Gene Regulatory Network Inference Using Maximal Information Coefficient

M. A. H. Akhand[1*], R. N. Nandi[1], S. M. Amran[1], K. Murase[2]

[1] Dept. of Computer Science and Engineering, Khulna University of Engineering and Technology, Khulna-9203, Bangladesh.

[2] Dept. of Human and Artificial Intelligent Systems, University of Fukui, 3-9-1 Bunkyo, Fukui 910-8507, Japan.

* Corresponding author. Tel.: +880-41-774318, +880-1926-203027; email: akhand@cse.kuet.ac.bd, akhandkuet@yahoo.com

Manuscript submitted July 1, 2015; accepted September 10, 2015.
doi: 10.17706/ijbbb.2015.5.5.296-310

Abstract: Gene Regulatory Network (GRN) plays an important role to understand the interactions and dependencies of genes in different conditions from gene expression data. An information theoretic GRN method first computes dependency matrix from the given gene expression dataset using an entropy estimator and then infer network using individual inference method. A number of prominent methods use Mutual Information (MI) and its variants for dependency measure because MI is an efficient approach to detect nonlinear dependencies. But MI does not work well for continuous multivariate variables. In this paper, we have investigated the recently proposed association detector method Maximal Information Coefficient (MIC), instead of MI, in inferring GRN. It is reported that MIC can detect effectively most forms of statistical dependence between pairs of variables. We have integrated MIC with two prominent MI based GRN inference methods Minimal Redundancy Network and Context Likelihood of Relatedness. The experimental studies on DREAM3 Yeast data, SynTReN generated synthetic data and SOS E. Coli real gene expression data revealed that inferred network with MIC based proposed methods outperformed their counter MI based standard methods in most of the cases, especially for large sized problem.

Key words: Gene regulatory network, mutual information, maximal information coefficient, nonlinear dependence.

1. Introduction

Inferring Gene Regulatory Network (GNR) is the reverse engineering approach to uncover the dynamic and intertwined nature of gene regulation in cellular systems. Tremendous amounts of gene expression data are available now-a-days due to modern high throughput technologies that helps to explore underlying regulatory mechanism of cellular systems [1], [2]. GNR inference is still a challenging task due to combinatorial nature of the problem as well as the poor information content in the data [3] and remains an open challenge in the field of System Biology. DREAM (Dialog for Reverse Engineering Assessments and Methods), a community based effort, offers various challenges [4]-[6] to develop noble GNR inference techniques that attracts research communities to develop distinct methods using DREAM's data.

A number of approaches have been investigated to infer GRNs from gene expression data with the aim of improving the network inference accuracy and scalability [7]. Basically, the methods can be categorized into two types: model based approaches and information theoretic approaches [8]. In a model based approach

nonlinear differential equations are used to express the chemical reaction of transcription, translation and other cellular processes. Parameters involved in nonlinear differential equations represent the regulation strengths of the regulators and a method estimates the parameter values. Representative algorithms in this category include multiple linear regression [9]-[12], singular value decomposition method [13], [14], network component analysis [15], [16], linear programming [17], particle swarm optimization [18] and immune algorithm [19].

In the information theoretic approach, the network is inferred through measuring the dependences or causalities between transcription factors and target genes [17]. A number of prominent methods in this category use Mutual Information (MI) and its variants because MI is an efficient approach to detect nonlinear dependencies that is the most vital thing to detect the regulatory mechanism. The popular methods based on MI are Relevance Network [20], MRNET [21], CLR [22], MRNETB [23], ARACNE [24], PCA-CMI [25], NARROMI [26], PCA-CMI and MIT Score [27] etc. Even though the MI is quite popular, it has some limitations. For example, MI evaluation usually involves the probability or density estimator which is challenging, especially for multivariate variables. The MI estimation is not also so easy when the variables are continuous; the commonly used strategy is discretize the data first and then estimate the MI from the discretized data [28]. Furthermore, MI fails to distinguish indirect regulators from direct ones and tends to overestimate the number of regulators targeting the gene [26].

In this work, we have investigated Maximal Information Coefficient (MIC) [29], the recently proposed association detector method, in inferring GRN. MIC is a measure of two-variable dependence that designed specifically for rapid exploration of many-dimensional datasets. It is reported that MIC can detect some rare associations as well as critical characteristics between data and may use as a good alternative of MI. To identify the effectiveness of MIC in GRN inference, we have incorporated it into MRNET and CLR, two popular GRN methods. The experimental studies on DREAM3 Yeast data, generated Synthetic data and Real Gene Expression data revealed that proposed MIC based methods outperformed their counter standard methods in most of the cases, especially for large sized problem.

Most recently, MIC have been incorporated with clustering strategy for GRN inference and identified effectiveness of MIC in GRN inference [30]. In the method, the genes with maximum similarity are grouped into same clusters and the interaction between two genes with different clusters is calculated using the weight of interaction between their corresponding medoids. In this study, MIC is used instead of MI for dependency matrix calculation in MRNET and CLR. The proposed method seems relatively simple and straight forward with respect to the clustering based one.

The rest of the paper is organized as follows. Section 2 first gives brief description of MI and MIC for better understanding and then explains MIC based two proposed GRN inference methods. Section 3 is for experimental studies: gives description of benchmark data and presents outcomes of the proposed method comparing with the counter standard methods on the data. At last, Section 4 gives a brief conclusion of this study with some future directions of works that open from it.

2. Maximal Information Coefficient and Its Integration to GRN

The aim of this study is to investigate MIC, instead of MI, for dependency measure in GRN inference. This section first briefly explains MI and MIC to make the paper self-contained and then presents proposed GRN inference methods incorporating MIC.

2.1. Mutual Information (MI)

MI is a measuring tool of mutual dependencies between two variables and is defined as

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right), \quad (1)$$

where X and Y are discrete variables; $p(x)$ and $p(y)$ are the marginal probabilities distribution; and $p(x,y)$ is the joint probability function of X, Y [31]. For continuous random variables, the MI is

$$I(X,Y) = \int_y \int_x p(x,y)\, log\left(\frac{p(x,y)}{p(x)p(y)}\right). \quad (2)$$

Here $p(x, y)$ is the joint probability density function of X and Y; and $p(x)$ and $p(y)$ are the marginal probability density functions of X and Y, respectively.

