An Order Independent Algorithm for Inferring Gene Regulatory Network Using Quantile Value for Conditional Independent Tests

Mir Hadi Mahmoodi
Shahid Beheshti University
Rosa Aghdam
Shahid Beheshti University
Changiz Eslahchi (ch-eslahchi@sbu.ac.ir)
Shahid Beheshti University https://orcid.org/0000-0002-8913-3904

Research article

Keywords: Inferring Gene Regulatory Network, Bayesian Network, Path Consistency Algorithm, Order Dependent Algorithm

Posted Date: January 23rd, 2020

DOI: https://doi.org/10.21203/rs.2.21700/v1

RESEARCH

An Order Independent Algorithm for Inferring Gene Regulatory Network Using Quantile Value for Conditional Independent Tests

Sayyed Hadi Mahmoudi [1], Rosa Aghdam[1†] and Changiz Eslahchi[1,2*]

*Correspondence:
ch-eslahchi@sbu.ac.ir
[1]Faculty of Mathematical Sciences, Department of Computer Science, Shahid Beheshti University., Tehran, Iran
Full list of author information is available at the end of the article
[†]Co-corresponding author

Abstract

Background: In recent years, due to the difficulty and inefficiency of experimental methods, numerous computational methods have been introduced for inferring structure of Gene Regulatory Networks (GRNs). Bayesian network is one of the popular methods in this field, however, still has many drawbacks and there is still a great space to be improved. For example, the Path Consistency (PC)-based algorithms as Bayesian network methods are still sensitive to the ordering of nodes i.e. different node orders results in different network structures. The second is that the networks inferred by these methods are highly dependent on the threshold used for independence testing. Also, it is still a challenge to select the set of conditional genes in an optimal way, which affects the performance and computation complexity of the PC-based algorithm.

Results: We introduce a novel algorithm, namely Order Independent PC-based algorithm using Quantile value (OIPCQ), which improves the accuracy of the learning process of GRNs and solves the order dependency issue. The quantile-based thresholds are considered for different order of CMI test. For conditional gene selection, we consider the paths between genes with length equal or greater than 2 while others well-known PC-based methods only considers the paths of length 2. We applied OIPCQ on the various networks of the DREAM3 and DREAM4 in silico challenges. As a real-world case study, we used OIPCQ to reconstruct SOS DNA network obtained from *Escherichia coli* and GRN for acute myeloid leukemia based on the RNA sequencing data from The Cancer Genome Atlas. The results show that OIPCQ produces the same network structure for all the permutations of the genes and improves the resulted GRN through accurately quantifying the causal regulation strength in comparison with others well-known PC-based methods.

Conclusions: According to the GRN constructed by OIPCQ, for acute myeloid leukemia, two regulators BCLAF1 and NRSF reported by Zhang et al are significantly important. However, the highest degree nodes in this GRN are ZBTB7A and PU1 which play a significant role in cancer, and especially in leukemia. OIPCQ is freely accessible at https://github.com/haammim/OIPCQ-and-OIPCQ2.

Keywords: Inferring Gene Regulatory Network; Bayesian Network; Path Consistency Algorithm; Order Dependent Algorithm

Background

Identifying regulations between genes is an important issue for better understanding the biological processes [1–4]. It provides information on what genes of particular

interest are over-expressed or under-expressed by different environmental conditions. Experimental methods for inference of Gene Regulatory Networks (GRN) are expensive, tedious, time-consuming and sometimes not reproducible. Recently, gene expression data is accessible through high-throughput sequencing technologies, which provides an insight on the regulatory mechanism [1, 3, 5, 6]. In GRN, genes are denoted as nodes while the goal is to detect interactions between them, referred to as edges. Computational methods developed to reconstruct GRNs are generally categorized in either machine-learning-based or model-based methods [7–35]. In literature, Pearson correlation coefficients [36, 37] and information theory [15, 18–27, 29–32, 38] are widely used to measure the regulation strength between genes. Both information theory and Pearson correlation coefficient can infer large-scale networks, but Mutual Information (MI) has the capability to measure non-linear dependencies which is a suitable measure to distinguish the relation between genes [36, 39]. As one of the most highly regarded network models, the Bayesian network is a factorization of a probability distribution alongside a Directed Acyclic Graph (DAG). The methods used to reconstruct the structure of Bayesian networks normally fall into three categories including constraint–based, score-based searching and hybrid methods. The Path Consistency Algorithm (PCA) is commonly used as a constraint-based method to infer GRNs. Methods based on PCA such as Fast Causal Inference (FCI), Really Fast Causal Inference (RFCI), PC Algorithm based on Conditional Mutual Information (PCA-CMI) and their modifications [25, 40–47] have two common drawbacks. The first is that these methods are not consistent for different sequential node orders [48]. The second is that the networks inferred by these methods are highly dependent on the threshold used for independence testing. Consensus Network (CN) [15], introduced Sequential ORDERing (SORDER) algorithm to selects a suitable sequential ordering of genes. It also improves the accuracy of the obtained results by taking the consensus of different networks. In [18], Zhang et al. introduced Conditional Mutual Inclusive Information (CMI2), which improves the GRN skeleton by utilizing interventional probability and Kullback-Leibler (KL) divergence. One of the issues so far unresolved in the literature is the order-dependency restriction of the algorithms, which the current work aims to address. Also, in conditional-independent tests, the proper selection of a collection of nodes which contains the separator sets significantly influences the performance of constraint-based methods. In the proposed method, a strategy for an effective selection of nodes based on existing paths between any pair of genes is devised in order to improve the network results. Also, our method is an order independent algorithm to reconstruct GRNs from gene expression data to overcome restrictions of order-dependent algorithms. The rest of the paper is organized as follows. In section Methods, a summary of related concepts including PCA, PCA-CMI and CMI2-based Network Inference (CMI2NI) [18] is provided as well as the proposed algorithm, called OIPCQ, for inferring GRNs. In section Results, the results of OIPCQ on the Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenges and the SOS DNA network in *Escherichia coli* was compared with the results of three state-of-the-art approaches including PCA-CMI, CN and CMI2NI. Finally, a case study was provided to evaluate the performance of OIPCQ for inferring a network of Acute Myeloid Leukemia (AML). The gene expression data is available in The Cancer Genome Atlas (TCGA) website at http://cancergenome.nih.gov/.

