Reconstruction, Topological and Gene Ontology Enrichment Analysis of Cancerous Gene Regulatory Network Modules

Khalid Raza*

Department of Computer Science, Jamia Millia Islamia (Central University), New Delhi, India

Khalid Raza

Abstract: The availability of large set of high throughput biological data needs algorithm that automatically reconstructs gene regulatory networks from these datasets. Cancerous regulatory network modules when analyzed critically may reveal the underlying mechanism of cancer, which may help in better diagnosis. Identification of cancerous genes and their regulation is an important research area in cancer systems biology. In this paper, we introduced an algorithm to infer cancerous gene regulatory network modules from gene expression profiles. The proposed algorithm has been applied to gene expression dataset of colon cancer patients and several network modules have been identified. We performed topological analysis of inferred network modules in terms of network density, degree distribution, clustering coefficient, average path length, network heterogeneity, and centrality measures. Further, GO-based enrichment analysis of the inferred network has been performed. To validate the proposed algorithm, it has been tested on benchmark dataset taken from DREAM3 challenge project.

Keywords: Gene regulatory network, biological network, systems biology, GO enrichment analysis, topological analysis, cancerous network modules.

1. INTRODUCTION

Cancer is a kind of disease that involves unregulated cell growth. The cancerous cells divide and grow uncontrollably and infest the nearby part of the body. The possible means to diagnose cancer are chemotherapy, radiotherapy and surgery, but unfortunately, these methods of treatment often damage healthy cells and tissues. Therefore, identification of molecular markers of cancers may be an alternative approach to diagnose the human cancer which may be fruitful for the development of novel therapies. Although, various significant genes responsible for the genesis of different tumors have been revealed, but fundamental molecular interactions are still unknown and remain a challenge for the researchers.

Due to rapid growth in microarray technology, gene expression of tens of thousands of genes can be measured simultaneously in a single experiment using a small amount of test sample which enables the researchers to detect cancerous molecular markers [1]. Microarrays have been successfully used in many biomedical applications including gene discovery, drug discovery, disease diagnosis, and toxicology. Typical microarray gene expression data is a matrix R with N rows and M columns, where rows represent genes and column indicate the samples (or environmental conditions or time-point). Due to experimental limitations, major problem with microarray data are dimensionality problem (M<<N) and presence of noise in expression values.

Microarray-based cancer prediction is a new and growing area of research. A gene regulatory network (GRN) tries to model the complex regulatory interactions within the living cells and gives a realistic representation of gene regulation. Microarray gene expression profiles of whole genome are used to better understand the interaction mechanism in cancer and infer cancer-specific GRN modules. The changes in expression profile of genes across various samples provides information that can be used to filter differentially expressed genes (DEGs) between normal and tumor samples and helps to find regulatory interactions between gene-pairs, which lead the reconstruction of GRN. Mapping the topology of GRNs is the key issue in systems biology research [2]. Also, accurate computational methods to infer genome-scale GRN from gene expression profiles are required to explore available experimental data in a new and integrative manner. Biological networks, including GRNs, are much complex and any disruption in the network architecture leads to a kind of disease. Analysis of properties of these networks offers better insights to understand fundamental mechanisms that control cellular processes and disease pathways. Network topological properties often reveal lots of biologically significant information. Biological networks usually follow some patterns and rules and have a specific kind of topology, which let the researchers investigate deeper for knowledge extraction [3]. The complexity of biological network interactions in human, also known as "human interactome", is daunting. The human interactome contains around 1,00,000 interacting molecules that includes ~25,000 protein coding genes, ~1,000 metabolites and an undefined number of proteins and functional RNAs [4]. The advancement in graph and network theory has helped us in getting better insights about the properties of biological networks, especially disease networks. From the studies of networks, it is revealed that networks in biological, social and technological systems are not functioning randomly, but are organized by some set of

*Address correspondence to this author at the Department of Computer Science, Jamia Millia Islamia (Central University), New Delhi, India;
Tel: +91-9891478255; E-mail: kraza@jmi.ac.in

1574-8936/16 $58.00+.00

principles. These network principles help us to extract some of the basic properties of genes involved in a disease. The analysis of biological network has got a boost due to availability of huge biological experiment network data sources.

A comprehensive comparative evaluation of many state-of-the-art GRN inference methods has been done by Madhamshettiwar and his colleagues [5] and finally, best-performing method has been applied to infer GRN of ovarian cancer. Many other attempts have been made to infer GRN of various cancers including colon cancer, ovarian cancer, lungs cancer and breast cancer [1, 6-8]. Raza and Jaiswal [9] inferred GRN of prostate cancer from gene expression profiles using simple statistical approaches such as t-test, fold-change and Pearson correlation coefficient. Also, topological analysis of inferred cancerous network has been done. However, the analysis was limited to degree and centrality measures only. Yang and colleagues [8] proposed a differential network-based framework to detect biologically meaningful cancer-related genes. Yang and colleagues [8] applied boosting regression based on likelihood score and informative prior for improving the accuracy, and identified six genes (namely, TSPYL5, CD55, CCNE2, DCK, BBC3, and MUC1) susceptible to breast cancer.

The work presented in this paper is an extended version of our previously published conference paper [10]. In this work, firstly differentially expressed genes (DEGs) have been identified using a t-test statistics and then one of information theoretic approaches, called mutual information, was used to compute regulatory relationships between gene-pairs [10]. We applied this approach to infer gene regulations in colorectal cancer (CRC), the third leading cause of cancer mortality world-wide. To validate the proposed method, we applied it on simulated benchmark data of yeast taken from DREAM3 challenge. Further, we performed topological analysis and GO-based enrichment analysis of the inferred colon cancer network modules. Rest of the paper is organized as follows. Section 2 briefly describes the methods of identifying differentially expressed genes. Section 3 describes information theory, entropy and mutual information. Section 4 briefly discusses some of the most widely used topological properties of biological networks. Section 5 introduces GO-based enrichment analysis. Section 6 describes the proposed algorithm and its working. Section 7 presents results and discussions and finally paper has been concluded.

2. IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

An important purpose for monitoring expression level of genes is to identify those genes which are differentially expressed across two kinds of tissue samples, or samples observed under two different experimental conditions. Set of genes differentially expressed over two different samples, i.e., normal and cancerous tissue, are expected to give clues about cancer mechanism. A large variety of methods exist for finding differentially expressed genes and most of these methods are based on statistical techniques, such as fold-change [11], t-test statistics [12, 13], ANOVA [14], rank product [15], Significant Analysis of Microarray [16], Random Variance Model [17], and Limma [18], in addition to several others. Here, fold change and t-statistics methods have been briefly covered. Description of the rest of the methods can be found in the concerned literature [19, 20].

The fold change is the simplest method that can be computed as a ratio of averages from control and test sample. The levels of fold change are observed, and genes under or above a threshold are selected. For example, fold change below 0.5 is considered as down-regulated, whereas fold change above 2.0 is considered as up-regulated. The fold change method is not completely reliable as statistical variability in the data is not considered. The two-sample t-test is widely used as a parametric hypothesis testing method for identification of differentially expressed genes. The t-statistics gives a probability value (p-value) for each gene. A small p-value indicates that genes are differentially expressed under the hypothesis that there is no differential expression, which is not true. The t-test for unpaired data and for both equal and unequal variance can be computed as,

$$t_i = \frac{\bar{y}_i - \bar{x}_i}{\sqrt{\frac{g_i^2}{n_1} + \frac{h_i^2}{n_2}}} \quad (1)$$

where x_i and y_i are the means, g_i and h_i are the variances, and n_1 and n_2 are the sizes of the two groups of the sample (conditions) tumor and normal, respectively, of gene expression profile i.

3. INFORMATION THEORY

The information theory has been popularly used in communication systems. But now, it is being applied in many other areas such as signal processing, natural language processing, statistical inference, cryptography and other forms of data analysis. Information theory is closely associated with a large number of pure and applied areas such as artificial intelligence, adaptive systems, machine learning, and complex systems to name a few. It is basically based on the probability theory and statistics. Entropy and mutual information are the two fundamental constituents of information theory, which are described as follows:

3.1. Entropy

Entropy is a key measure used in information theory that quantifies the uncertainty involved in predicting the value of a random variable. It is a measure of uncertainty that measures the amount of information needed on the average to describe the random variable. The entropy H(X) of a discrete random variable X can be defined as,

$$H(X) = -\sum_x p(x) \log p(x) \quad (2)$$

where, p(x) is the probability mass function. The entropy is non-negative because $0 \leq p(x) \leq 1$, which implies $\log(1/p(x)) \geq 0$. The definition of entropy can be extended for a pair of random variables, called joint entropy. The joint entropy H(X,Y) of two discrete random variables (X,Y) with join probability distribution $p(x,y)$ can be defined as,

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y) \quad (3)$$

The conditional entropy of a random variable, given another random variable, can be defined as,

$$\mathrm{H}(\mathrm{X} \mid \mathrm{Y}) = \sum_y p(y)\, H(X \mid Y = y) \quad (4)$$

$$= -\sum_y p(y) \sum_x p(x \mid y) \log p(x \mid y) \quad (5)$$

$$= -\sum_x \sum_y p(x,y) \log p(x \mid y) \quad (6)$$

In fact, the joint entropy of a pair of random variables is the entropy of one plus conditional probability of other variable, such as,

$$\mathrm{H}(\mathrm{X},\mathrm{Y}) = H(X) + H(Y \mid X) \quad (7)$$

3.2. Mutual Information

Mutual information is another widely used measure in information theory. It accounts for the amount of information that one random variable holds about another random variable. The mutual information *MI*(*X*;*Y*) of two random variables *X* and *Y* corresponds to the intersection of the information in *X* with the information in *Y*. It can be computed as,

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \quad (8)$$

where entropy *H* and joint entropy *H*(*X*,*Y*) can be computed using equation (2) and equation (3), respectively.