MI measures the shared information of these two variables and determines the contribution of knowing one of these variables reduces the uncertainty of others. If the variables are independent, there is no effect to reducing the uncertainty then $I(X, Y) = 0$; on the other hand, if there is a relation then $I(X, Y)>0$.

2.2. Maximal Information Coefficient (MIC)

MIC is the recently proposed dependency measure approach based on the idea that if a relationship exists between two variables then a grid can be drawn on the scatterplot of the variables partitioning the data to encapsulate the relationship [29]. To calculate MIC, a characteristics matrix is considered which is populated with the maximum mutual information gains for different particular sizes. The maximum of value the characteristic matrix is considered as the Maximal Information Coefficient, i.e., MIC.

If D is a set of ordered pairs x and y, the values may partitioned into grids with cells. For a grid G, $D|_G$ means the probability distribution made by the Data D of the cells of G. The maximum information gain for all the grids sized of x, y can be represented as

$$I^*(D,x,y) = max_G I(D|_G), \quad (3)$$

where $I(D|_G)$ denotes the mutual information of $D|_G$. Finally, MIC is the maximum value of the normalized form characteristic matrixes with Eq. 3 and may express as

$$MIC(D) = \max_{xy<B(|D|)} \frac{I^*(D,x,y)}{log_2 min\{x,y\}}, \quad (4)$$

where B is a growing function satisfying $B(n) = O(n)$.

MIC has two properties, Generality and Equitability. Generality means the capability of detecting different association with sufficient sample size to all functional relationships [32]. Equitability means to give similar scores for equally noisy relationships of different types. The detail description of MIC is available in [29] and [32].

It is reported that MIC is a good alternative for any other correlation measure methods like Mutual Information, Maximal Coefficient and Distance Correlation. Since MIC can detect some rare associations and critical characteristics in data and easily applicable for continuous multi variable [29], incorporation of MIC in GRN inference might improve GRN performance. The coming section will explain proposed MIC based GRN methods where MIC is used instead of MI.

2.3. Integration of MIC with MRNET and CLR

Minimal Redundancy Network (MRNET) [21] and Context Likelihood of Relatedness (CLR) [22] are two popular GRN methods based on MI. Both the methods compute dependency matrix from the given gene expression data using MI and infer network using their individual inference method. MRNET [21] infers a network by using Maximum Relevance/Minimum Redundancy (MRMR) feature selection procedure. MRMR selects a set of variables that both have high MI with the target variable (maximum relevance) and are low

MI between them (minimum redundancy). MRMR returns a score according to each target gene and MRNET considers network deleting all edges whose score lies below a given threshold. On the other hand, CLR first calculates MI of each pair of genes and then derives a score from the empirical distribution of the calculated MI for all the gene pairs. The method considers a link between genes when the calculated score exceeds a given threshold. Finally, quality of an inferred network is measured comparing with the given true network [23]. The detail descriptions of MRNET and CLR are available in [21] and [22], respectively.

To integrate MIC with MRNET and CLR, MIC is used instead of MI as an entropy estimator to get the dependency matrix. Using MIC as dependency calculation in MRNET and CLR hereafter call MRNET-MIC and CLR-MIC, respectively. Algorithm 1 and Algorithm 2 present major steps of proposed MRNET-MIC and CLR-MIC, respectively. Algorithm 1 uses MRNET for network inference (Step 3) using Dependency Matrix (*DM*) that calculated using MIC (Step 2). Standard MRNET uses MI to produce the *DM*; therefore, uses of *MI* instead of *MIC* in Step 2 will turn Algorithm 1 as standard MRNET. To validate Inferred Network (*IN*), it first generates Confusion Matrix (*CM*) comparing *IN* with given True Network (*TN*). Then, receiver operator characteristic (ROC) and precision-recall (PR) curves are drawn (Step 4.b); and ROC and PR areas are calculated (Step 4.c) from the *CM* to evaluate a method. Algorithm 2 is differ from Algorithm 1 in only network inference stage (Step 3): Algorithm 2 uses CLR for network inference, whereas Algorithm 1 uses MRNET. Similar to Algorithm 1, Algorithm 2 will be standard CLR if *MI* is used instead of *MIC* in Step 2.

Algorithm 1: MRNET-MIC Algorithm

Input: Gene Expression Dataset *D*, True Network *TN*

Output: Inferred Network, ROC curve, PR curve, ROC area, PR area

1. Load data
2. Compute Dependency Matrix, *DM = MIC(D)* // Use *MIC*
3. Generate Inferred Network, *IN = MRNET(DM)* // Use MRNET
4. Validate *IN* comparing with *TN*
   a. Generate Confusion Matrix (*CM*)
   b. Draw ROC curve and PR curve from *CM*
   c. Calculate ROC area and PR area from *CM*

Algorithm 2: CLR-MIC Algorithm

Input: Gene Expression Dataset *D*, True Network *TN*

Output: Inferred Network, ROC curve, PR curve, ROC area, PR area

1. Load data
2. Calculate Dependency Matrix, *DM = MIC(D)* // Use *MIC*
3. Generate Inferred Network, *IN = CLR(DM)* // Use *CLR*
4. Validate *IN* comparing with *TN*
   a. Generate Confusion Matrix (*CM*)
   b. Draw ROC curve and PR curve from *CM*
   c. Calculate ROC area and PR area from *CM*

3. Experimental Studies

This section first explains the benchmark data that used in this study, experimental setup and validation methods. Then, performance of MIC based proposed GRN inference methods on the benchmark data is presented comparing with their counter MI based standard methods.

3.1. Datasets

In this study, both synthetic and real gene expression benchmark data are considered. The gene expression data is available in a two dimensional matrix form in which each column represents an individual gene and each row represents the expression level of all genes within an experiment. Table 1 shows the brief description of the datasets which shows a considerable variety in the number of types, gene number, sample size; and thus provides a suitable experimental test bed. Yeast datasets (with 10, 50 and 100 genes) contain noise free synthetic data from DREAM3 challenge [4]-[6]. We employed SynTReN (Synthetic Transcriptional Regulatory Networks) [33] network generator to generate synthetic data. SynTReN is well accepted software to create synthetic transcriptional regulatory network and to generate respective simulated data from the source network with different level of noise. We have generated three datasets SynTReN1, SynTReN2 and SynTReN3 based on Escherichia Coli (i.e., E. Coli) source with biological and experimental noise levels of 0.1, 0.2 and 0.3, respectively. On the other hand, the SOS data is the well-known SOS DNA repair network dataset of real E. Coli [34]. The selected datasets are popular for GRN inference and are employed in many exiting studies as benchmark [4]-[6], [35], [36].