Results

In this section, the performance of OIPCQ and OIPCQ2 are benchmarked against other well-known methods (PCA-CMI, CN and CMINI) using both simulated (DREAM project) and real data (SOS DNA and AML). The DREAM project is an in silico network challenge introduced in 2006. In this work, we used DREAM3 and DREAM4 datasets. DREAM3 contains three sub-challenges of size 10, 50 and 100 genes. Each sub-challenge contains five gold standard networks (Ecoli1, Ecoli2, Yeast1, Yeast2 and Yeast3) and for each, there are three gene expression sets (heterozygous knockdown, null-mutants (steady state) and trajectories (time courses)). Among these sets, Yeast1, which is a steady-state dataset, is used. DREAM4 contains three sub-challenges of size 10, 100 and 100-multifactorial, among which, we have used all five networks of the 100-multifactorial sub-challenge.

To benchmark the performance of OIPCQ and OIPCQ2 against well-known algorithms, True Positive (TP), False Positive (FP), True Positive Rate (TPR), Positive Predictive Value (PPV), False Positive Rate (FPR), False Discovery Rate (FDR), overall ACCuracy (ACC), F-measure and Matthews Correlation Coefficient (MCC) are calculated. They are defined as follows:

$$\begin{aligned} TPR &= \frac{TP}{TP+FN},\ FDR=\frac{FP}{FP+TP},\ FPR=\frac{FP}{FP+TN} \\ ACC &= \frac{TP+TN}{TP+FP+TN+FN}, F-measure=2\frac{PPV\times TPR}{PPV+TPR}, \\ PPV &= \frac{TP}{TP+FP},\ MCC=\frac{TP\times TN-FP\times FN}{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}. \end{aligned}$$

Results for DREAM3:

On the DREAM3 datasets, OIPCQ with two thresholds (θ_1 for MI and θ_2 for CMI) were implemented and compared with PCA-CMI, CN and CMI2NI. For OIPCQ and OIPCQ2 the same parameters (θ_1 and θ_2) are considered. For PCA-CMI and CMI2NI algorithms one parameter is considered. CN algorithm requires two thresholds, one for producing a consensus network ($\mathrm{CN}_{Consensus}$) and interval threshold for independent tests ($\mathrm{CN}_{Ind.Test}$). The selected thresholds for the mentioned methods are shown in Table 1. The benchmark results for DREAM3 are summarized in Tables 2, S1 and S2 in Supplementary file. The F-measure values for mentioned algorithms are illustrated in Figure 1. The results show that OIPCQ and OIPCQ2 consistently perform better than all other algorithms in terms of PPV, ACC, MCC and F-measure criteria. OIPCQ and OIPCQ2 algorithms were benchmarked with CMI2NI using DREAM3 datasets with sizes 10, 50 and 100. In all sizes, OIPCQ and OIPCQ2 had better performance compared to CMI2NI. For size 10, FP was improved from 1 to 0 with no change in TP. For size 50, FP was decreased from 40 to 35 and TP was increased from 39 to 40. For size 100, FP was changed from 38 to 34 and TP was improved from 64 to 71.

Results for DREAM4:

Similar to the DREAM3 case, for DREAM4, OIPCQ and OIPCQ2 with two thresholds (θ_1 for MI and θ_2 for CMI) were implemented and compared with PCA-CMI, CN and CMI2NI. The selected thresholds for the mentioned methods are shown in

Table S3 in Supplementary file. The benchmark results for DREAM4 for five networks of the 100-multifactorial sub-challenge are summarized in Tables S4 through S8 and Figure 2. Similar to DREAM3, the results show that OIPCQ and OIPCQ2 consistently perform better than all other algorithms in terms of PPV, ACC, MCC and F-measure metrics.

Case Study: SOS-DNA and AML:

The efficiency of OIPCQ and OIPCQ2 algorithms are also tested on real datasets *E. coli*(SOS-DNA) and AML. The real network for SOS-DNA consists of a network with 9 genes and 24 edges. Thresholds used for the implemented algorithms are listed in Table S9 in Supplementary file. The benchmark results on the SOS-DNA dataset are presented in Figure 3 and Table S10 in Supplementary file.

The results show that OIPCQ performs consistent with the benchmark algorithms. The AML network contains 81 genes, of which 65 are target and 16 are regulatory genes. RACER algorithm, presented in [49], was developed to infer the GRN in AML dataset and is referred to as a Golden Standard for this dataset. Zhang's CMI2NI algorithm [18] produced a network with 549 edges, of which 113 are common with RACER. In order to compare the networks constructed by OIPCQ and OIPCQ2 with CMI2NI, we have selected thresholds such that the constructed network has equal edges to the CMI2NI's network. Hence, the produced network contains 549 edges from which 114 are in common with RACER. The results of OIPCQ and OIPCQ2 on AML dataset are summarized in Table S11 in the supplementary file. Figures 4 and 5 show the GRN constructed by OIPCQ and OIPCQ2 on AML, respectively. The central nodes in these figures show the 16 regulators. In these networks, the highest degrees belonged to ZBTB7A and PU1 regulators with respective values of 53 and 47 (see Table S11). These values are significantly higher compared to that of RACER network with 12 and 9 degrees. Previous studies [50, 51] have shown that ZBTB7A functions as a transcriptional suppressor. ZBTB7A was also proven to play a critical role in AML as a transcription factor in [52]. According to [53, 54], AML is also influenced by the slow decline of the transcript factor PU1. In addition to the mentioned two regulators, BCLAF1 and NRSF are reported by Zhang et al. in [18] as significant regulators with significant role in cancer. In both of our networks, BCLAF1 and NRSF (Figures 4 and 5) had high degrees as well, which is consistent with the results of CMI2NI [18] . In order to verify these findings, the pathway enrichment was done separately for each of the four regulators ZBTB7A, PU1, BCLAF1 and NRSF, along with their target genes. The pathway enrichment was done in the cancer annotation system CaGe (http://mgrc.kribb.re.kr/cage/). The results of the pathway enrichments are presented in Tables S13 to S20, which include pathways with p-values of higher than a certain threshold. In these tables the obtained p-values related to the pathways correspond to each of the four regulators and their target genes in the resulted networks from OIPCQ and OIPCQ2 algorithms are more meaningful than those of CMI2NI. In order to compare the importance of target genes of BCLAF1 and NRSF in OIPCQ and OIPCQ2 with CMI2NI, three most significant pathways were selected and compared (Table 3). The result of Table 3 indicate that the relationship between genes in our networks are more related to LEUKEMIA cancer in comparison with CMI2NI's network. In

summary, based on the results obtained in this study, we can claim that the regulators ZBTB7A and PU1 beside to BCLAF1 and NRSF play a significant role in cancer, and especially in leukemia.