Information theoretic approach, particularly mutual information, has been used in several GRN inference methods. It measures the interactions and relations between genes within a cell using gene expression profiles. The earliest mutual information (MI) based method for GRN inference, called relevance network (RN), was developed by Butte and Kohane [21]. The edges are assigned to gene-pairs, only if corresponding mutual information value is above the threshold. The networks inferred from relevance network are association networks due to the fact that edges represent the association between two genes but are not necessarily a causal effect. Other type of inference methods is also available that intend to find out causal regulatory interactions between genes and their products that can be validated with biological experiments, available databases and literatures. Till now, we do not have any well accepted procedure for GRN inference and their analysis at molecular level. Essential data preprocessing is required to make them ready for GRN inference. These preprocessing steps involve standardized procedures for the normalization of gene expression data within and between samples [22].

4. TOPOLOGICAL PROPERTIES OF THE NETWORK

It is interesting to observe that despite the difference in nature of network; most real-world networks share common properties. The topological properties of the network give us valuable insights about internal organization of a biological network, for instance, partitioning regulatory networks into a functional and feasible structure. Some of the basic network properties, which are commonly analyzed are describe in the following section. Detailed discussions on complex networks, its structure and dynamics can be found in [23].

4.1. Network Density

The network density shows how dense or sparse a network is based on number of connection per node. It is defined as a ratio of number of connections (*E)* to number of possible connections (i.e., *N*(*N*-1)/2 for *N* node network), which is given by,

$$Density = \frac{2E}{N(N-1)} \quad (9)$$

4.2. Degree Distribution

The degree of a node is the most elementary network measure which gives the number of connections (*k)* that a node has. The degree of overall network is the average degree <*K*> of the network. Since, average degree does not capture potential variation in the network; hence, it is better to measure the degree distribution *P*(*k*), which is the number of nodes having exactly *k* connections.

Let *k* be degree of node, then degree distribution is given by,

$$P(K = k) = f(k) = \frac{N_k}{N} \quad (10)$$

where, *f*(*k*) is a probability distribution, N_k is number of nodes with degree *k*=1,2,… and *N* is total number of nodes in the network. The degree distribution of several real networks including some biological networks follow power-law [24], which is given by,

$$P(K = k) \approx ak^{-\gamma} \quad (11)$$

where, γ is a constant representing degree exponent.

4.3. Clustering Coefficient

Nodes of several real networks exhibit a tendency to cluster. This behavior of the network can be quantified using clustering coefficient [25]. This measure shows propensity of a network to be grouped into clusters, where a cluster is a subset of vertices containing numerous edges which connect them to one another. In other words, the clustering coefficient measures the degree to which neighbors of a particular node are connected to each other. The clustering coefficient *C* of node *i* can be computed as,

$$C_i = \frac{2n_i}{k_i(k_i-1)} \quad (12)$$

where n_i represents number of connections joining k_i neighbors of node *i* to each other. Thus, C_i measures the ratio of number of connections between neighbors of node *i* to total possible connections. Clustering coefficient always takes values as $0 \leq C_i \leq 1$. It is considered that closer the value of C_i is to 1, more likely network would form clusters.

Similarly, average clustering coefficient <*C*> can be defined as,

$$< C >= \frac{1}{N}\sum_{i=1}^{N} C_i \quad (13)$$

Biological networks have been found to have higher average clustering coefficient showing that they are naturally modular [3].

4.4. Average Shortest Path Length

In most of the networks, there may be several paths between two nodes. An appropriate distance measure between two nodes, *i* and *j,* would be length of the shortest path, l_{ij}. The average path length <*l*> can be defined as,

$$< l >= \frac{2}{N(N-1)} \sum_{i,j=1}^{N} l_{ij} \quad (14)$$

The average path length more specifically can be defined as the average number of connections between nodes that must cross shortest path between two nodes. The average shortest path length is also known as *characteristic path length*. The most popular algorithms for finding shortest paths are Dijkastra's and Floyd's algorithms [26].

4.5. Network Diameter and Radius

Eccentricity is defined as the largest geodesic distance of a node. The network diameter is the maximum eccentricity of all the nodes of a network. The network diameter gives the number of steps that are sufficient to travel between any two nodes. On the other hand, network radius is the minimum eccentricity of all the nodes in the network.

4.6. Centralization

Centralization is a measure that assesses if a network has star-like structure, or if each node of the network has same degree, on average. A network centralization value close to 1 indicates that the network has star-like topology, whereas a value close to zero signifies that there is a higher likelihood for the network to have same connectivity. It can be defined as,

$$Centralization = \frac{n}{n-2}\left(\frac{\max(k)}{n-1} - Density\right) \quad (15)$$

4.7. Neighborhood Connectivity Distribution

The neighborhood connectivity of a node *i* can be defined as the average connections of all the neighbors of *i*. The neighborhood connectivity distribution depicts the average neighborhood connections of all nodes *i* with *k* neighbors for *k*=0,1,…

4.8. Network Heterogeneity

Network heterogeneity shows the propensity of a network to contain hub nodes. This property also shows several unique properties of scale-free networks; for example, scale-free networks possess robust-yet-fragile property, which means that the networks are robust against random failure of nodes but fragile to intentional attack [27]. Hence, the measure of heterogeneity is important for understanding the behavior and function of networks. Network heterogeneity is based on variance of connectivity. It can be defined as coefficient of variation of the connectivity distribution given by [28],

$$Heterogeniety = \frac{\sqrt{variance(k)}}{mean(k)} \quad (16)$$

Biological networks tend to be very heterogeneous in the sense that some genes are highly connected and most genes have lesser number of connections.

4.9. Shared Neighborhood

Let $P(i, j)$ be the number of partners shared between nodes *i* and *j*, i.e., nodes which are neighbors of both *i* and *j*, then shared neighbors distribution gives the number of node pairs (i, j) with $P(i, j)=k$, for k=1,2,… If a motif is over-represented in a biological network, then it may be determined using shared neighborhood distribution.

4.10. Topological Coefficient

Topological coefficient is a relative measure that quantifies the extent to which a node shares neighbors with others. Generally, it is calculated for all the nodes with more than one neighbor. Nodes having one or zero neighbors are assigned a topological coefficient of zero. The topological coefficient of a node *i* having k_i neighbors can be computed as,

$$T_i = \frac{avg(P(i,j))}{k_i} \quad (17)$$

where, $P(i, j)$ is defined for all nodes *j* which share at least one neighbor with *i*.

4.11. Network Centrality

The idea of centrality was originally pursued by social network analysts to find the most popular person in the group, or the person who stands at the center of attention. In biological networks, finding central molecules is important for the reason that these would interact with several other molecules and directly affect the network topology. There are a number of measures of centrality, which have been described below:

4.11.1. Degree Centrality

It is well known that more choices mean more opportunities and less dependence. Similarly, the more connection nodes of a network have, the more power they may have. The degree centrality is a measure that shows the number of connections a node has. When connections are directed, it can be calculated as the total number of in-degree and out-degree of the node. Nodes having very high degree of centrality are known as hubs because they are connected to several neighbors. Any perturbation in central nodes may lead to serious consequences on the topology of the network. As reported, biological networks are robust against random perturbations, but any disruption of hubs may lead to system collapse [27].

4.11.2. Closeness Centrality

Degree centrality takes into account only the immediate connections that a node has rather than indirect connections to all other nodes. One node might be connected to a large number of other nodes, but those other nodes might be

disconnected from the network as a whole. Hence, the node would be quite central, but only in a local neighborhood. Closeness centrality emphasizes on the distance (farness) of a node to all others in the network by focusing on the distance from each node to all others. If the distance of a node is the sum of its distances to all other nodes in the network then the closeness centrality can be defined as the inverse of the distances, as

$$Closenss\ Centrality, C_{close}(i) = \frac{1}{\sum dist(i,j)} \quad (18)$$

where, *dist(i,j)* represents distance (or shortest path) between the nodes *i* and *j*. Therefore, it can be stated that the more central a node is, the lower is its total distance to all other nodes. Closeness centrality measure can be used to assess the time it takes to spread information from a given node to all other nodes in the network. It can be applied to the biological network for the identification of top central metabolites, proteins and genes in large-scale networks.