Table 1. Benchmark Datasets for GRN Inference

Dataset	Origin	Data Type	Genes	Samples
Yeast10	DREAM3 challenge	Synthetic noise free	10	10
Yeast50		Synthetic noise free	50	50
Yeast100		Synthetic noise free	100	100
SynTReN1	SynTReN network generator	Synthetic with noise level 0.1	200	100
SynTReN2		Synthetic with noise level 0.2	200	100
SynTReN3		Synthetic with noise level 0.3	200	100
SOS	Real E. Coli	Real Gene Expression Data	9	9

3.2. Experimental Setup and Validation Method

We followed a common general experimental setup that does not favor any particular method. We implemented the methods and simulated the results in *R* language, the well-known open source statistical analysis tool. We employed *minevra* & *minet* packages in *R* for GRN inference; the packages are freely available in the site of [37] and [38]. The performance evaluated by receiver operator characteristic (ROC) curve and precision-recall (PR) curve. In general, ROC curve is a graphical tool for depicting true hit rate along the vertical axis (the number of target events correctly classified as targets) as compared to false alarm rate along the horizontal axis (the number of target events incorrectly classified as non-targets). In GRN inference evaluation, the ROC curve is created by plotting the fraction of true positive rate (i.e., true positives out of the total actual positives) vs. false positive rate (i.e., the fraction of false positives out of the total actual negatives), at various threshold settings. The following equations are used to calculate true positive rate (*TPR*) and false positive rate (*FPR*).

$$TPR = TP/(TP + FN) \tag{5}$$

$$FPR = FP/(FP + TN) \tag{6}$$

Here TP=True Positive (i.e., links are correctly identified), FP=False Positive (i.e., identified links are not correct), TN=True Negative (i.e., correctly identified that there is no links between genes), FN=False Negative (i.e., failed to identify links between genes). TPR is also known as sensitivity or recall in machine learning. The areas under ROC curve are then calculated.

We also evaluated the methods based on Precision – Recall (PR) curve. The PR curve is recommended to be an alternative to the ROC curves [39]. Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. The equation to calculate Precision and Recall are as follows.

$$Precision = TP / (TP + FP) \quad (7)$$

$$Recall = TP / (TP + FN) \quad (8)$$

The areas under PR curve are then calculated. It is notable that higher values of ROC and PR areas indicate better proficiency of a method.

3.3. Results and Discussion

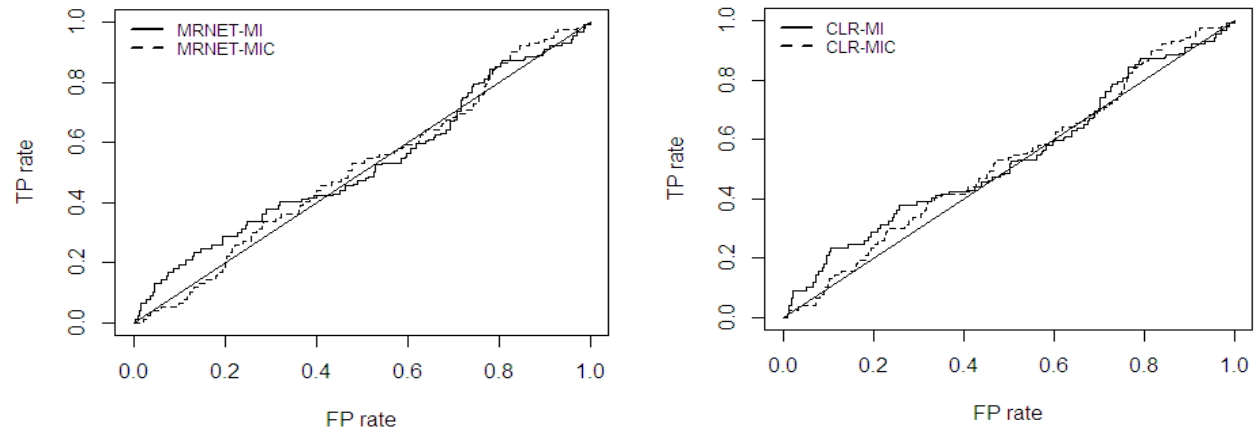
This section evaluates performance of MIC based proposed GRN inference methods (i.e., MRNET-MIC and CLR-MIC) with standard MRNET and CLR methods (may call MRNET-MI and CLR-MI) on the basis of ROC and PR curves for the datasets of Table 1. Tables 2-4 present ROC and PR areas of the methods for the datasets. On the other hand, ROC and PR curves are presented in Figs. 1-3 for three selected datasets from different types. In a table, better result between a proposed MIC based method and the corresponding MI based standard method indicated with italic type. The result with bold face type represents the best outcome among the four methods. In addition, effect of MIC is presented for better understanding as a rate of improvement with respect to the standard MI based method. A positive (+) sign indicates proposed MIC method outperformed standard MI based method; whereas negative (-) sign means MIC was not effective for the problem.