Discussion

Survival of living organisms depends on the interaction between thousands of genes. GRN are schematic representations of interactions among all gene pairs in a given cell. The functions and dynamics of various cells can be figured out through reconstructing the GRNs. Recently, PC-based methods were proposed for learning structure of GRNs. However, these methods are still sensitive to the ordering of nodes. In this study, we mentioned that PC-based methods are order-dependent algorithms and the GRN results may change if the order of the nodes change. Considering the maximum value of CMI's, for removing edges in network can be increasing both FP and FN. The quantile-based threshold assume independency between genes based on the distribution of the CMI values and yield better results with respect to consider constant value for threshold. In the present study, OIPCQ and OIPCQ2 algorithms were proposed to infer GRNs from the gene expression data.

Conclusions

The results show that OIPCQ and OIPCQ2 outperform other algorithms on simulated datasets. Also, the OIPCQ and OIPCQ2 results on the AML data shows more similarities with RACER compared to some popular inferring network methods. Based on the outcomes provided in Result section, it can be observed that OIPCQ and OIPCQ2 outperform other methods in most datasets (8 out of 10 cases) while resolving the order-dependency issue. Similar results are concluded by using OIPCQ and OIPCQ2 algorithms, therefore, applying Equation 5 instead of 4 for calculating CMI, the constructed GRN do not change the result significantly. Finally, the main advantage of the proposed method is that it is applicable to all constraint-based and in particular PC-based methods.

Methods

In this section, a theoretical background on information theory, entropy, MI and CMI is provided. Following that, the details of the proposed OIPCQ algorithm for inferring the structure of GRN are introduced.

Mutual Information and Conditional Mutual Information

Both MI and CMI are proven to be effective for inferring GRNs due to their capability to measure nonlinear dependencies between variables [55]. MI and CMI are defined as follows [56, 57]:

$$MI(X,Y) = \int_{\mathbb{R}}\int_{\mathbb{R}} p(x,y)\ \log\frac{p(x,y)}{p(x)\ p(y)}\ dx\ dy, \tag{1}$$

$$CMI(X,Y) = \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} p(x,y,z)\ \log\frac{p(x,y|z)}{p(x|z)\ p(y|z)}\ dx\ dy\ dz, \tag{2}$$

where $p(x, y)$, $p(x)$ and $p(y)$ represent the joint distribution of X and Y, marginal distribution of X, marginal distribution of Y, respectively. $p(x, y, z)$, $p(x, y|z)$, $p(x|z)$ and $p(y|z)$ indicate joint distribution of X, Y and Z, the conditional density distribution of X and Y given Z, the conditional density distribution of X given Z and the conditional density distribution of Y given Z, respectively. Under the assumption that gene expression data follows a Gaussian distribution, MI for two continuous variables X and Y can be calculated as:

$$MI(X, Y) = \frac{1}{2} log \frac{|C(X)||C(Y)|}{|C(X, Y)|}, \tag{3}$$

where $C(A)$ and $|C(A)|$ indicate the covariance matrix of vector A and determinant of $C(A)$, respectively. Similarly, $CMI(X, Y|\mathbf{Z})$ is defined as:

$$CMI(X, Y|\mathbf{Z}) = \frac{1}{2} log \frac{|C(X, \mathbf{Z})||C(Y, \mathbf{Z})|}{|C(\mathbf{Z})||C(X, Y, \mathbf{Z})|}. \tag{4}$$

Conditional Mutual Inclusive Information (CMI2)

The CMI2 uses both KL divergence and interventional and is defined as:

$$\begin{aligned} CMI2(X, Y|Z) &= \frac{DKL(P\|P_{X\to Y}) + DKL(P\|P_{Y\to X})}{2} \\ &= \sum_{x,y,z} p(x, y, z) \ln \frac{p(x, y, z)}{p(x, z) \sum_x p(y|z, x)p(x) + p(y, z) \sum_y p(x|z, y)p(y)}, \end{aligned} \tag{5}$$

where $p(x, y, z)$ is the joint probability distribution of X, Y and Z, $P_{X\to Y} = P_{X\to Y}(X, Y, Z)$ and $P_{Y\to X} = P_{Y\to X}(X, Y, Z)$ are the interventional probability distributions of X, Y and Z for removing edges $X \to Y$ and $Y \to X$, respectively. $DKL(P\|P_{X\to Y})$ and $DKL(P\|P_{Y\to X})$ are KL divergences from P to $P_{X\to Y}$, and from P to $P_{Y\to X}$, respectively.

PC Algorithm based on Conditional Mutual Information (PCA-CMI)

In PCA-CMI [25], a network is initiated with a completely undirected graph. Then, through an iterative process, the skeleton gets updated as edges are removed based on the results of the independent tests between adjacent nodes. Finally, the algorithm makes the skeleton which is fully undirected.

Let S_i be a skeleton of ith order with i starting from -1. So, S_{-1} denotes a completely undirected graph from which the algorithm starts. For two adjacent nodes X and Y in S_{i-1}, a set $V_{XY} = ADJ(X) \bigcap ADJ(Y)$ is defined where $ADJ(X)$ being a set of adjacent vertices of X in S_{i-1}. $CMI(X, Y|\mathbf{M})$ is calculated for each i-subset $\mathbf{M}$ of V_{XY}. For calculating $MI(X, Y)$ and $CMI(X, Y|\mathbf{M})$, equation 3 and 4 are used respectively. For removing the edge between two adjacent nodes X and Y in S_{i-1}, $CMI_{max}(X, Y|Z)$ as $\max_{\mathbf{M}} CMI(X, Y|\mathbf{M})$ and θ as the threshold for independent test are considered. The edges for which $CMI_{max}(X, Y|Z) < \theta$ are removed from S_{i-1}.