4.11.3. Betweenness Centrality

Several times a node depends on other node to make connections with others. Betweenness centrality measures the number of times a node acts as a bridge for two communicating nodes in a network. Suppose, σ_{ij} be total shortest paths between distinct nodes *i* and *j,* and $\sigma_{ij}(w)$ be the total shortest paths between *i* and *j via* node *w*. Let *V*(*i*) be the set of all ordered pairs, (*i,j*) such that *i*, *j* and *w* are distinct, then the betweeness centrality can be computed as,

$$Betweeness\ Centrality, C_{bet}(w) = \sum_{(i,j)\in V(w)} \frac{\sigma_{ij}(w)}{\sigma_{ij}} \quad (19)$$

The proteins and genes having high betweenness centralities are termed as "bottlenecks" for their role as key connectors.

4.12. Small World Property

Small world property is one of the important organizing principles of a network that tells that two nodes are likely to be connected by a relatively short path length. In other words, a network having relatively small diameter is often known as a small world network [29]. For instance, social networks are so rich in short path length and form a small world network, also known as "Six degrees of separation". Several real world networks including food chains, road maps, scientists' collaboration network, electrical power grids, neural networks of brain and metabolic networks form the 'small world' architecture.

4.13. Scale-Free Property

One of the important properties of network topology is degree distribution, i.e., the distribution of how many edges each node has. A network is called scale-free if its degree distribution follows the power-law. It is reported in literature that several types of biological networks are scale-free [30]. A scale-free network has several characteristics such as, i) there is a short path between any two nodes of the network (called 'small-world' property), ii) there are several nodes with few connections, and few nodes with large connections (called hubs), iii) scale-free networks are robust to random breakdowns and invariant to change in scale and iv) these networks are sensitive to attack on hubs [27].

5. GENE ONTOLOGY ENRICHMENT ANALYSIS

Ontology can be defined as representation of things, which are observable, and the relationships between those things. Gene Ontology (GO) is a well-defined structure vocabulary for biological terms. It is *de facto* standard for gene functionality prediction. It is widely used for gene functional annotation and enrichment analysis that describes molecular function, biological process and cellular components of genes and gene products in a species-neutral manner. The GO is being used for gene function understanding, pathway analysis and network modeling, to name a few. The advantage of GO is that it provides information exchange between different biological communities. The GO ontology files may be freely accessed in a number of formats by using popular GO browser, e.g. AmiGo[1].

BiNGO [31] is a software tool that determines which GO categories are statistically overrepresented in a given set of genes or a subgraph of a biological network. BiNGO is available as a plugin in Cytoscape[2] software tool, which is an open source network platform for network visualization and integration of molecular interaction networks. BiNGO maps the function of given set of genes on GO hierarchy and outputs the generated mapping as a graph in Cytoscape. An important advantage of BiNGO over other GO tools is that it may be applied directly and interactively on interaction graphs and can take full advantages of Cytoscape's versatile visualization environment. BiNGO plugin applies Benjamini & Hochberg correction to limit false discovery rate (FDR). In this paper, we have used BiNGO to perform GO based enrichment analysis for the identified network modules to find out groups of genes sharing common biological pathways. BiNGO has been used to determine the GO terms that are significantly overrepresented in selected set of genes.

6. PROPOSED ALGORITHM

The proposed algorithm for inferring gene regulatory network modules is outlined as follows. A sketch of the proposed algorithm is shown in Fig. (**1**).

6.1. Data Preprocessing and Normalization

Gene expression profiles, generated by microarray technology, are noisy due to experimental limitations. Before using gene expression profiles, we must go through some preprocessing steps to handle duplicate attributes, missing values, outliers, and normalizing the data to bring it into a particular range. Missing values have been replaced by the average value of gene profile. We eliminated duplicate gene profiles and also eliminated those gene profiles whose expression values are missing for more than half the number of samples. Finally, data are normalized using Min-Max normalization technique to scale in the range [0,1]. Normalized expression value of e_i for a sample E in the i^{th} row can be calculated using following equation:

[1]http://www.amigo.geneontology.org
[2]http://www.cytoscape.org/

Algorithm: *Inference and analysis of cancerous gene regulatory modules*

Step 1. *Preprocess microarray data*:
handling duplicate attributes, missing & noisy values, and data normalization

Step 2. *Filtering most significant genes using t-test statistics*:
p-value<=threshold

Step 3. *Computation of pair-wise MI & elimination of weak correlation links*:
MI(x,y)>=threshold

Step 4. *Computation of adjacency matrix & network generation*:

Source	Target	Interactions
G1	G2	1
G1	G3	0
…	…	…
Gn	Gm	1

Step 5. *Validation of the results*:
Biological databases & literature search

Step 6. *Topological analysis*:
Identification of hub genes, network centrality, heterogeneity, etc.

Step 7. *Gene Ontology analysis*:
Identification of GO terms.

Fig. (1). Proposed algorithm for the inference and analysis of cancerous gene regulatory modules.

$$Normalized(e_i) = \frac{e_i - E_{min}}{E_{max} - E_{min}} \quad (20)$$

where, E_{max} and E_{min} are maximum and minimum values for sample E, respectively.

6.2. Identification of Significant Genes

Before gene expression data is analyzed, first it is ensured that the dataset includes genes that differ in their expression level significantly between two classes of samples. Many methods are available for identification of differentially expressed genes, as discussed in Section 2. Due to wide applications and significant results of t-test statistics for samples having two different classes (e.g. cell types, cancer types, or experimental conditions), we applied t-test to identify differentially expressed genes between two classes, i.e., normal and tumor.

6.3. Computation of Pairwise Mutual Information Among Gene Pairs

The Mutual Information (MI) is the generalisation of pairwise correlation coefficient used to compare expression profiles of a set of microarrays and to measure the degree of independence between two genes. For each pair of genes (x, y), their mutual information MI(x, y) can be computed using equation (2), equation (3) and equation (8). The edges $a_{xy}=a_{yx}$ of the network is set to either 0 or 1, depending on a significant threshold for MI value. From the definition of MI, it becomes zero if the two random variables, x and y, are statistically independent (i.e., p(x, y)=p(x)p(y)), as their joint entropy is given by H(x, y)= H(x) + H(y). A higher MI value specifies that the two genes are non-randomly connected to each other. MI describes an undirected graph because it is symmetric, i.e., MI(x, y)=MI(y, x). MI is more generalized than the Pearson correlation coefficient (PCC) because it quantifies only linear dependencies between variables. According to the definition of MI, it needs each sample to be statistically independent from others, and thus, this approach can be used for steady-state as well as time-series gene expression data. The topological properties and GO-enrichment analysis are already discussed in the previous sections.

7. RESULTS & DISCUSSIONS

We applied the proposed algorithm on microarray data of circulating plasma RNA of colorectal cancer (CRC) consisting of 20 samples, taken from different donors. Out of 20 samples, 12 belong to colon tumors and 8 have been taken from normal biopsies. The dataset contains expression profiles of 15,552 genes. Their expression values were obtained by measuring relative abundance of different RNA species in plasma using cDNA microarray hybridization technique, by comparing RNA isolation and amplified from CRC patients and healthy patients. This dataset was published by Collado and his co-workers [32] and we downloaded the full dataset from Gene Expression Omnibus (Accession Number: GSE4988). Fig. (**2**) shows the comparative view of gene expression of different samples for both colon cancer samples and normal samples. The expression of sample profiles in normal tissue is higher in comparison to that of cancer tissue in most of the cases. Many of the cancer sample profiles are down-regulated, as shown in Fig. (**2**).