3.3.1. Evaluation on DREAM3 challenge yeast data

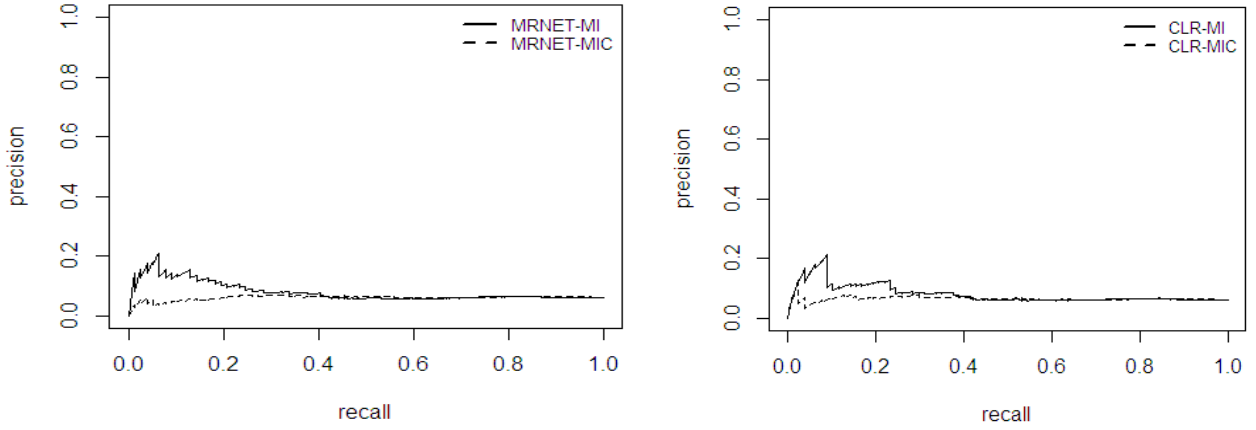
Table 2 compares ROC areas and PR areas between MIC and MI based MRNET and CLR on Yeast datasets with size 10, 50 and 100. Since ROC and PR curves are found almost similar characteristics for the methods in three Yeast datasets (i.e., Yeast10, Yeast50 and Yeast100), curves for only Yeast50 dataset are presented in Fig. 1. It is observed from the figure that a MIC based method is competitive to standard MI based method for both MRNET and CLR. Therefore, ROC area and PR area are also found competitive for both MI and MIC based methods. MIC is shown to improve CLR slightly for Yeast10 dataset only; 3.05 % and 14.19 % in ROC and PR areas, respectively. On the other hand, MRNET-MIC is found inferior to standard MRNET for all three problems. DREAM3's Yeast data is a synthetic data and contains few genes as well as few samples; competitive performance of MIC is acceptable because MIC is reported well for high dimension data.

Table 2. The ROC Areas and the PR Areas of Different Methods on DREAM3 Yeast Datasets with Size 10, 50 and 100

Method	MRNET	MRNET-MIC	Effect of MIC	CLR	CLR-MIC	Effect of MIC
ROC area						
Yeast10	***0.6444***	0.4719	- 26.77 %	0.5731	*0.5906*	+ 3.05 %
Yeast50	*0.5239*	0.5111	- 2.45 %	***0.5353***	0.5261	- 1.72 %
Yeast100	***0.5557***	0.5312	- 4.42 %	*0.5513*	0.5459	- 0.98 %
PR area						
Yeast10	***0.3220***	0.1829	- 43.18 %	0.2671	*0.3050*	+ 14.19 %
Yeast50	*0.0799*	0.0608	- 23.81 %	***0.0800***	0.0654	- 18.26 %
Yeast100	***0.0421***	0.0363	- 13.90 %	*0.0410*	0.0408	- 0.49 %



(a) ROC curve



(b) PR curve

Fig. 1. The performance of different methods on DREAM3 challenge Yeast dataset with size 50.

3.3.2. Evaluation on synthetic data from SynTReN

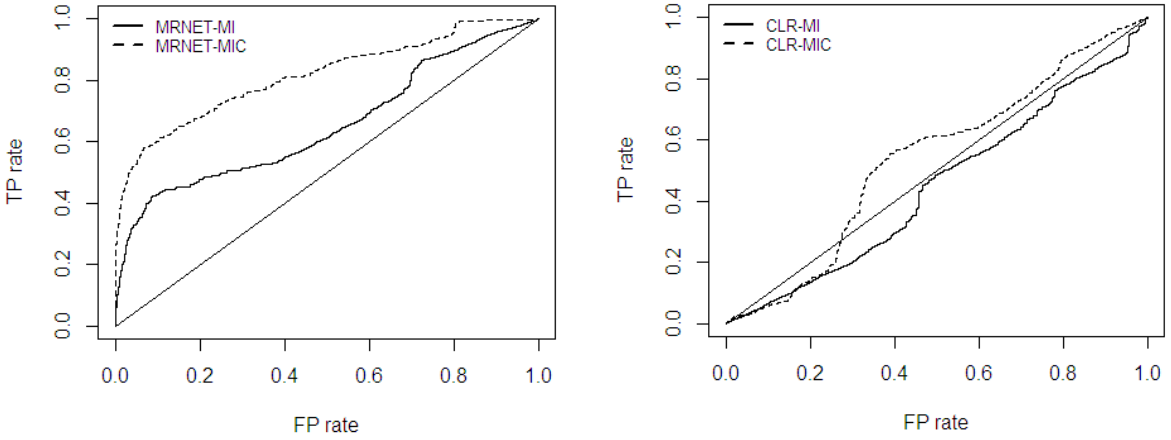
Table 3 compares ROC areas and PR areas between MIC and MI based MRNET and CLR on SynTReN generated datasets. Since ROC and PR curves are found almost similar characteristics for the methods in three synthetic datasets (i.e., SynTReN1, SynTReN2 and SynTReN3), curves for only SynTReN2 dataset are presented in Fig. 2. According to the figure, both MRNET-MIC and CLR-MIC are shown to outperform their counter standard MRNET and CLR methods, respectively and clearly identifies the effectiveness of MIC in GRN inference from the datasets. Consequently, MIC is found to improve ROC area and PR area values significantly (as seen in Table 3) for both MRNET and CLR methods in all three datasets. As an example, for SynTReN1 dataset the ROC area values for standard MRNET and CLR are 0.6515 and 0.4471, respectively. For the same dataset, the proposed MRNET-MIC and CLR-MIC are shown ROC area values of 0.8142 and 0.5750, respectively. Thus, improvements of MRNET and CLR due to MIC were 24.97% and 28.96%, respectively. Again, On the basis of PR area for SynTReN1, MIC based proposed MRNET-MIC and CLR-MIC also outperformed their counter standard MI based methods. Similar proficiency of MIC in GRN inference are also found for SynTReN2 and SynTReN3 datasets. Among the four methods, proposed MRNET-MIC is found as the best GRN inference method for all three SynTReN generated datasets. It is notable that the generated datasets are based on E. Coli source and gene size is 200. Therefore, for such large number of genes, performance improvement with MIC is logical and justify the proficiency of MIC. Moreover, the performance improvement for the synthetic data is more acceptable to validate a method because in this

case true network is defined whereas network structure is unknown for most of the real phenomena [40]. Finally, the performance of proposed MRNET-MIC on SynTReN generated data clearly revealed the effectiveness of MIC in network inference.