CMI2NI: GRN inference method based on CMI2

Given an expression dataset with n genes and m samples, CMI2NI infers its underlying GRN. In CMI2NI, after obtaining MI and CMI2 with Equations 3 and 5, the PCA-CMI algorithm was used to remove the (conditional) indirect edges from the complete graph. GRN inference is performed by removing those edges without strong causal regulations recursively until there is no change in the network topology. For more details of the CMI2NI algorithm, see [18].

Drawbacks of PC-based Algorithms

In this section, we discuss about drawbacks of PC-based algorithms. The first is that these methods are not robust for different sequential node orders. The second is that the results by these methods are highly dependent on the maximum value of CMI tests used for removing edges in each step of the algorithms. In addition, the proper selection of a collection of nodes which contains the separator sets significantly influences the performance of these methods.

PC-based Algorithms are order-dependent.

The network topology resulted from PC-based algorithms are dependent on the order of genes taken as input. In each order of PC-based algorithms, the edges are removed based on conditional mutual information tests which explained in section "PC Algorithm based on Conditional Mutual Information (PCA-CMI)". Therefore, the adjacent of the vertices are updated as a result of sequential removing edges from the network. According to this method, V_{XY} depends on the initial order of the nodes and a different nodes order may eventually result in a different final network. To illustrate this, we implemented the PCA-CMI on the DREAM3 dataset with 10, 50 and 100 genes. For each of these sets, 1000 different gene order permutations were generated and tested. The threshold was set as $\theta = 0.05$ for CMI tests. Each node in Figure 6 part (a), (b) and (c) shows the TP against FP for each randomly-generated permutation for DREAM3 dataset with 10, 50 and 100 genes, respectively. Figure 6 part (d) illustrates the F-measure values for these datasets. The study on different sequential node ordering results in different TP and FP. It is concluded that, by considering different sequential node ordering, the resulted networks are also different. This figure illustrated that larger networks are more affected by the order of the input genes and the order dependency is a minor concern in networks with fewer variables (networks with less that 10 genes).

Construction of Separator Sets

In PCA-CMI, CMI2NI, and CN algorithms, the separator set is extracted from V_{XY}. So, these algorithms in each order only considers the paths of length 2 and ignores any existing connections with length greater than 2. One way of dealing with this constraint is to use $U_{XY} = ADJ(X) \bigcup ADJ(Y)$ for order greater than one ($i > 1$). For $i > 1$, by using U_{XY} instead of V_{XY}, the decision will be made by more information considering all the paths between X and Y. For example, in Figure 7 by using $V_{XY} = \{M, N\}$ only the $CMI(X, Y|M, N)$ for order $i = 2$ is calculated and only two paths of length 2, $X - N - Y$ and $X - M - Y$, between X and Y are considered. By using $U_{XY} = \{M, N, Z, W\}$, we also considered the path

of length 3, $X - Z - W - Y$, for checking the dependency between X and Y. The results show that, by considering more paths, OIPCQ helps to keep more reliable edges compared to other methods.

Removing Edges in PC-based Algorithms is threshold-dependent.
In PC-based algorithms, the decision for removing edges from a network strictly depends on the value of $CMI_{max}(X, Y|Z)$ and threshold θ as the criterion for removing the edges. In fact, the edge XY is removed if for each Z in separator X and Y, $CMI(X, Y|Z) < \theta$. This method results in many FN. On the other hand, if the most $CMI(X, Y|Z)$ are close to zero and only one of them is greater than θ, PCA-CMI, CN and CMI2NI keep the edge XY in the network. Our investigation show that most of such edges are FP. So, considering the distribution of $CMI(X, Y|Z)$, quantile-based criterion for removing an edge is more effective and yield better results than using constant value as a threshold for removing edge.

The OIPCQ Algorithm

The OIPCQ starts from a complete graph and iterates the following process to extract skeleton S_i from S_{i-1}.
Step 0: **Initialization**: Generate a complete network with number of nodes equal to the number of genes.
Step 1: **Calculate MI**: Compute MI values for each pair of genes.
Step 2: **Remove Edges**: Eliminate corresponding edges for which MI values are smaller than θ_1 (θ_1 denotes the threshold for MI test). The resulted network in this step is denoted by S_0.
Step 3: **Calculate CMI of order 1**: If there exists any edges between X and Y in S_0, find all genes Z which are adjacent to both X and Y, and then calculate their $CMI(X, Y|Z)$ for Z belongs to V_{XY}. In this step, the paths of length 2 are considered between X and Y.
Step 4: **Remove Edges**: Define $CMI_{70}(X, Y|Z)$ as all 70th percentile of all $CMI(X, Y|Z)$ values. If $CMI_{70}(X, Y|Z)$ is less than θ_2 (θ_2 is the threshold for CMI test), remove the edge between X and Y. The resulted network in this step is denoted by S_1.
Step 5: **Calculate CMI of order 2**: Do the steps above for X and Y and calculate $CMI(X, Y|Z, W)$, where Z and W belong to U_{XY}. In this step, the paths of length at least 2 are considered between X and Y.
Step 6: **Remove Edges**: Define $CMI_{70}(X, Y|Z, W)$ as the 70th percentile of all $CMI(X, Y|Z, W)$ values. If $CMI_{70}(X, Y|Z, W)$ is less than θ_2, remove the edge between X and Y.
An example of the OIPCQ algorithm is illustrated in Figure 8 for network with 5 genes. The other version of OIPCQ, named OIPCQ2 was introduced using CMI2 for detecting dependency between genes. In both OIPCQ and OIPCQ2 algorithms, in each order of algorithms for each X and Y, U_{XY} and V_{XY} are defined and fixed. Then, at the end of each order the algorithm decides to remove edges based on threshold θ_1 and θ_2. By using this method and fixing the U_{XY} and V_{XY} in each order of algorithms, the order dependency issue is solved and both OIPCQ and OIPCQ2 algorithms are order independent.

Additional Files

Additional file 1 — The Supplementary file includes Tables.