Gene expression data contains a large number of genes, the majority of which may not be relevant for analysis. We

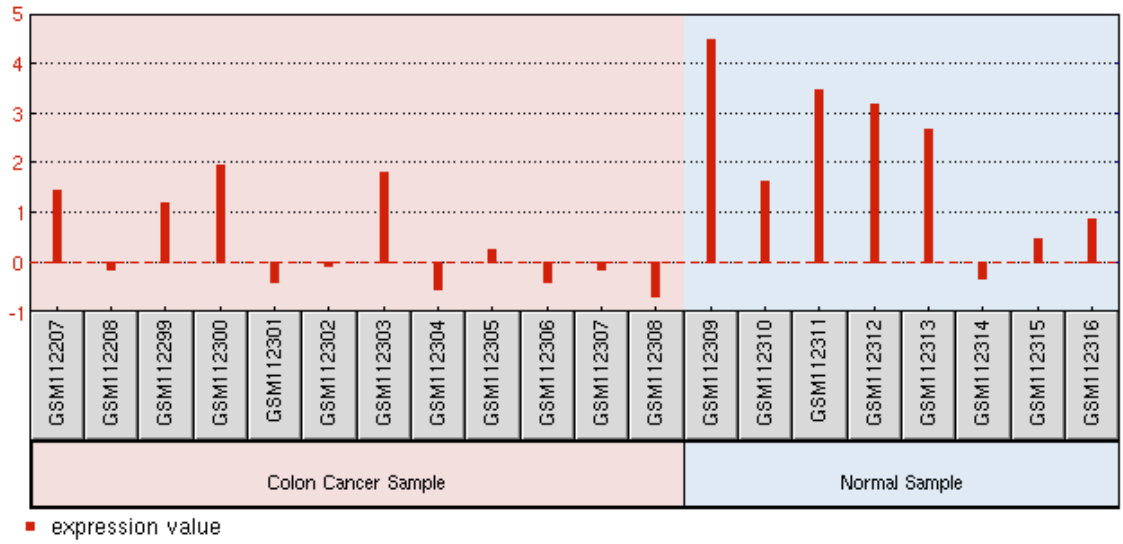


Fig. (2). Sample profile graph showing expression values in cancer and normal sample.

applied t-test statistics to select most significant genes whose p-values<0.01. In this way, we found 101 most significant genes that have been selected for further analysis. To find the regulatory interactions among the selected significant genes, mutual information between gene pairs has been computed using equation (2), (3) and (8). Now, gene interaction network has been constructed, where nodes correspond to gene names and pair-wise mutual information is allocated to the edge between genes. Initially, we took the top 30 highest pair-wise MI values (can be assumed as interaction weight) for the network construction and found a network of 22 genes, which is shown in Fig. (**3**).

From Fig. (**3**), it is clear that gene ACAT2 is highly connected with a degree of 9 and is regulating a large number of genes. Similarly, we constructed nine other networks by taking top 40, 50, 60, 70, 80, 90, 100, 250 and 500 MI values and observed the five highly connected genes in each case. The observation of five highly connected genes in each of the network is shown in Table **1**. From Table **1**, it is clear that as the numbers of interactions increase, the degree of each hub genes also increases. The network 10 considers 500 interactions that involves 79 genes, in which five highly connected genes are ACAT2(54), CYP1B1(50), NPM1(48), COX15(46), CREM(42), where numbers in parenthesis shows the connection degrees.

The identification of highly connected genes (hubs) may play a vital role in cancer diagnosis and therapies. The extracted genes have been validated with the available biological databases and literature. It was found that most of the identified genes including ACP, LDHA, SPARCL, EPAS1, MVP, OXA1L, and RPL10A, in additional to several others, are involved in the colon cancer. All the identified interactions among genes including highly connected genes need biological experimental validation for its reliability.

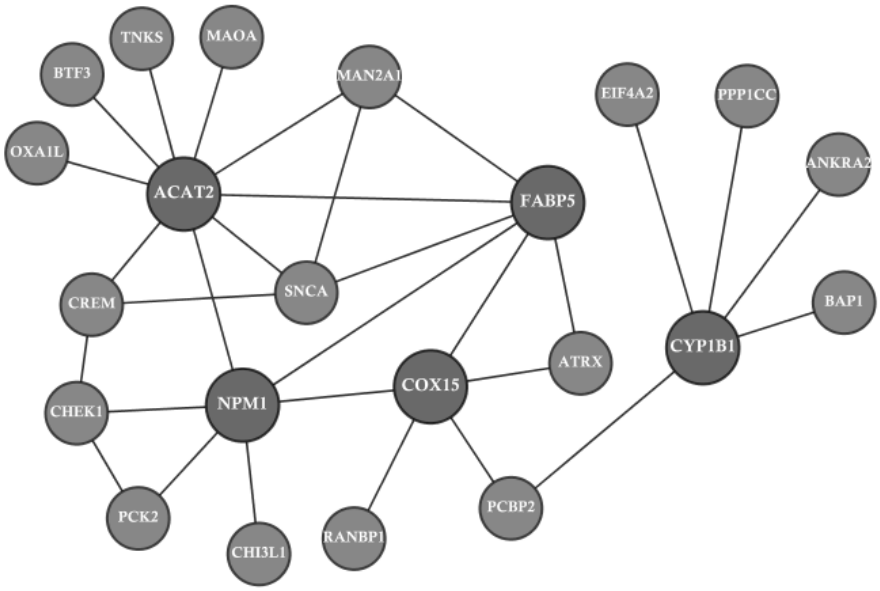


Fig. (3). Inferred network of 22 genes and 30 interactions.

Table 1. Ten different networks, number of interactions, number of genes involved in each, five highly connected genes with their degrees.

Network No.	No. of Interactions	Number of Genes	Top Five Genes with Highest Degree
Network 1	30	22	ACAT2 (9), FABP5 (6), NPM1(6), COX15(5), CYP1B1(5)
Network 2	40	25	ACAT2(9), FABP5(9), NPM1(8), CYP1B1(8), CREM(6)
Network 3	50	27	ACAT2(12), NPM1(11), CYP1B1(10), FABP5(10), SNCA (7)
Network 4	60	29	ACAT2(14), NPM1(13), CYP1B1(11), FABP5(10), SNCA(7)
Network 5	70	30	ACAT2(14), NPM1(13), COX15(12), CYP1B1(12), FABP5(10)
Network 6	80	31	CYP1B1(16), ACAT2(14), NPM1(13), COX15(12), SNCA(11)
Network 7	90	35	ACAT2(17), CYP1B1(16), NPM1(14), SNCA(13), TNKS(13)
Network 8	100	36	NPM1(18), ACAT2(17), CYP1B1(16), SNCA(13), TNKS(13)
Network 9	250	56	CYP1B1(34), NPM1(33), COX15(30), ACAT2(30), CREM(28)
Network 10	500	79	ACAT2(54), CYP1B1(50), NPM1(48), COX15(46), CREM(42)

7.1. Analysis of Inferred Colon Cancer Networks

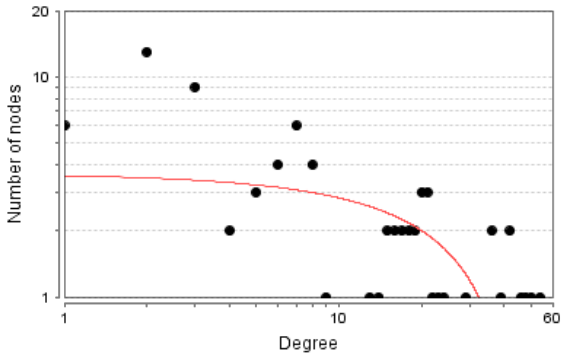
We have used Cytoscape tool with NetworkAnalyzer plugin for the analysis of inferred networks. NetworkAnalyzer [33] computes a list of simple as well as complex topological parameters for both directed and undirected networks. In this section, we have done topological analysis of colon cancer network modules that we inferred using proposed algorithm. Here, we have done analysis of three networks only with number of interactions as 100, 500 and entire inferred interactions of 5050 among 101 genes. Different network properties along with their values for three different networks are shown in Table **2**.

As we know, high clustering coefficient is the signature of network's modularity. From the Table **2**, we can observe that clustering coefficient of the network increases and reaches a value of 1.0 as number of interactions is increased. Hence, we can say that inferred networks are modular. The network diameter decreases from 4.0 to 1.0 and the radius also decreases from 3.0 to 1.0, which indicates that genes are more approachable when the number of interactions is increased. The network centralization and heterogeneity value gear up slightly for second network, but finally, it became zero for the third network. The shortest path between nodes of a network is the path connecting all the nodes with smallest number of steps. The shortest paths of the three networks are 1260, 6162 and 10100. As the number of interactions increase, characteristic path length decreases and corresponding average number of neighbors increases. The number of isolated nodes and self-loops are zero for all the three considered networks, which shows that no genes in the network are isolated and there is no self-regulation. All the three networks are scale-free because its degree distribution follows power-law and their degree exponents (γ) are shown in Table **2**.

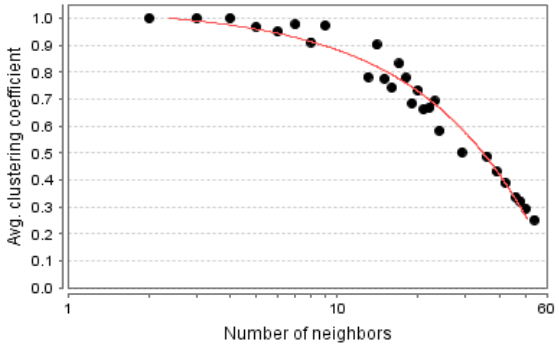
The Fig. (**4a-h**) depict the number of analyses for the second network, i.e. network having 500 interactions

Table 2. Simple parameters of three different colon cancer network modules.