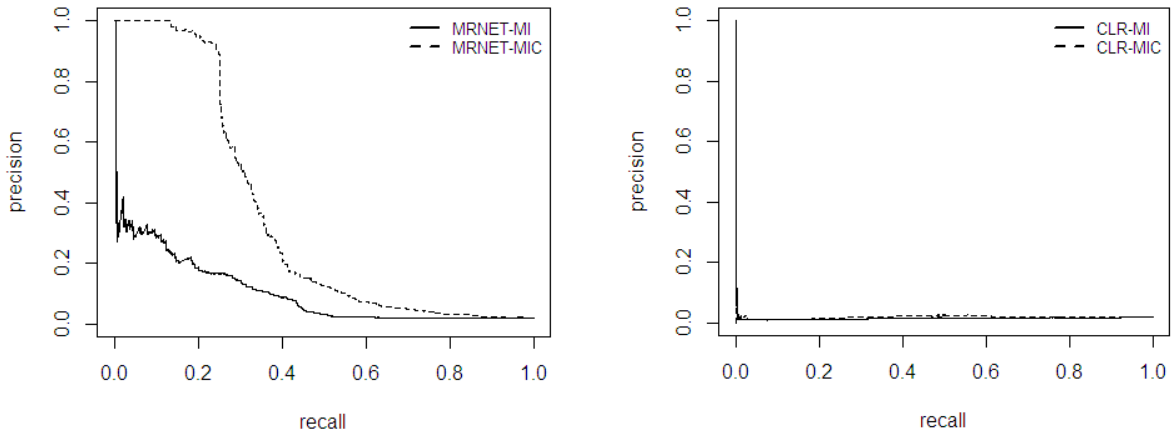
Table 3. The ROC Areas and the PR Areas of Different Methods on SynTReN Generated Datasets with Size 200 for Noise Levels 0.1, 0.2 and 0.3.

Method	MRNET	MRNET-MIC	Effect of MIC	CLR	CLR-MIC	Effect of MIC
ROC area						
SynTReN1	0.6515	***0.8142***	+ 24.97 %	0.4471	*0.5750*	+ 28.60 %
SynTReN2	0.6478	***0.7958***	+ 22.85 %	0.4471	*0.5373*	+ 20.16 %
SynTReN3	0.6467	***0.7839***	+ 21.21 %	0.4471	*0.5556*	+ 24.26 %
PR area						
SynTReN1	0.1045	***0.3772***	+ 261.07 %	0.0159	*0.0242*	+ 52.35 %
SynTReN2	0.0998	***0.3507***	+ 251.49 %	0.0159	*0.0197*	+ 23.83 %
SynTReN3	0.1054	***0.3556***	+ 237.24 %	0.0159	*0.0218*	+ 37.24 %

3.3.3. Evaluation on SOS E. coli real gene expression data



(a) ROC curve

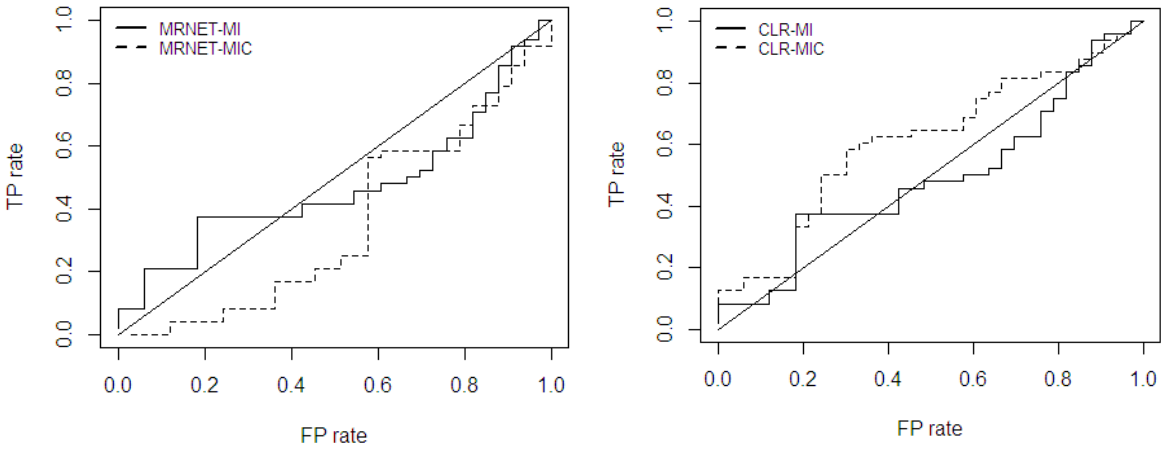


(b) PR curve

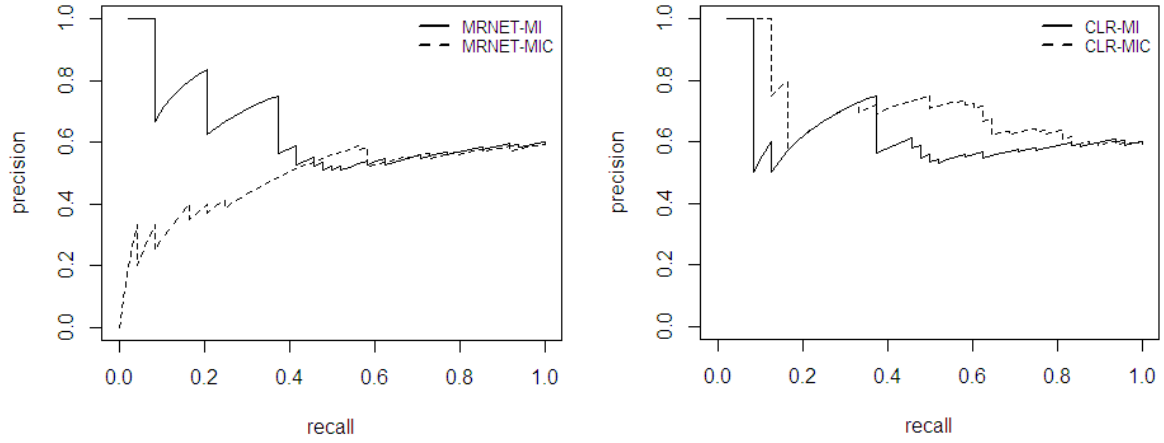
Fig. 2. The performance of different methods on SynTReN generated synthetic SynTReN2 dataset with size 200.