Abbreviations

Gene Regulatory Networks (GRNs); Path Consistency (PC); Order Independent PC-based algorithm using Quantile value (OIPCQ); Mutual Information (MI); Directed Acyclic Graph (DAG); Path Consistency Algorithm (PCA); Fast Causal Inference (FCI); Really Fast Causal Inference (RFCI), PC Algorithm based on Conditional Mutual Information (PCA-CMI); Sequential ORDERing (SORDER); Consensus Network (CN); Mutual Inclusive Information (CMI2); Kullback-Leibler (KL); CMI2-based Network Inference (CMI2NI); Dialogue for Reverse Engineering Assessments and Methods (DREAM); Acute Myeloid Leukemia (AML); The Cancer Genome Atlas (TCGA); True Positive (TP); False Positive (FP); True Positive Rate (TPR); Positive Predictive Value (PPV); False Positive Rate (FPR); False Discovery Rate (FDR); overall ACCuracy (ACC); Matthews Correlation Coefficient (MCC)

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Datasets and the codes of the OIPCQ and OIPCQ2 algorithms are available in our github repository, [https://github.com/haammim/OIPCQ-and-OIPCQ2.]

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this study.

Author's contributions

RA, SHM and CE developed the method and performed computational and statistical analysis. RA, SHM and CE design and wrote the paper. RA and CE contributed in writing and editing the manuscript. All authors read and approved the final manuscript

Acknowledgements

Rosa Aghdam is a postdoctoral fellow at Faculty of Mathematical Sciences, Department of Computer Science, the Shahid Beheshti University under support of Iranian National Science Foundation (INSF97009558).

Author details

[1]Faculty of Mathematical Sciences, Department of Computer Science, Shahid Beheshti University., Tehran, Iran.
[2]School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.

References

1. Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Floreano, D., Stolovitzky, G.: Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the national academy of sciences **107**(14), 6286–6291 (2010)
2. Bansal, M., Belcastro, V., Ambesi-Impiombato, A., Di Bernardo, D.: How to infer gene networks from expression profiles. Molecular systems biology **3**(1), 78 (2007)
3. De Smet, R., Marchal, K.: Advantages and limitations of current network inference methods. Nature Reviews Microbiology **8**(10), 717 (2010)
4. Chai, L.E., Loh, S.K., Low, S.T., Mohamad, M.S., Deris, S., Zakaria, Z.: A review on the computational approaches for gene regulatory network construction. Computers in biology and medicine **48**, 55–65 (2014)
5. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., *et al.*: Functional discovery via a compendium of expression profiles. Cell **102**(1), 109–126 (2000)
6. Hurley, D., Araki, H., Tamada, Y., Dunmore, B., Sanders, D., Humphreys, S., Affara, M., Imoto, S., Yasuda, K., Tomiyasu, Y., *et al.*: Gene network inference and visualization tools for biologists: application to new human transcriptome datasets. Nucleic acids research **40**(6), 2377–2398 (2011)
7. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. Journal of computational biology **7**(3-4), 601–620 (2000)
8. Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Aderhold, A., Bonneau, R., Chen, Y., *et al.*: Wisdom of crowds for robust gene network inference. Nature methods **9**(8), 796 (2012)
9. Zhang, X., Liu, K., Liu, Z.-P., Duval, B., Richer, J.-M., Zhao, X.-M., Hao, J.-K., Chen, L.: Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. Bioinformatics **29**(1), 106–113 (2012)
10. Zhu, H., Rao, R.S.P., Zeng, T., Chen, L.: Reconstructing dynamic gene regulatory networks from sample-based transcriptional data. Nucleic acids research **40**(21), 10657–10667 (2012)
11. Akutsu, T., Miyano, S., Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In: Biocomputing'99, pp. 17–28. World Scientific, ??? (1999)
12. Zhu, D., Li, H.: Improved bayesian network inference using relaxed gene ordering. International journal of data mining and bioinformatics **4**(1), 44–59 (2010)