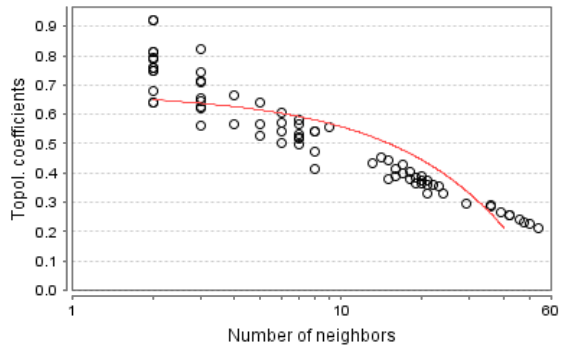
Network Properties	Module 1: 100 Interactions	Module 2: 500 Interactions	Module 3: 5050 Interactions
Clustering Coefficient	0.4905	0.7657	1.0
Network Diameter	4	3	1
Network Radius	3	2	1
Network Centralization	0.3764	0.5438	0
Shortest Paths	1260 (100%)	6162 (100%)	10100 (100%)
Characteristic Path Length	2.2349	1.9925	1.0
Avg. Number of Neighbors	5.5555	12.6582	100
Network Density	0.1587	0.1622	1
Network Heterogeneity	0.9143	1.0542	0
Isolated Nodes	0	0	0
Number of Self-loop	0	0	0
Scale-free	Yes	Yes	Yes
Degree Exponent (γ)	0.658	0.560	0.470



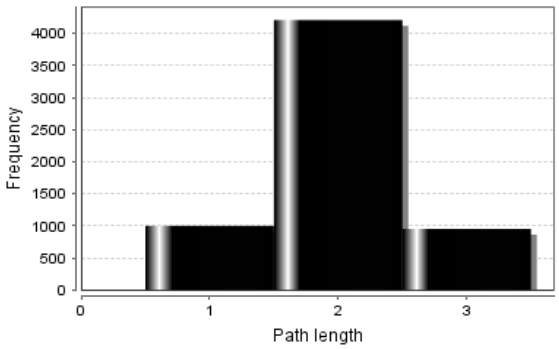
(**a**) Degree distribution



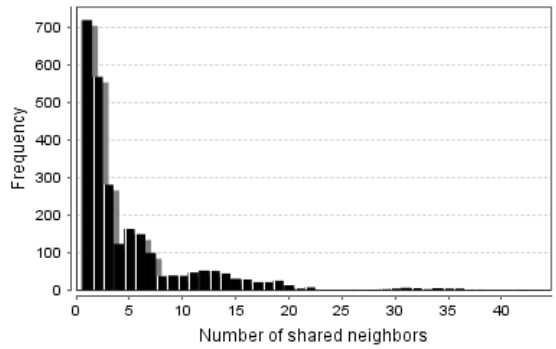
(**b**) Average clustering coefficient



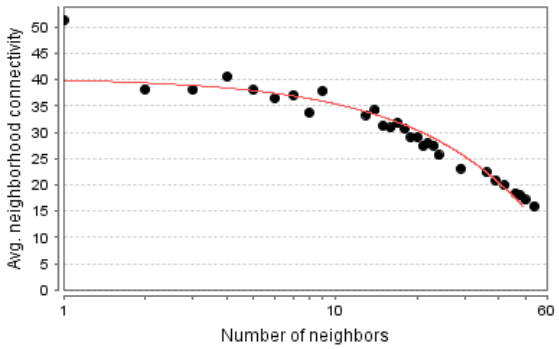
(**c**) Topological coefficient



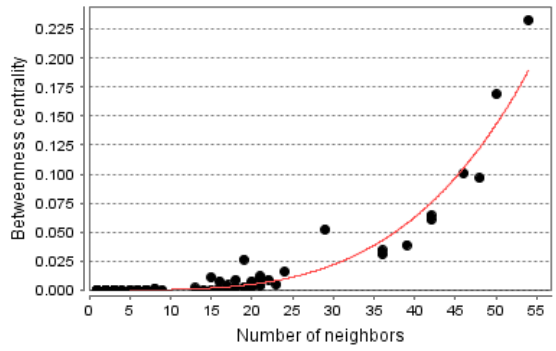
(**d**) Shortest path length distribution



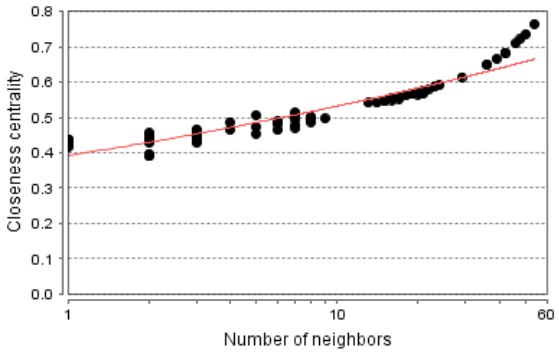
(**e**) Shared neighbours distribution



(**f**) Average neighbourhood connectivity distribution



(**g**) Betweenness centrality



(**h**) Closeness centrality

Fig. (4a-h). Showing various topological measures of inferred network module consisting of 500 interactions involving 79 genes.

involving 79 genes. Fig. (**4a**) shows that most of the genes have less degree and only few are having highest degree. Hence, we can say that network has few densely connected genes (hubs) and mostly sparsely connected genes. Fig. (**4b**) shows average clustering coefficient distribution, which is the average of the clustering coefficients for all genes in the network. The diagram shows decreasing trends in average clustering coefficient as number of neighboring genes increases. Fig. (**4c**) shows topological coefficient, which is a relative measure showing the extent to which a gene shares its neighbors with other genes. The topological coefficient shows a decreasing trend, as neighbors are increased. Fig. (**4d**) shows shortest path length distributions that provides the number of node pairs (i,j) with $L(i, j)=k$ for k=1,2,… The shortest path length distribution indicates the small world properties of the network. There are 6,162 number of shortest path, out of which more than 4,000 paths have a length of 2 or we can say that majority of the genes have path lengths of 2 only. Around 1,000 paths have length of 1 and 3, as shown in Fig. (**4d**).

The shared neighbor distribution gives the number of node pairs (i, j) with $P(i,j)=k$ for k=1,2,…, as shown in Fig. (**4e**). Fig. (**4f**) shows average neighborhood connectivity distribution which provides average of neighborhood connectivity of all nodes with k neighbors for k=0,1,… Fig. (**4f**) is showing a decreasing function of k, edges between low connected and highly connectives prevail the network. Biologically, betweenness centrality indicates the relevance of a gene (or protein) as functionally capable of holding together interacting genes. Fig. (**4g**) shows an increasing trend in betweenness centrality, which indicates that genes are functionally capable of holding together interacting genes. Fig. (**4h**) shows closeness centrality of all the nodes plotted against the number of neighbors. It measures how fast information spreads from a given gene to other reachable genes in the network. Fig. (**4h**) shows an increasing trend in closeness centrality, varying from 0.4 to 0.8.

7.2. GO Enrichment Analysis of the Inferred Network Modules

The Tables **3-5** list the overrepresented GO categories for three identified modules: module 1, module 2 and module 3, respectively. The columns include the GO-ID terms of the category, enrichment significance p-value, Corr. p-value, cluster frequency, total cluster frequency, description and sets of genes annotated to that category.

7.3. Results on Simulated Benchmark Datasets

To validate the proposed method, we took simulated benchmark dataset of yeast, which is a part of DREAM3 challenge, consisting of 10 genes and 25 interactions. The algorithm is applied on this dataset, except step 2 because it is a benchmark dataset in which all the genes are significant. The predicted results were evaluated in terms of true positives (TP), true negative (TN), false positive (FP) and false negative (FN). The used evaluation measures are as follows:

$$Recall/TPR/Sensitivity = \frac{TP}{TP+FN} \quad (21)$$

$$Specificity = \frac{TN}{TN+FP} \quad (22)$$

$$Precision = \frac{TP}{TP+FP} \quad (23)$$

$$FPR = \frac{FP}{FP+TN} \quad (24)$$

$$F-Score = \frac{2*Precision*Recall}{Prevision+Recall} \quad (25)$$

where, TPR is true positive rate and FPR is false positive rate. Another popularly used metric is Area Under Receiver Operating Characteristics (AU-ROC) curve. It is a simple metric used to determine how an algorithm performs over the

Table 3. GO terms significantly enriched with identified module 1 with 100 interactions. Overrepresentation, using Hypergeometric statistical test, Benjamini & Hochberg False Discovery Rate (FDR) correction and significance level (p-value) <= 0.01.