It is interesting to observe the effectiveness of the proposed MIC based methods on the real gene expression data. The SOS data is the well-known real SOS DNA repair network dataset in E. Coli [34] and is

used in many GRN inference studies up to now [35], [36]. Fig. 3 presents ROC and PR curves of the proposed MRNET-MIC and CLR-MIC methods in comparison with the standard MRNET and CLR with MI, respectively, on SOS dataset. According to the figure, at lower FP rate values MRNET-MIC is inferior to standard MRNET but competitive at the higher values. Similar scenario is also found in PR curve. On the other hand, proposed CLR-MIC is shown to outperform its counter standard CLR in most of the cases. Table 4 compares the ROC areas and the PR areas for both proposed and standard methods for the SOS dataset. Although incorporation of MIC in MRNET did not improve its performance, MIC was found to improve CLR performance at significant level. As an example, the ROC and PR area values for standard CLR are 0.4937 and 0.6089, respectively. On the other hand, using MIC the proposed CLR-MIC achieved ROC and PR area values 0.6086 (i.e., improved 23.27%) and 0.6877 (i.e., improved 12.94%), respectively. Among the four methods, the proposed CLR-MIC is significantly better than any other methods on the basis of ROC and PR areas. SOS data is very small in size heaving only nine samples for nine genes. Therefore, ineffectiveness of MIC in MRNET is logical. At the same time, achievement of significantly better outcome with proposed CLR-MIC is interesting. It also indicates the chance of getting better result with MIC for real gene expression data regardless the size.



(a) ROC Curve

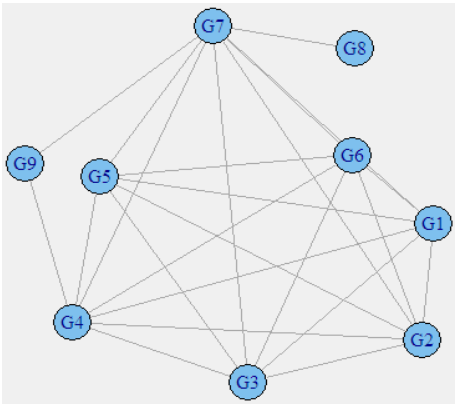


(b) PR Curve

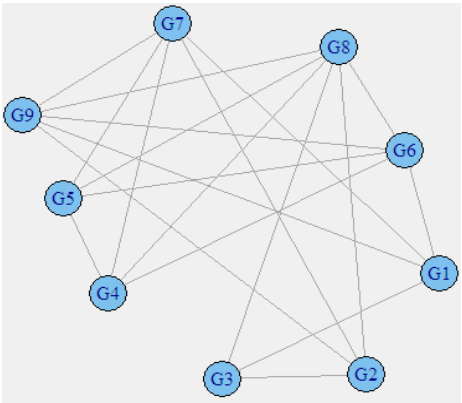
Fig. 3. The performance of different methods on SOS real gene expression data of E. Coli.

Fig. 4 shows the pictorial representation of inferred networks through different methods as well as the true network assumed behind the SOS data. In the figure, connection between two genes indicates relation in their activities. C3Net [41] is used to plot a network from regulatory values of its inferred network in which non-zero values are considered as connection. Fig. 4 identifies difference in network inference due to

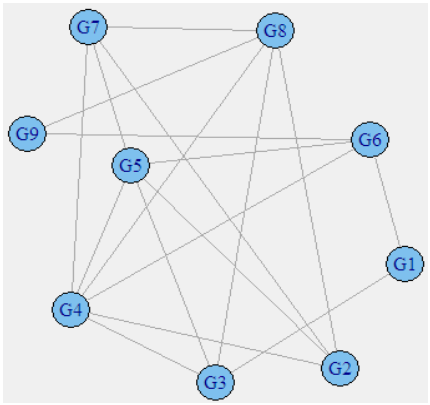
the use of MIC and MI as dependency matrix calculation. Although the use of MIC in MRNET did not improve its performance as of ROC and PR areas, some modification in inferred network are found interesting. As an example, there is a link between G7 (i.e., Gene 7) and G8 (i.e., Gene 8) in the true network (Fig. 4(a)) and the proposed MRNET-MIC identified it truly (Fig. 4(b)) that was missed by standard MRNET (Fig. 4(b)). On the other hand, the false connection between G6 and G8 in network of MRNET was rectified in MRNET-MIC. MIC employment in CLR is also found affirmative on the basis of several gene to gene connections. CLR-MIC (Fig. 4(e)) identified true connection between G3 and G4 that was missed in standard CLR (Fig. 4(d)). In the true network, G8 only connected with G7 but CLR inferred five connections with G8. The proposed CLR-MIC reduced the number into three.



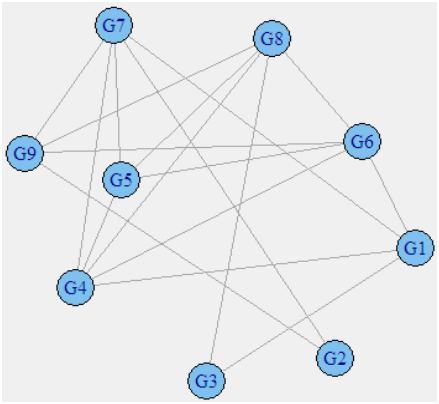
(a) True network



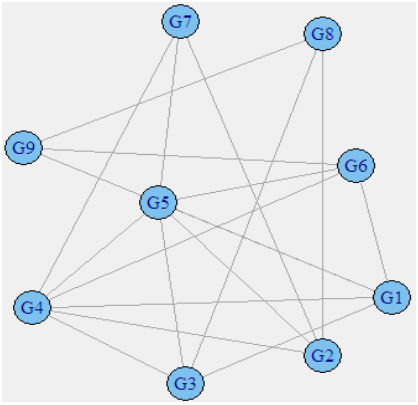
(b) Inferred network with standard MRNET



(c) Inferred network with MRNET-MIC



(d) Inferred network with standard CLR



(e) Inferred network with CLR-MIC

Fig. 4. Pictorial representation of inferred networks through different methods as well as the true network for E. Coli SOS data with 9 genes.

Table 4. The ROC Areas and the PR Areas of Different Methods on SOS Dataset of E. Coli.