13. Krämer, N., Schäfer, J., Boulesteix, A.-L.: Regularized estimation of large-scale gene association networks using graphical gaussian models. BMC bioinformatics **10**(1), 384 (2009)
14. Menéndez, P., Kourmpetis, Y.A., ter Braak, C.J., van Eeuwijk, F.A.: Gene regulatory networks from multifactorial perturbations using graphical lasso: application to the dream4 challenge. PloS one **5**(12), 14147 (2010)
15. Aghdam, R., Ganjali, M., Zhang, X., Eslahchi, C.: Cn: a consensus algorithm for inferring gene regulatory networks using the sorder algorithm and conditional mutual information test. Molecular BioSystems **11**(3), 942–949 (2015)
16. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. Bioinformatics **19**(17), 2271–2282 (2003)
17. Nair, A., Chetty, M., Wangikar, P.P.: Improving gene regulatory network inference using network topology information. Molecular BioSystems **11**(9), 2449–2463 (2015)
18. Zhang, X., Zhao, J., Hao, J.-K., Zhao, X.-M., Chen, L.: Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. Nucleic acids research **43**(5), 31–31 (2014)
19. Chaitankar, V., Ghosh, P., Perkins, E.J., Gong, P., Zhang, C.: Time lagged information theoretic approaches to the reverse engineering of gene regulatory networks. In: BMC Bioinformatics, vol. 11, p. 19 (2010). BioMed Central
20. Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., Gardner, T.S.: Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biology **5**(1), 8 (2007)
21. Liang, S., Fuhrman, S., Somogyi, R.: Reveal, a general reverse engineering algorithm for inference of genetic network architectures (1998)
22. Luo, W., Hankenson, K.D., Woolf, P.J.: Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. BMC bioinformatics **9**(1), 467 (2008)
23. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A.: Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In: BMC Bioinformatics, vol. 7, p. 7 (2006). BioMed Central
24. Wang, K., Saito, M., Bisikirska, B.C., Alvarez, M.J., Lim, W.K., Rajbhandari, P., Shen, Q., Nemenman, I., Basso, K., Margolin, A.A., *et al.*: Genome-wide identification of post-translational modulators of transcription factor activity in human b cells. Nature biotechnology **27**(9), 829 (2009)
25. Zhang, X., Zhao, X.-M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J.-K., Liu, Z.-P., Chen, L.: Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. Bioinformatics **28**(1), 98–104 (2011)
26. Zhou, C., Zhang, S.-W., Liu, F.: An ensemble method for reconstructing gene regulatory network with jackknife resampling and arithmetic mean fusion. International journal of data mining and bioinformatics **12**(3), 328–342 (2015)
27. Villaverde, A., Ross, J., Banga, J.: Reverse engineering cellular networks with information theoretic methods. Cells **2**(2), 306–329 (2013)
28. López-Kleine, L., Leal, L., López, C.: Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data. Briefings in functional genomics **12**(5), 457–467 (2013)
29. Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A.: Reverse engineering of regulatory networks in human b cells. Nature genetics **37**(4), 382 (2005)
30. Modi, S.R., Camacho, D.M., Kohanski, M.A., Walker, G.C., Collins, J.J.: Functional characterization of bacterial srnas using a network biology approach. Proceedings of the National Academy of Sciences **108**(37), 15522–15527 (2011)
31. Wang, J., Chen, B., Wang, Y., Wang, N., Garbey, M., Tran-Son-Tay, R., Berceli, S.A., Wu, R.: Reconstructing regulatory networks from the dynamic plasticity of gene expression by mutual information. Nucleic acids research **41**(8), 97–97 (2013)
32. Sumazin, P., Yang, X., Chiu, H.-S., Chung, W.-J., Iyer, A., Llobet-Navas, D., Rajbhandari, P., Bansal, M., Guarnieri, P., Silva, J., *et al.*: An extensive microrna-mediated network of rna-rna interactions regulates established oncogenic pathways in glioblastoma. Cell **147**(2), 370–381 (2011)
33. Li, Z., Li, P., Krishnan, A., Liu, J.: Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis. Bioinformatics **27**(19), 2686–2691 (2011)
34. Barabási, A.-L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature reviews genetics **12**(1), 56 (2011)
35. Rezaei Tabar, V., Zareifard, H., Salimi, S., Plewczynski, D.: Learning directed acyclic graphs by determination of candidate causes for discrete variables. Journal of Statistical Computation and Simulation **89**(10), 1957–1970 (2019)
36. De La Fuente, A., Bing, N., Hoeschele, I., Mendes, P.: Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics **20**(18), 3565–3574 (2004)
37. Saito, S., Hirokawa, T., Horimoto, K.: Discovery of chemical compound groups with common structures by a network analysis approach (affinity prediction method). Journal of chemical information and modeling **51**(1), 61–68 (2010)
38. Aghdam, R., Ganjali, M., Eslahchi, C.: Ipca-cmi: an algorithm for inferring gene regulatory networks based on a combination of pca-cmi and mit score. PloS one **9**(4), 92600 (2014)
39. Meyer, P.E., Lafitte, F., Bontempi, G.: minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information. BMC bioinformatics **9**(1), 461 (2008)
40. Spirtes, P., Meek, C., Richardson, T.: Causal inference in the presence of latent variables and selection bias. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 499–506 (1995). Morgan Kaufmann Publishers Inc.
41. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, Prediction, and Search vol. 81. MIT press, ??? (2000)
42. Spirtes, P.: An anytime algorithm for causal inference. In: Proc. of the Eighth International Workshop on

Artificial Intelligence and Statistics, pp. 213–221 (2001). Citeseer
43. Zhang, J.: On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence **172**(16), 1873–1896 (2008)
44. Colombo, D., Maathuis, M.H., Kalisch, M., Richardson, T.S., *et al.*: Learning high-dimensional directed acyclic graphs with latent and selection variables. The Annals of Statistics **40**(1), 294–321 (2012)
45. Claassen, T., Mooij, J., Heskes, T.: Learning sparse causal models is not np-hard. arXiv preprint arXiv:1309.6824 (2013)
46. Aghdam, R., Ganjali, M., Niloofar, P., Eslahchi, C.: Inferring gene regulatory networks by an order independent algorithm using incomplete data sets. Journal of Applied Statistics **43**(5), 893–913 (2016)
47. Aghdam, R., Alijanpour, M., Azadi, M., Ebrahimi, A., Eslahchi, C., Rezvan, A.: Inferring gene regulatory networks by pca-cmi using hill climbing algorithm based on mit score and sorder method. International Journal of Biomathematics **9**(03), 1650040 (2016)
48. Aghdam, R., Rezaei Tabar, V., Pezeshk, H.: Some node ordering methods for the k2 algorithm. Computational Intelligence **35**(1), 42–58 (2019)
49. Li, Y., Liang, M., Zhang, Z.: Regression analysis of combined gene expression regulation in acute myeloid leukemia. PLoS computational biology **10**(10), 1003908 (2014)
50. Liu, X.-S., Haines, J.E., Mehanna, E.K., Genet, M.D., Ben-Sahra, I., Asara, J.M., Manning, B.D., Yuan, Z.-M.: Zbtb7a acts as a tumor suppressor through the transcriptional repression of glycolysis. Genes & development **28**(17), 1917–1928 (2014)
51. Liu, X., Liu, Z., Gerarduzzi, C., Choi, D., Ganapathy, S., Pandolfi, P., Yuan, Z.: Somatic human zbtb7a zinc finger mutations promote cancer progression. Oncogene **35**(23), 3071 (2016)
52. Hartmann, L., Dutta, S., Opatz, S., Vosberg, S., Reiter, K., Leubolt, G., Metzeler, K.H., Herold, T., Bamopoulos, S.A., Bräundl, K., *et al.*: Zbtb7a mutations in acute myeloid leukaemia with t (8; 21) translocation. Nature communications **7**, 11733 (2016)
53. Rosenbauer, F., Wagner, K., Kutok, J.L., Iwasaki, H., Le Beau, M.M., Okuno, Y., Akashi, K., Fiering, S., Tenen, D.G.: Acute myeloid leukemia induced by graded reduction of a lineage-specific transcription factor, pu. 1. Nature genetics **36**(6), 624 (2004)
54. Steidl, U., Rosenbauer, F., Verhaak, R.G., Gu, X., Ebralidze, A., Otu, H.H., Klippel, S., Steidl, C., Bruns, I., Costa, D.B., *et al.*: Essential role of jun family transcription factors in pu. 1 knockdown–induced leukemic stem cells. Nature genetics **38**(11), 1269 (2006)
55. Brunel, H., Gallardo-Chacón, J.-J., Buil, A., Vallverdú, M., Soria, J.M., Caminal, P., Perera, A.: Miss: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. Bioinformatics **26**(15), 1811–1818 (2010)
56. Altay, G., Emmert-Streib, F.: Revealing differences in gene network inference algorithms on the network level by ensemble methods. Bioinformatics **26**(14), 1738–1744 (2010)
57. Darbellay, G.A.: An estimator of the mutual information based on a criterion for conditional independence. Computational Statistics & Data Analysis **32**(1), 1–17 (1999)

Figure 1: Comparison of F-measure values of OIPCQ and OIPCQ2 algorithms with other methods for learning DREAM3 Challenge with 10 genes, DREAM3 Challenge with 50 gene and DREAM3 Challenge with 100 genes.