GO-ID	P-Value	Corr P-Value	Cluster Frequency	Total Frequency	Description	Genes in Test Set
44237	0.0000	0.0031	26/36=72.22%	4989/14303=34.88%	cellular metabolic process	ADSS|OXA1L|CYP1B1|CSNK1G2|MAOA|CREM|SNCA|CCNT1|ANXA1|RPS15A|BAP1|TOPBP1|CHEK1|PPP1CC|PCK2|ACAT2|ATRX|MAN2A1|PGM5|EIF4A2|PCBP2|BTF3|TNKS|KPNA2|RPS23|COX15
6066	0.0000	0.0031	8/36=22.22%	433/14303=3.03%	alcohol metabolic process	MAN2A1|PGM5|MAOA|CREM|SNCA|PPP1CC|PCK2|ACAT2
8152	0.0000	0.0031	28/36=77.78%	5957/14303=41.65%	metabolic process	ADSS|OXA1L|CYP1B1|CREM|CCNT1|SNCA|BAP1|RPS15A|CHEK1|ACAT2|PCBP2|BTF3|TNKS|RPS23|COX15|CSNK1G2|MAOA|ANXA1|CHI3L1|TOPBP1|PCK2|PPP1CC|ATRX|MAN2A1|PGM5|EIF4A2|KPNA2|FABP5
44238	0.0000	0.0038	26/36=72.22%	5286/14303=36.96%	primary metabolic process	ADSS|CYP1B1|CREM|CCNT1|SNCA|BAP1|RPS15A|CHEK1|ACAT2|PCBP2|BTF3|TNKS|RPS23|CSNK1G2|MAOA|ANXA1|CHI3L1|TOPBP1|PCK2|PPP1CC|ATRX|MAN2A1|PGM5|EIF4A2|KPNA2|FABP5

Table 4. GO terms significantly enriched with identified module 2 with 500 interactions. Overrepresentation, using Hypergeometric statistical test, Benjamini & Hochberg False Discovery Rate (FDR) correction and significance level (p-value) <= 0.01.

GO-ID	P-Value	Corr P-Value	Cluster Frequency	Total Frequency	Description	Genes in Test Set
44237	0.0000	0.0002	50/78=64.1%	4989/14302=34.8%	cellular metabolic process	XRCC5|NDUFB3|ADSS|HLF|CAV1|LDHA|OXA1L|CYP1B1|CREM|USP9X|CCNT1|SNCA|TGFB3|RPS15A|BAP1|CHEK1|ACAT2|ACP1|CBR1|FRG1|PCBP2|PKD2|BTF3|FAU|TNKS|RPL10A|PIK3R1|RPS23|GLRX|COX15|EPAS1|CSNK1G2|MAOA| ANXA1|TOPBP1|BAD|CRAT|DECR1| PCK2| PPP1CC|SMN2|ATRX|MAN2A1|PGM5|RPS6KA1|TBCA|EIF4A2|RELN| KPNA2|ALOX12
8152	0.0000	0.0016	53/78=67.95%	5957/14302= 41.65%	metabolic process	XRCC5|HLF|OXA1L|LDHA|SNCA|CCNT1|TGFB3|BAP1|PCBP2|FAU|COX15|CSNK1G2|LYZ|TOPBP1|DECR1|PPP1CC|SMN2|MAN2A1|PGM5|TBCA|EIF4A2|RELN|KPNA2|ALOX12|NDUFB3|ADSS|CAV1|CYP1B1|USP9X|CREM|RPS15A|CHEK1|ACAT2|ACP1|CBR1|FRG1|PKD2|BTF3|TNKS|RPL10A|RPS23|PIK3R1|GLRX|EPAS1|MAOA|CHI3L1|ANXA1|BAD|CRAT|PCK2|ATRX|RPS6KA1|FABP5
6066	0.0000	0.0017	12/78=15.38%	434/14302= 3.03%	alcohol metabolic process	MAN2A1|HLF|LDHA|PGM5|EPAS1|MAOA|CREM|SNCA|BAD|PCK2|ACAT2|PPP1CC
44238	0.0000	0.0050	47/78= 60.26%	5286/14302= 36.96%	primary metabolic process	XRCC5|HLF|ADSS|LDHA|CAV1|CYP1B1|CREM|USP9X|CCNT1|SNCA|TGFB3|RPS15A|BAP1|CHEK1|ACAT2|ACP1|FRG1|PCBP2|PKD2|BTF3|FAU|TNKS|RPL10A|PIK3R1|RPS23|EPAS1|CSNK1G2|MAOA|ANXA1|CHI3L1|TOPBP1|BAD|CRAT|DECR1|PCK2|PPP1CC|SMN2|ATRX|MAN2A1|PGM5|RPS6KA1|TBCA|EIF4A2|RELN|KPNA2|FABP5|ALOX12
30258	0.0000	0.0050	5/78=6.41%	78/14302= 0.45%	lipid modification	CRAT|DECR1|ACAT2|PIK3R1|ALOX12
1765	0.0000	0.0050	2/78=2.56%	2/14302=0.01%	membrane raft assembly	CAV2|CAV1
70836	0.0000	0.0050	2/78=2.56%	2/14302=0.01%	caveola assembly	CAV2|CAV1
6584	0.0000	0.0050	4/78=5.13%	35/14302=0.24%	catecholamine metabolic process	HLF|EPAS1|MAOA|SNCA
34311	0.0000	0.0050	4/78=5.13%	35/14302=0.24%	diol metabolic process	HLF|EPAS1|MAOA|SNCA
9712	0.0000	0.0050	4/78=5.13%	35/14302=0.24%	catechol metabolic process	HLF|EPAS1|MAOA|SNCA
18958	0.0000	0.0050	4/78=5.13%	36/14302=0.25%	phenol metabolic process	HLF|EPAS1|MAOA|SNCA

(Table 4) contd…..

GO-ID	P-Value	Corr P-Value	Cluster Frequency	Total Frequency	Description	Genes in Test Set
9987	0.0000	0.0050	67/78=85.90%	9367/14302=65.49%	cellular process	XRCC5|HLF|COX11|OXA1L|LDHA|SNCA|CCNT1|TGFB3|BAP1|CBX1|TNFRSF11B|SERPINA5|PCBP2|AP3B2|FAU|RANBP1|COX15|CSNK1G2|SLC25A5|LYZ|MGP|STXBP3|TOPBP1|DECR1|PPP1CC|SMN2|MAN2A1|PGM5|TBCA|EIF4A2|RELN|KPNA2|MVP|ALOX12|NDUFB3|ADSS|CAV2|CAV1|CYP1B1|MCL1|USP9X|CREM|RPS15A|CHEK1|ACP2|ACAT2|ACP1|CBR1|FRG1|NPM1|PKD2|BTF3|HBP1|TNKS|RPL10A|PIK3R1|RPS23|GLRX|EPAS1|MAOA|ANXA1|BAD|CRAT|PCK2|CDKN1C|ATRX|RPS6KA1
19318	0.0001	0.0068	7/78=8.97%	185/14302=1.29%	hexose metabolic process	MAN2A1|LDHA|PGM5|CREM|BAD|PCK2|PPP1CC
65003	0.0001	0.0068	13/78=16.67%	676/14302=4.73%	macromolecular complex assembly	CAV2|CAV1|OXA1L|COX11|TBCA|NPM1|MGP|BAP1|STXBP3|DECR1|ACAT2|SMN2|COX15
7005	0.0001	0.0071	6/78=7.69%	132/14302=0.92%	mitochondrion organization	CAV2|MAN2A1|HLF|OXA1L|EPAS1|SNCA
32844	0.0001	0.0071	6/78=7.69%	133/14302=0.93%	regulation of homeostatic process	CAV1|TNFRSF11B|SNCA|PKD2|TNKS|ALOX12
6461	0.0001	0.0071	11/78=14.10%	507/14302=3.54%	protein complex assembly	CAV2|CAV1|OXA1L|COX11|TBCA|NPM1|MGP|BAP1|STXBP3|DECR1|COX15
70271	0.0001	0.0071	11/78=14.10%	507/14302=3.54%	protein complex biogenesis	CAV2|CAV1|OXA1L|COX11|TBCA|NPM1|MGP|BAP1|STXBP3|DECR1|COX15
6916	0.0001	0.0076	7/78=8.97%	199/14302=1.39	anti-apoptosis	MYD88|MCL1|BNIP3L|NPM1|SNCA|ANXA1|ALOX12
44281	0.0001	0.0076	19/78=24.36%	1371/14302=9.59%	small molecule metabolic process	HLF|ADSS|LDHA|CAV1|EPAS1|MAOA|CREM|SNCA|BAD|CRAT|DECR1|PCK2|PPP1CC|ACAT2|ATRX|MAN2A1|CBR1|PGM5|ALOX12
43933	0.0001	0.0085	13/78=16.67%	724/14302=5.06%	macromolecular complex subunit organization	CAV2|CAV1|OXA1L|COX11|TBCA|NPM1|MGP|BAP1|STXBP3|DECR1|ACAT2|SMN2|COX15
15980	0.0001	0.0085	6/78=7.69%	145/14302=1.01%	energy derivation by oxidation of organic compounds	NDUFB3|OXA1L|SNCA|CRAT|PPP1CC|COX15
6006	0.0001	0.0085	6/78=7.69%	146/14302=1.02%	glucose metabolic process	LDHA|PGM5|CREM|BAD|PCK2|PPP1CC
32846	0.0001	0.0085	4/78=5.13%	49/14302=0.34%	positive regulation of homeostatic process	CAV1|SNCA|TNKS|ALOX12
5996	0.0002	0.0098	7/78=8.97%	216/14302=1.51%	monosaccharide metabolic process	MAN2A1|LDHA|PGM5|CREM|BAD|PCK2|PPP1CC

Table 5. GO terms significantly enriched with identified module 3 with 5050 interactions. Overrepresentation, using Hypergeometric statistical test, Benjamini & Hochberg False Discovery Rate (FDR) correction and significance level (p-value) <= 0.01.