Method	MRNET	MRNET-MIC	Effect of MIC	CLR	CLR-MIC	Effect of MIC
ROC area						
SOS	*0.4722*	0.3510	- 25.67 %	0.4937	***0.6086***	+ 23.27 %
PR area						
SOS	*0.6224*	0.4809	- 22.74 %	0.6089	***0.6877***	+ 12.94 %

Table 5. Summary of Inferred Connections through MRNET, MRNET-MIC, CLR and CLR-MIC Comparing with the True Network for SOS E. Coli

Connection Status	True Network	MRNET	MRNET-MIC	CLR	CLR- MIC
True Positive (TP)	24	11	13	11	14
True Negative (TN)	12	3	7	5	7
False Positive (FP)	-	9	5	7	5
False Negative (FN)	-	13	11	13	10

Table 5 compares summary of connections of inferred networks presented in Fig. 4 for better evaluation of the networks comparing with the true network of SOS data. There are 24 connections in the true network from 36 possible connections for nine genes. Total connections in the inferred network by MRNET is 20; but the connections matched with the true network (i.e., TP) are only 11 and remaining nine connections are not available in the true network (i.e., FP). MRNET correctly identified that there is no links between genes (i.e., TN) for only three cases out of 12 cases of true network but failed to identify 13 true links (i.e., FN). With MIC, inferred network of proposed MRNET-MIC seems better than that of MRNET: true values (i.e., TP= 13 and TN= 7) are more and false values (i.e., FP= 9 and FN= 11) are less than the corresponding values for standard MRNET network. In comparison to MRNET, CLR inferred network shows same TP (i.e., 11) and FN (i.e., 13) values but is better than MRNET with more TN and less FP values. On the other hand, proposed CLR-MIC inferred network outperformed that of CLR: both true connection values (i.e., TP and TN) are more and false connection values (i.e., FP and FN) are less than the corresponding values of standard CLR inferred network. At a glance, the proposed CLR-MIC inferred network is more alike to the true network than other three methods showing the highest TP value (i.e., 14) and the lowest FN value (i.e., 10).

4. Conclusion

Dependency Matrix (DM) calculation is a common step in any information theoretic method of Gene Regulatory Network (GRN) inference. A number of prominent methods use Mutual Information (MI) technique for DM calculation. MI is popular for measuring nonlinear dependencies but the recently proposed association detector method Maximal Information Coefficient (MIC) is shown to perform better than MI in several aspects. In this study, MIC has been investigated for GRN inference and verified the effectiveness of it.

MIC integrated with two prominent information theoretic GRN methods (i.e., MRNET and CLR) and proposed MRNET-MIC and CLR-MIC. In a proposed method MIC is used for DM calculation instead of MI of its standard form. The outcomes of proposed methods were evaluated and compared with their counter standard methods for DREAM3 Yeast data, generated Synthetic data and SOS real gene expression data. The performance of the methods were measured on the basis of ROC curve, PR curve, ROC area and PR area. A MIC based proposed method was shown competitive performance with its standard MI based method for Yeast dataset that is small sized synthetic noise free data. On the other hand for E. Coli based SynTReN generated data with 200 genes, each of MRNET-MIC and CLR-MIC always outperformed its counter MI

based standard method MRNET-MI and CLR-MI, respectively. Among the four tested methods, proposed MRNET-MIC was shown significant result in GRN inference for such large datasets. More interestingly, proposed CLR-MIC was shown the best suited method for SOS real gene expression data although its size is very small. Finally, the experimental results reveal that MIC is a good choice for GRN inference.

A potential future direction is also opened from this study. In this study MIC is incorporated with two popular GRN methods; and MIC incorporation with other information theoretic methods [24]-[27], [35] may give better performance that remain as future study. Moreover, an alternate version of MIC, Generalized MIC (GMIC) [32], might also be interesting to use in GRN inference.

References

[1] Hughes, T., Marton, M., Jones, A., Roberts, C., Stoughton, R., *et al.* (2000). Functional discovery via a compendium of expression profiles. *Cell, 102*, 109–126.

[2] Sanchez-Osorio, I., Ramos, F., Mayoga, P., & Dantan, E. (2014). Fundations for modeling the dynamics of gene regulatory network: A multilevel perspective review. *Journal of Bioinformatics and Computational Biology, 12(1)*, 1340003.

[3] Someren, E. P., Wessels, L. F., Backer, E., & Reinders, M. J. (2002). Genetic network modeling. *Pharmacogenomics*, *3(4)*, 507–525.

[4] Marbach, D., Schaffter, T., Mattiussi, C., & Floreano, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology, 16(2)*, 229–239.

[5] Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., & Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences, 107(14)*, 6286-6291.

[6] Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., *et al.* (2010). Towards a rigorous assessment of systems biology models: The dream3 challenges. *PloS ONE, 5(2)*, e9202.

[7] Smet, R. D., & Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews: Microbiology, 8*, 717-729.

[8] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., & Bernardo, D. (2007). How to infer gene network from expression profiles. *Molecular Systems Biology, 3*, 78.

[9] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58(1),* 267-288.

[10] Gardner, T. S., Bernardo, D., Lorenz, D., & Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science, 301,* 102-105.

[11] Cantone, I., Marucci, L., Iorio, F., Ricci, M. A., Belcastro, V., *et al.* (2009). A yeast synthetic network for in vivo assessment of reverse engineering and modeling approaches. *Cell, 137(1)*, 172-181.

[12] Honkela, A., Girardot, C., Gustafson, E. H., Liu, Y., Furlong, E. E. M., *et al.* (2010). Model-based method for transcription factor target identification with limited data. *Proceedings of National Academy of Science, USA*, *107(17),* 7793-7798.

[13] Yeung, M. K. S., Tegne´r, J., & Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of National Academy of Science, USA, 99(9),* 6163-6168.

[14] Bernardo, D., Thompson, M. J., Gardner, T. S., Chobot, S. E., Eastwood, E. L., *et al.* (2005). Chemogenomic

profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechnology, 23(3)*, 377-383.

[15] Liao, J. C., Boscolo, R., Yang, Y., Tran, L. M., Sabatti, C., & Roychowdhuryet, V. P. (2003). Network component analysis: reconstruction of regulatory signals in biological systems, *Proceedings of National Academy of Science, USA*, *100(26)*, 15522–15527.

[16] Chen, L., Xuan, J., Riggins, R. B., Wang, Y., Hoffman, E. P., & Clarke, R. (2010). Multilevel support vector regression analysis to identify condition-specific regulatory networks. *Bioinformatics*, *26(11)*, 1416-1422.