Figure 2: Comparison of F-measure values of OIPCQ and OIPCQ2 algorithms with other methods for learning DREAM4 challenge for five networks of the 100-multifactorial sub-challenge.

Figure 3: Comparison of F-measure values of OIPCQ and OIPCQ2 algorithms with other methods for learning SOS-DNA and AML.

Figure 4: GRN of AML that infer by OIPCQ.

Figure 5: GRN of AML that infer by OIPCQ2.

Figure 6: Results of True Positive against False Positive for 1000 randomly-generated permutation for DREAM3 dataset with (a)10 genes, (b) 50 and (c)100 genes. The F-measure values for DREAM3 dataset with 10, 50 and 100 genes is represented in part (d).

Figure 7: Example of Paths of length 2 and greater than 2 between X and Y.

Figure 8: An example of the OIPCQ algorithm. MI and CMI denote the mutual information and conditional mutual information. $CMI_{70}(X,Y|Z,W)$ indicates the 70th percentile of the CMI values. $V_{XY} = ADJ(X) \bigcap ADJ(Y)$ and $U_{XY} = ADJ(X) \bigcup ADJ(Y)$.

Table 1: Threshold values for methods used on three sets of DREAM3-Yeast1-null-mutant dataset

	Dataset		
Algorithm	10 genes, 10 edges	50 genes, 77 edges	100 genes, 166 edges
PCA-CMI	0.03	0.03	0.05
$CN_{Consensus}$	0.6	0.6	0.6
$CN_{Ind.Test}$	(0.02 , 0.05)	(0.02 , 0.05)	(0.03 , 0.05)
CMI2NI	0.03	0.04	0.06
$OIPCQ_{\theta 1}$	0.05	0.02	0.03
$OIPCQ_{\theta 2}$	0.01	0.05	0.05

Table 2: Results for DREAM3-size10-Yeast1

Algorithm	TP	FP	PPV	TPR	ACC	F-measure	FPR	FDR	MCC
PCA-CMI	9	1	0.9	0.9	0.95556	0.9	0.02857	0.1	0.87143
CN	9	1	0.9	0.9	0.95556	0.9	0.02857	0.1	0.87143
CMI2NI	9	1	0.9	0.9	0.95556	0.9	0.02857	0.1	0.87143
OIPCQ	9	**0**	**1**	0.9	**0.97778**	**0.94737**	**0**	**0**	**0.93541**
OIPCQ2	9	**0**	**1**	0.9	**0.97778**	**0.94737**	**0**	**0**	**0.93541**

Table 3: Comparison of CMI2NI, OIPCQ and OIPCQ2 methods based on the three important pathways resulted by KEGG. Among the significant pathways, three most significant ones were selected. The pathway enrichment was done separately for each of the four regulators ZBTB7A, PU1, BCLAF1 and NRSF. Column 1 indicates the name of pathways and number of genes in pathways. The name of regulators and algorithms are in columns 2 and 3, respectively. The overlap between gene sets represented as pathways shows in columns 4. The related p-values and q-values are collected in columns 5 and 6, respectively.

No.	Pathway (number of genes in pathways)	Regulator	Method	Genes overlapped	p-value	q-value
1	CHRONIC MYELOID LEUKEMIA (73 genes)	BCLAF1	CMI2NI/OIPCQ/OIPCQ2	22/**25**/**25**	2.90e-35/2.17e-41/2.17e-41	2.41e-32/1.80e-38/1.80e-38
2	ACUTE MYELOID LEUKEMIA (60 genes)	BCLAF1	CMI2NI/OIPCQ/OIPCQ2	18/**20**/**20**	1.94e-28/3.03e-32/ 3.03e-32	5.38e-26/8.41e-30/ 8.41e-30
3	PATHWAYS IN CANCER (328 genes)	BCLAF1	CMI2NI/OIPCQ/OIPCQ2	28/**31**/**31**	5.05e-30/2.41e-34/2.41e-34	2.10e-27/1.00e-31/1.00e-31
1	CHRONIC MYELOID LEUKEMIA (73 genes)	NRSF	CMI2NI/OIPCQ/OIPCQ2	18/**26**/**26**	2.94e-29/8.90e-44/8.90e-44	2.45e-26/7.42e-41/ 7.42e-41
2	ACUTE MYELOID LEUKEMIA (60 genes)	NRSF	CMI2NI/OIPCQ/OIPCQ2	**15**/14/14	3.32e-24/6.78e-20/6.78e-20	1.38e-21/1.88e-17/1.88e-17
3	PATHWAYS IN CANCER (328 genes)	NRSF	CMI2NI/OIPCQ/OIPCQ2	22/**28**/**28**	1.02e-23/5.47e-29/ 5.47e-29	2.83e-21/2.28e-26/2.28e-26