GO-ID	P-Value	Corr P-Value	Cluster Frequency	Total Frequency	Description	Genes in Test Set
44237	0.0000	0.0007	59/99=59.60%	4989/14301=34.89%	cellular metabolic process	XRCC5|HLF|LDHA|OXA1L|SNCA|CCNT1|TGFB3|BAP1|PTEN|PCBP1|PCBP2|FAU|COX15|CSNK1G2|TOPBP1|DECR1|PPP1CC|SMN2|MAN2A1|PGM5|TBCA|EIF4A2|PSMA3|RELN|CA2|CA1|KPNA2|ALOX12|NDUFB3|ADSS|CAV1|CYP1B1|USP9X|CREM|RPS15A|CHEK1|ACAT2|ATP5G3|ACP1|UBE2D3|CBR1|FRG1|PKD2|BTF3|TNKS|RPL10A|RPS23|PIK3R1|GLRX|NDUFA4|LPL|EPAS1|MAOA|ANXA1|BAD|CRAT|PCK2|ATRX|RPS6KA1
30258	0.0000	0.0035	6/99=6.06%	65/14301=0.45%	lipid modification	CRAT|DECR1|ACAT2|PTEN|PIK3R1|ALOX12
8152	0.0000	0.0035	63/99=63.64%	5957/14301=41.65%	metabolic process	XRCC5|HLF|LDHA|OXA1L|SNCA|CCNT1|TGFB3|BAP1|PTEN|PCBP1|PCBP2|FAU|COX15|CSNK1G2|LYZ|TOPBP1|DECR1|PPP1CC|SMN2|MMP12|MAN2A1|PGM5|TBCA|EIF4A2|PSMA3|RELN|CA2|CA1|KPNA2|ALOX12|NDUFB3|ADSS|CAV1|CYP1B1|USP9X|CREM|RPS15A|CHEK1|ACAT2|ATP5G3|ACP1|UBE2D3|CBR1|FRG1|PKD2|BTF3|TNKS|RPL10A|PIK3R1|RPS23|GLRX|NDUFA4|LPL|EPAS1|MAOA|CHI3L1|ANXA1|BAD|CRAT|PCK2|ATRX|RPS6KA1|FABP5
6066	0.0000	0.0035	13/99=13.13%	434/14301=3.03%	alcohol metabolic process	HLF|LDHA|EPAS1|CREM|MAOA|SNCA|BAD|PCK2|PPP1CC|ACAT2|PTEN|MAN2A1|PGM5
44281	0.0000	0.0048	24/99=24.24%	1371/14301= 9.59%	small molecule metabolic process	LPL|HLF|ADSS|CAV1|LDHA|EPAS1|MAOA|CREM|SNCA|BAD|CRAT|DECR1|PCK2|ACAT2|PPP1CC|ATP5G3|PTEN|ATRX|MAN2A1|CBR1|PGM5|CA2|CA1|ALOX12
32846	0.0000	0.0057	5/99=5.05%	49/14301=0.34%	positive regulation of homeostatic process	CAV1|SNCA|TNKS|CA2|ALOX12
32844	0.0000	0.0083	7/99=7.07%	133/14301=0.93%	regulation of homeostatic process	CAV1|TNFRSF11B|SNCA|PKD2|TNKS|CA2|ALOX12
1765	0.0000	0.0083	2/99=2.02%	2/14301=0.01%	membrane raft assembly	CAV2|CAV1
70836	0.0000	0.0083	2/99=2.02%	2/14301=0.01%	caveola assembly	CAV2|CAV1
6091	0.0001	0.0093	10/99=10.10%	312/14301=2.18%	generation of precursor metabolites and energy	NDUFA4|NDUFB3|OXA1L|LDHA|SNCA|CRAT|PPP1CC|ATP5G3|GLRX|COX15
15980	0.0001	0.0093	7/99=7.07%	145/14301=1.01%	energy derivation by oxidation of organic compounds	NDUFA4|NDUFB3|OXA1L|SNCA|CRAT|PPP1CC|COX15
9712	0.0001	0.0100	4/99=4.04%	35/14301=0.24%	catechol metabolic process	HLF|EPAS1|MAOA|SNCA
6584	0.0001	0.0100	4/99=4.04%	35/14301=0.24%	catecholamine metabolic process	HLF|EPAS1|MAOA|SNCA
34311	0.0001	0.0100	4/99=4.04%	35/14301=0.24%	diol metabolic process	HLF|EPAS1|MAOA|SNCA

(Table 5) contd…..

GO-ID	P-Value	Corr P-Value	Cluster Frequency	Total Frequency	Description	Genes in Test Set
9987	0.0001	0.0100	82/99=82.83%	9366/14301=65.49%	cellular process	XRCC5|HLF|LDHA|COX11|OXA1L|SNCA|CCNT1|TGFB3|BAP1|CBX1|PTEN|TNFRSF11B|PCBP1|SERPINA5|PCBP2|AP3B2|FAU|RANBP1|RARB|KCNQ2|COX15|SLC25A5|CSNK1G2|LYZ|MGP|TOPBP1|STXBP3|DECR1|PPP1CC|SMN2|MAN2A1|PGM5|TBCA|EIF4A2|PSMA3|RELN|CA2|CA1|KPNA2|MVP|ALOX12|NDUFB3|ADSS|CAV2|CAV1|CYP1B1|MCL1|CREM|USP9X|RPS15A|CHEK1|ACP2|ITM2B|ACAT2|ACP1|ATP5G3|CBR1|UBE2D3|FRG1|NPM1|PKD2|CHD1|BTF3|TNKS|HBP1|RPL10A|TRIP10|PIK3R1|RPS23|GLRX|NDUFA4|LPL|EPAS1|MAOA|ANXA1|BAD|CRAT|PCK2|ATRX|CDKN1C|RPS6KA1|DGKZ
7585	0.0001	0.0100	4/99=4.04%	36/14301=0.25%	respiratory gaseous exchange	MAN2A1|COX11|SFTPA2|COX15
18958	0.0001	0.0100	4/99=4.04%	36/14301=0.25%	phenol metabolic process	HLF|EPAS1|MAOA|SNCA

whole space showing relative trade-offs between TPs (benefits) and FPs (costs). An ROC curve is defined by FPR and TPR as x and y axes respectively. Because TPR is equivalent to sensitivity and FPR = (1-specificity), the ROC plot is also known as sensitivity *versus* (1 − specificity) plot. The AU-ROC can be calculated by using trapezoidal areas created between each ROC point.

Table 6. Confusion matrix.

		Predicted	
		1	0
Actual	1	TP=22	FN=3
	0	FP=3	TN=17

When proposed algorithm is applied on the said simulated benchmark dataset, out of 25 interactions, 22 interactions were truly identified. The confusion matrix is shown in Table **6**. The AU-ROC has been generated using Gene Network Weaver software tool [34] as shown in Fig. (**5**).

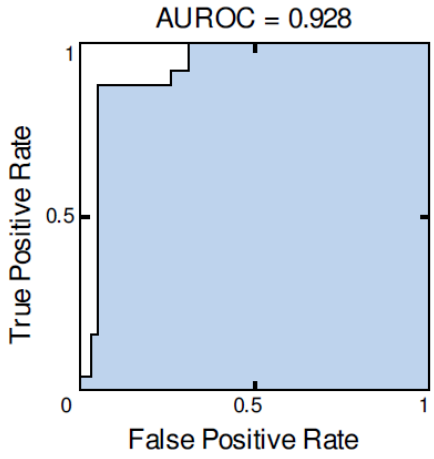


Fig. (5). Area Under ROC Curve (AUROC).