[17] Küffner, R., Petri, T., Tavakkolkhah, P., Windhager, L., & Zimmer, R. (2012). Inferring gene regulatory networks by ANOVA. *Bioinformatics, 28(10)*, 1376-1382.

[18] Palafox, L., Noman, N., & Iba, H. (2013). Reverse engineering of gene regulatory networks using dissipative particle swarm optimization. *IEEE Trans. on Evolutionary Computation, 17(4)*, 577-586.

[19] Tomoyoshi, N. , Shigeto, S., & Takenaka, Y. (2011). Inference of S-system models of gene regulatory networks using immunue algorithm. *Journal of Bioinformatics and Computational Biology, 9*, 75–86.

[20] Butte, A. J., & Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Proceedings of Pacific Symposium on Biocomputing, 5*, 418-429.

[21] Meyer, P. E., Kontos, K., & Bontempi, G. (2007). Biological network inference using redundancy analysis, bioinformatics research and development. *Lecture Notes in Computer Science, 4414,* 16-27.

[22] Faith, J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., *et al.* (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PloS Biology, 5(1),* e8.

[23] Meyer, P. E., Marbach, D., Roy, S., & Kellis, M. (2010). Information-theoretic inference of gene networks using backward elimination. *Proceedings of International Conference on Bioinformatics and Computational Biology (BioComp)*.

[24] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., *et al.* (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinformatics, 7*, S7.

[25] Zhang, X., Zhao, X., He, K., Lu, L., Cao, Y., *et al.* (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics, 28(1)*, 98-104.

[26] Zhang, X., Liu, K., Liu, Z., Duval, B., Richer, J., *et al.* (2013). NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference, *Bioinformatics, 29(1)*, 106–113.

[27] Aghdam, R., Ganjali, M., & Eslahchi, C. (2014). IPCA-CMI: An algorithm for inferring gene regulatory networks based on a combination of PCA-CMI and MIT score. *PloS ONE, 9(4)*, e92600.

[28] Simoes, R. M., & Emmert-Streib, F. (2012). Bagging statistical network inference from large-scale gene expression data, *PloS ONE, 7(3)*, e33624.

[29] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., *et al. (2011).* Detecting novel associations in large data sets, *Science, 334*, 1518-1524.

[30] Dimitrakopoulos, G. N., Maraziotis, I. A., Sgarbas, K., & Bezerianos, A. (2014). A Clustering based Method Accelerating Gene Regulatory Network Reconstruction. *Procedia Computer Science, 29,* 1993-2002.

[31] MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, ISBN 0-521-64298-1.

[32] Luedtke, A. R., & Tran, L. (2013). *The Generalized Mean Information Coefficient*, Tech. Rep. arXiv preprint arXiv: 1308, 5712.

[33] Bulcke, T. V., Leemput, K. V., Naudts, B., Remortel, P., Ma, H., *et al.*(2006). SynTRreN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics, 7*, 43.

[34] Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation topnetwork of escherichia coli. *Nature Genetics, 31*, 64-68.

[35] Zhang, X., Zhao, J., Hao, J., Zhao, X., & Chen, L. (2015). Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Research*, *43 (5)*, e31.

[36] Aghdam, R., Ganjali, M., Zhang, X., & Eslahchi, C. (2015). CN: A Consensus Algorithm for Inferring Gene Regulatory Networks Using SORDER Algorithm and Conditional Mutual Information Test. *Molecular BioSystems, 11(3),* 942-949.

[37] Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., & Furlanello, C. (2013). Minerva and minepy: A C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics, 29(3), 407-418.

[38] Meyer, P. E., Latte, F., & Bontempi, G. (2008). Minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics, 9*, 461.

[39] Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. *Proceedings of the 23rd International Conference on Machine learning* (pp. 233-240). ACM.

[40] Kabir, M., Noman, N., & Iba, H. (2010). Reverse engineering gene regulatory network from microarray data using linear time-variant model. *BMC Bioinformatics, 11*, S56.

[41] Altay, G., & Emmert-Streib, F. (2010). Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology, 4*, 132.

M. A. H. Akhand received the B.Sc. degree in electrical and electronic engineering from Khulna University of Engineering and Technology (KUET), Bangladesh in 1999, the M.E. degree in human and artificial intelligent systems in 2006, and the doctoral degree in system design engineering in 2009 from University of Fukui, Japan. He joined as a lecturer at the Department of Computer Science and Engineering at KUET in 2001, and is now a Professor. He is also head of the Computational Intelligence Research Group of this department. He is a member of Institution of Engineers, Bangladesh (IEB) and some other profession bodies. His research interests include artificial neural networks, evolutionary computation, pattern recognition, bioinformatics, swarm intelligence and other bio-inspired computing techniques.

R. N. Nandi received the B.Sc. degree in computer science and engineering from Khulna University of Engineering & Technology in 2014. At present, he is an iOS developer in IPvision Canada Inc where he works on Social Networking Applications. He is interested in bioinformatics research especially to work with biological network inference problems. He is also interested about deep learning application in the field of computational biology and computer vision.

S. M. Amran received the B.Sc. degree in computer science and engineering from Khulna University of Engineering & Technology, Bangladesh in 2014. At present, he is working in Flyte Solutions Inc on Android development with Java EE backend. His research interest includes bioinformatics, data mining and big data analysis. Currently, he is working for gathering meaningful data from large datasets, structuring and optimization.

K. Murase is a professor at the Graduate School of Engineering, University of Fukui, Fukui, Japan, since 1999. He received ME in electrical engineering from Nagoya University in 1978, PhD in biomedical engineering from Iowa State University in 1983. He Joined as a research associate at the Department of Information Science of Toyohashi University of Technology in 1984, as an associate professor at the Department of Information Science of Fukui University in 1988, and became the professor in 1992. He is a member of The Institute of Electronics, Information and Communication Engineers (IEICE), The Japanese Society for Medical and Biological Engineering (JSMBE), The Japan Neuroscience Society (JSN), The International Neural Network Society (INNS), and The Society for Neuroscience (SFN). He serves as a board of Directors in Japan Neural Network Society (JNNS), a councilor of Physiological Society of Japan (PSJ) and a councilor of Japanese Association for the Study of Pain (JASP).