
[20] W. Kong, J. Zhang, X. Mou, and Y. Yang, "Integrating gene expression and protein interaction data for signaling pathway prediction of Alzheimer's disease," *Comput. Math. Methods Med.*, vol. 2014, 2014, Art. no. 340758.
[21] J. Zhao, T.-H. Yang, Y. Huang, and P. Holme, "Ranking candidate disease genes from gene expression and protein interaction: A Katz-centrality based approach," *PLoS One*, vol. 6, no. 9, 2011, Art. no. e24306.
[22] Y. Li and J. Li, "Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data," *BMC Genomics*, vol. 13, no. suppl 7, 2012, Art. no. S27.
[23] J. Li, et al., "Mining disease genes using integrated protein-protein interaction and gene-gene co-regulation information," *FEBS Open Bio*, vol. 5, pp. 251–256, 2015.
[24] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou, "Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network," *PLoS One*, vol. 7, no. 4, 2012, Art. no. e33393.
[25] S. Paul and P. Maji, "Gene expression and protein-protein interaction data for identification of colon cancer related genes using f-information measures," *Natural Comput.*, pp. 1–15, 2015, Doi: 10.1007/s11047-015-9485-6.
[26] P. Maji, E. Shah, and S. Paul, "A new similarity measure for identification of disease genes," in *Proc. 6th Int. Conf. Pattern Recognition Mach. Intell.*, 2015, pp. 451–461.
[27] J. Meng, D. Liu, and Y. Luan, "Inferring plant microRNA functional similarity using a weighted protein-protein interaction network," *BMC Bioinf.*, vol. 16, no. 1, pp. 1–11, 2015.
[28] P. Bogdanov and A. Singh, "Molecular function prediction using neighborhood features," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 7, no. 2, pp. 208–217, Apr.–Jun. 2010.
[29] K.-L. Ng, J.-S. Ciou, and C.-H. Huang, "Prediction of protein functions based on function-function correlation relations," *Comput. Biol. Med.*, vol. 40, no. 3, pp. 300–305, 2010.
[30] U. Karaoz, et al., "Whole-genome annotation by using evidence integration in functional-linkage networks," *Proc. Nat. Academy Sci. USA*, vol. 101, no. 9, pp. 2888–2893, 2004.
[31] S. Letovsky and S. Kasif, "Predicting protein function from protein/protein interaction data: A probabilistic approach," *Bioinf.*, vol. 19, no. suppl 1, pp. i197–i204, 2003.
[32] J. I. F. Bass, et al., "Using networks to measure similarity between genes: Association index selection," *Nature Methods*, vol. 10, no. 12, pp. 1169–1176, Dec. 2013.
[33] P. Maji, "Mutual information based supervised attribute clustering for microarray sample classification," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 1, pp. 127–140, Jan. 2012.
[34] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statistics*, vol. AMS-33, pp. 1065–1076, 1962.
[35] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Cambridge, MA, USA: Academic Press, 1990.
[36] B. W. Silverman, *Density Estimation for Statistics and Data Analysis (Statistics and Applied Probability)*. London, U.K.: Chapman & Hall, 1986.
[37] D. Szklarczyk, et al., "The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res.*, vol. 39, no. suppl 1, pp. D561–D568, 2011.
[38] D. Santoni, et al., "An integrated approach (CLuster analysis integration method) to combine expression data and protein-protein interaction networks in agrigenomics: Application on arabidopsis thaliana," *OMICS: J. Integrative Biol.*, vol. 18, no. 2, pp. 155–165, 2014.
[39] J. L. Huret, P. Dessen, and A. Bernheim, "Atlas of genetics and cytogenetics in oncology and haematology," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 272–274, 2003.
[40] T. S. Keshava Prasad, et al., "Human protein reference database-2009 update," *Nucleic Acids Res.*, vol. 37, no. suppl. 1, pp. D767–D772, 2009.
[41] J. Sabates-Bellver, et al., "Transcriptome profile of human colorectal adenomas," *Mol. Cancer Res.*, vol. 5, no. 12, pp. 1263–1275, 2007.
[42] S. Nagaraj and A. Reverter, "A Boolean-based systems biology approach to predict novel genes associated with cancer: Application to colorectal cancer," *BMC Syst. Biol.*, vol. 5, no. 1, 2011, Art. no. 35.
[43] G. Bindea et al., "ClueGO: A cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks," *Bioinf.*, vol. 25, no. 8, pp. 1091–1093, 2009.
[44] G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He, "DOSE: An R/Bioconductor package for disease ontology semantic and enrichment analysis," *Bioinf.*, vol. 31, no. 4, pp. 608–609, 2015.
[45] C. R. Santos and A. Schulze, "Lipid metabolism in cancer," *FEBS J.*, vol. 279, no. 15, pp. 2610–2623, 2012.