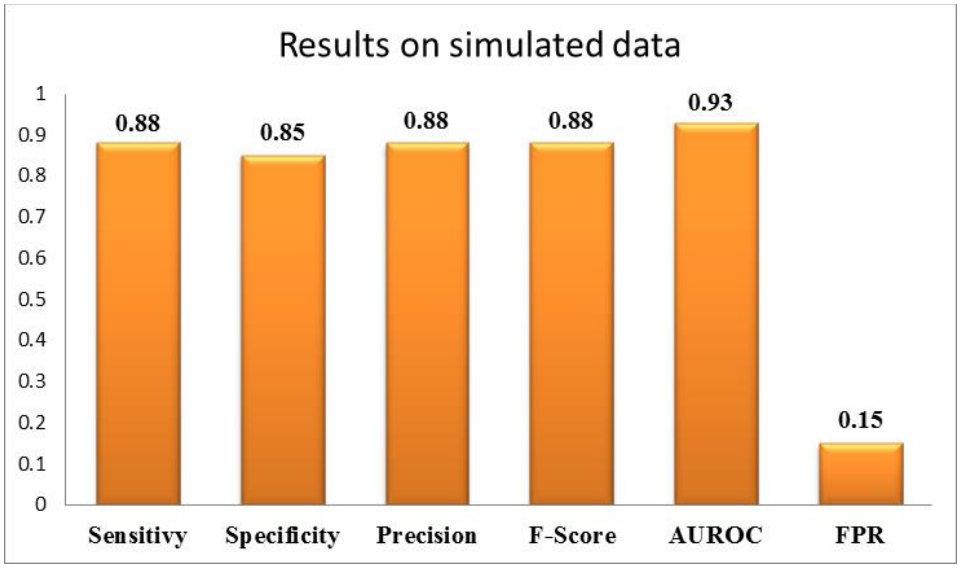


Fig. (6). Results on the simulated dataset.

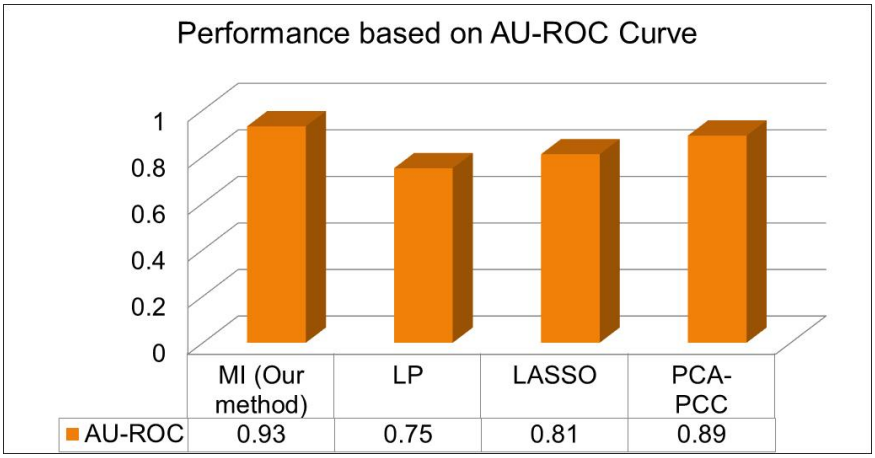


Fig. (7). Performance comparison of MI-based approach with other techniques based on AU-ROC measure.

The accuracy of the method in terms of sensitivity, specificity, precision, F-Score, AUROC and FPR is shown in Fig. (**6**). The sensitivity, precision, F-Score is equal to 0.88; specificity is 0.85 and AUROC is 0.93. The FPR is quite reasonable, i.e., 0.15. The performance of the method has been compared with other techniques such as linear programming (LP) [35], least absolute shrinkage and selection operator (LASSO) [36] and PCA-PCC [37]. The MI-based result has been found to be performing better than the rest, as shown in Fig. (**7**).

CONCLUSION

In this work, we have shown the application of information theoretic approach to colon cancer, demonstrating how this approach can reveal novel gene regulatory interactions in case of cancer. We constructed ten different networks by varying the number of interactions, ranging from 30 to 500. The identified signature in first network captures the regulatory relationships among 22 differentially expressed genes. In case of tenth network considering 500 interactions, it shows regulatory relationships among 79 differentially expressed genes. Our study has resulted three major outcomes. First, we identified differentially expressed genes in colon cancer patients, most of which were biologically verified and found to participate in colon cancer. Second, the interactions between differentially expressed genes have been identified, which need further biological validation for its reliability. Third, we identified genes regulating most of the other genes (hubs). The utility of our approach and the reliability of the obtained results need further experimental validation. These findings may help to reveal the common interaction mechanism of colon cancer and provide new insights into cancer diagnosis and therapy. Topological analysis of the inferred gene regulatory network modules has been done. We observed that inferred networks have a few number of highly connected genes (called hubs), and a majority of the genes were found to be sparsely connected. The high clustering coefficient shows that the networks have modular architecture. The shortest path length shows that if we randomly select two nodes, then there is a higher probability of them being connected. The shortest path also shows the robustness of all the analyzed networks. The increasing trend in betweeness centrality of colon cancer networks indicate that genes are functionally capable of holding together interacting genes. The path length distribution shows that the network follows 'small world' property. The degree distributions of all the analyzed networks follow power-law; hence these networks are scale-free. We also performed GO-based enrichment analysis of inferred network modules. Finally, the proposed method has also been tested on simulated network of yeast dataset taken from DREAM3 challenge, and AUROC has been found to be approximately 0.93, which is better than the other techniques such as LP, LASSO and PCA-PCC.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Wang X, Gotoh O. Microarray-based cancer prediction using soft computing approach. Cancer Inform 2009; 7: 123-39.

[2] Maetschke SR, Madhamshettiwar PB, Davis MJ, Ragan MA. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. Brief Bioinform 2014; 15(2): 195-211.

[3] Pavlopoulos GA, Secrier M, Moschopoulos CN, *et al*. Using graph theory to analyze biological networks. BioData Min 2011; 4(1): 10.

[4] Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011; 12(1): 56-68.

[5] Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Med 2012; 4(5): 1-16.

[6] Bonnet E, Tatari M, Joshi A, *et al.* Module network fnference from a cancer gene expression data set identifies microRNA regulated modules. PLoS ONE 2010; 5(4): e10162.

[7] Ahmad FK, Deris S, Othman NH. The inference of breast cancer metastasis through gene regulatory networks. J Biomed Inform 2012; 45(2): 350-62.

[8] Yang B, Zhang J, Yin Y, Zhang Y. Network-based inference framework for identifying cancer genes from gene expression data. Biomed Res Int 2013; 2013: 401649.

[9] Raza K, Jaiswal R. Reconstruction and analysis of cancer-specific gene regulatory networks from gene expression profiles. Intl J Bioinform Biosci 2013; 3(2): 25-34.

[10] Raza K, Parveen R. Reconstruction of gene regulatory network of colon cancer using information theoretic approach. Proceeding of 4th International Conference. Confluence 2013; 461-6.

[11] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270(5235): 467-70.

[12] Devore JL, Peck R. Statistics: The exploration and analysis of data. 3rd ed. Duxbury Press 1997.

[13] Draghici S. Data analysis for DNA microarrays. Chapman & Hall/CRC 2003.

[14] Kerr M, Martin M, Churchill G. Analysis of variance for gene expression microarray data. J Comput Biol 2000; 7: 819-37.

[15] Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 2011; 573(1-3): 83-92.

[16] Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2011; 98: 5116-21.

[17] Wright G, Simon R. A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics 2003; 19: 2448-55.

[18] Smyth G. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004; 3(1): Article No. 3.

[19] Pan W. A comparative review of statistical methods for discovery differentially expressed genes in replicated microarray experiments. Bioinformatics 2002; 18(4): 546-54.

[20] Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics 2006; 7(359): 1471-2105.

[21] Butte A, Kohane I. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000; 418-29.

[22] Simoes RM, Emmert-Streib F. Influence of statistical estimators of mutual information and data heterogeneity on the inference of gene regulatory networks. PLoS ONE 2011; 6(1): e29279.

[23] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU. Complex networks: structure and dynamics. Phys Rep 2006; 424(4): 175-308.

[24] Zhang J, Shakhnovich EI. Sensitivity-dependent model of protein-protein interaction networks. Phys Biol 2008; 5(3): 036011.

[25] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature 1998; 393(6684): 440-2.

[26] Rosen KH. Discrete mathematics and its applications, 7th ed. McGraw-Hill 2011.

[27] Albert R, Jeong H, Barabasi, AL. Error and attack tolerance of complex networks. Nature 2000; 406(6794): 378-82.

[28] Dong J, Horvath S. Understanding network concepts in modules. BMC Syst Biol 2007; 24.

[29] Milgram S. The small world problem. Psychol Today 1967; 2(1): 60-7.

[30] Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004; 5(2): 101-13.

[31] Maere S, Heymans K, Kuiper M. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 2005; 21(16): 3448-9.

[32] Collado M, Garcia V, Garcia JM, *et al.* Genomic profiling of circulating plasma RNA for the analysis of cancer. Clin Chem 2007; 53(10): 1860-3

[33] Assenov Y, Ramírez F, Schelhorn SE, Lengauer T, Albrecht M. Computing topological parameters of biological networks. Bioinformatics 2008; 24(2): 282-4.

[34] Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics 2011; 27(16): 2263-70.

[35] Wang Y, Joshi T, Zhang XS, Xu D, Chen L. Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics 2006; 22: 2413-20.

[36] Tibshirani R. Regression shrinkage and selection *via* the Lasso. J R Stat Soc 1996; 58: 267-88.

[37] Kalisch M, Bühlmann P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J Mach Learn Res 2007; 8: 613-36.

Received: September 18, 2014 Revised: January 18, 2015 Accepted: January 21, 2015