[46] L. Klampfer, "Cytokines, inflammation and colon cancer," *Current Cancer Drug Targets*, vol. 11, no. 4, pp. 451–464, 2011.
[47] T. Hirano, K. Ishihara, and M. Hibi, "Roles of STAT3 in mediating the cell growth, differentiation and survival signals relayed through the IL-6 family of cytokine receptors space," *Oncogene*, vol. 19, no. 21, pp. 2548–2556, 2000.
[48] H. Xiong, et al., "Inhibition of JAK1, 2/STAT3 signaling induces apoptosis, cell cycle arrest, and reduces tumor cell invasion in colorectal cancer cells," *Neoplasia*, vol. 10, no. 3, pp. 287–297, 2008.
[49] X. Wang and Y. Lin, "Tumor necrosis factor and cancer, buddies or foes?" *Acta Pharmacol Sin*, vol. 29, no. 11, pp. 1275–1288, 2008.
[50] M. Coskun, et al., "Involvement of CDX2 in the cross talk between TNF-α and Wnt signaling pathway in the colon cancer cell line Caco-2," *Carcinogenesis*, vol. 35, no. 5, pp. 1185–1192, 2014.
[51] P. Sarvaiya, D. Guo, I. Ulasov, P. Gabikian, and M. Lesniak, "Chemokines in tumor progression and metastasis," *Oncotarget*, vol. 4, no. 12, pp. 2171–2185, 2013.
[52] H.-L. Wang, C.-S. Deng, J. Lin, D.-Y. Pan, Z.-Y. Zou, and X.-Y. Zhou, "Expression of angiopoietin-2 is correlated with vascularization and tumor size in human colorectal adenocarcinoma," *Tohoku J. Exp. Med.*, vol. 213, no. 1, pp. 33–40, 2007.
[53] J. C. Hardwick, L. L. Kodach, G. J. Offerhaus, and G. R. van den Brink, "Bone morphogenetic protein signalling in colorectal cancer," *Nature Rev. Cancer*, vol. 8, no. 10, pp. 806–812, 2008.
[54] W. Rollinger, J. Karl, J. Kochan, M. Roessler, and M. Tacke, Seprase as a Marker for Cancer, 2009, wO Patent App. PCT/EP2008/010,385. [Online]. Available: http://www.google.co.in/patents/WO2009074275A1?cl=en
[55] L. Zhang, et al., "IL-23 selectively promotes the metastasis of colorectal carcinoma cells with impaired Socs3 expression via the STAT5 pathway," *Carcinogenesis*, vol. 35, no. 6, pp. 1330–1340, 2014.
[56] M. R. Zaidi and G. Merlino, "The two faces of interferon-γ in cancer," *Clinical Cancer Res.*, vol. 17, no. 19, pp. 6118–6124, 2011.
[57] G. Wu, Y. Z. Zhu, and J. Zhang, "Sox4 up-regulates Cyr61 expression in colon cancer cells," *Cellular Physiology Biochemistry*, vol. 34, no. 2, pp. 405–412, 2014.
[58] S. N. Klimosch, et al., "Functional TLR5 genetic variants affect human colorectal cancer survival," *Cancer Res.*, vol. 73, no. 24, pp. 7232–7242, 2013.
[59] C. Stolf, et al., "2-methoxy-5-amino-N-hydroxybenzamide sensitizes colon cancer cells to TRAIL-induced apoptosis by regulating death receptor 5 and survivin expression," *Mol. Cancer Therapeutics*, vol. 10, no. 10, pp. 1969–1981, 2011.

Pradipta Maji received the BSc degree in physics, the MSc degree in electronics science, and the PhD degree in the area of computer science from Jadavpur University, India, in 1998, 2000, and 2005, respectively. Currently, he is an associate professor in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. His research interests include pattern recognition, machine learning, computational biology and bioinformatics, medical image processing, and so forth. He has published more than 100 papers in international journals and conferences. He is author of a book published by Wiley-IEEE Computer Society Press and another book published by Springer-Verlag, London. He received the 2008 Microsoft Young Faculty Award from Microsoft Research Laboratory India Pvt., the 2009 Young Scientist Award from the National Academy of Sciences, India, the 2011 Young Scientist Award from the Indian National Science Academy, India, and the 2015 Young Faculty Research Fellowship from the Department of Electronics and Information Technology, Government of India. He was selected as the 2009 Young Associate of the Indian Academy of Sciences, India.

Ekta Shah received the BTech and MTech degrees in computer science from the West Bengal University of Technology, India, in 2012 and Indian Statistical Institute, Kolkata, in 2015, respectively. Currently, she is a research scholar in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. Her research interests include pattern recognition, computational biology and bioinformatics, medical image processing, and so forth. She has published a few papers in international journals and conferences.