Figures

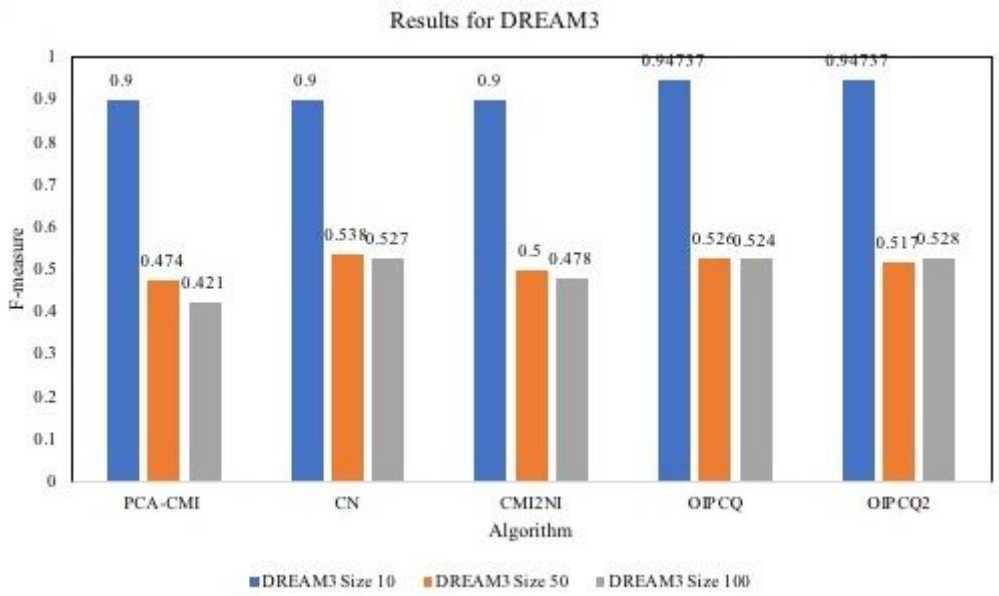

Figure 1

Comparison of F-measure values of OIPCQ and OIPCQ2 algorithms with other methods for learning DREAM3 Challenge with 10 genes, DREAM3 Challenge with 50 gene and DREAM3 Challenge with 100 genes.

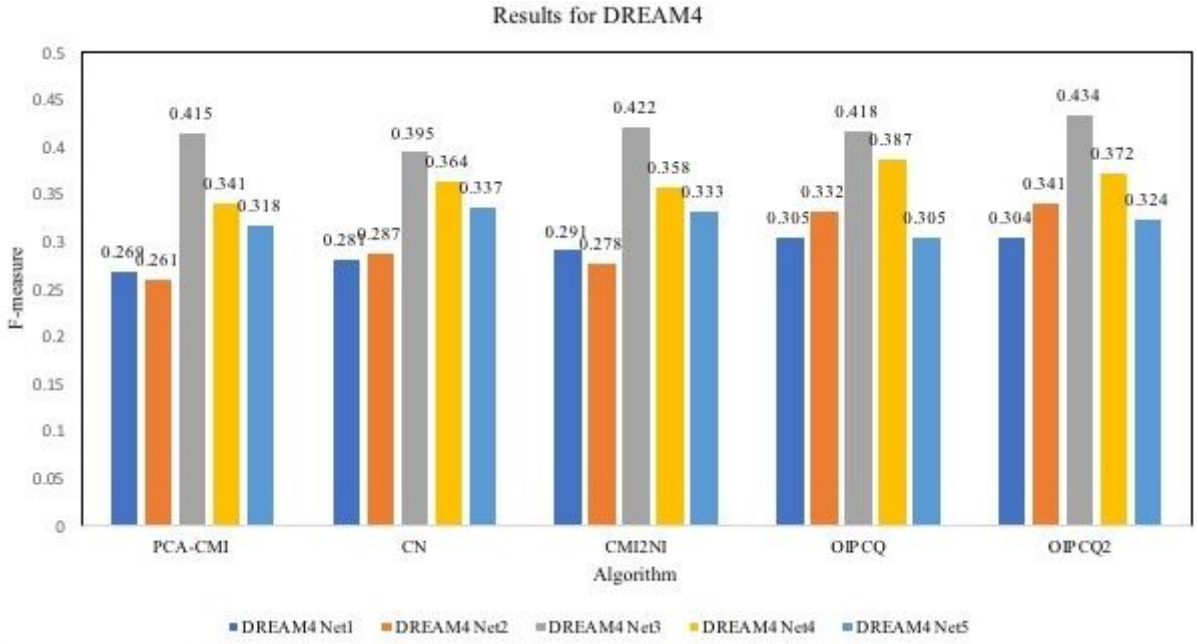

Figure 2

Comparison of F-measure values of OIPCQ and OIPCQ2 algorithms with other methods for learning DREAM4 challenge for five networks of the 100- multifactorial sub-challenge.

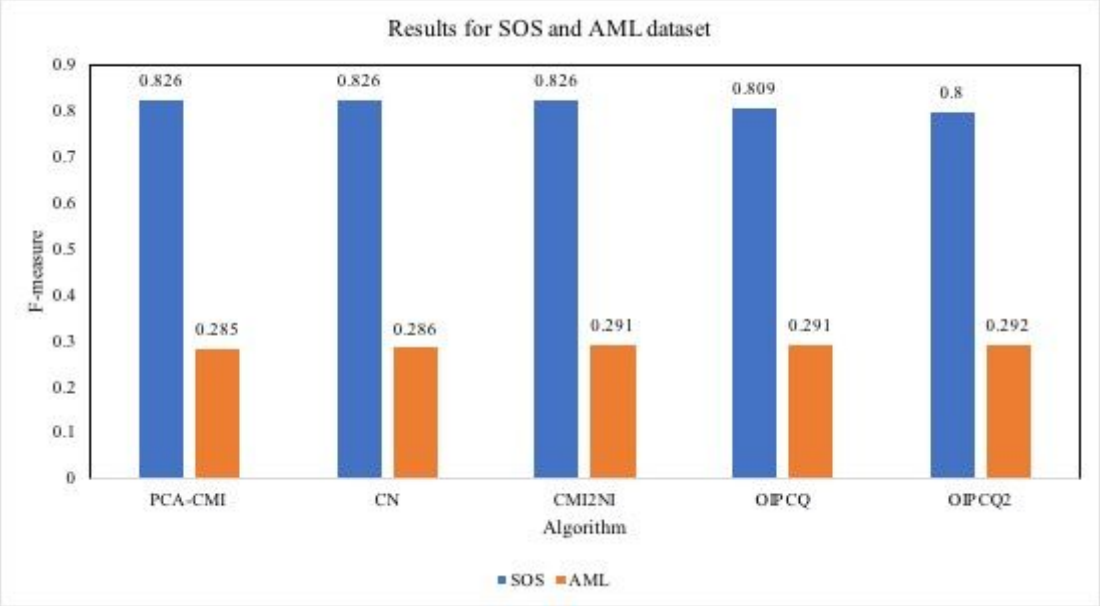

Figure 3

Comparison of F-measure values of OIPCQ and OIPCQ2 algorithms with other methods for learning SOS-DNA and AML.

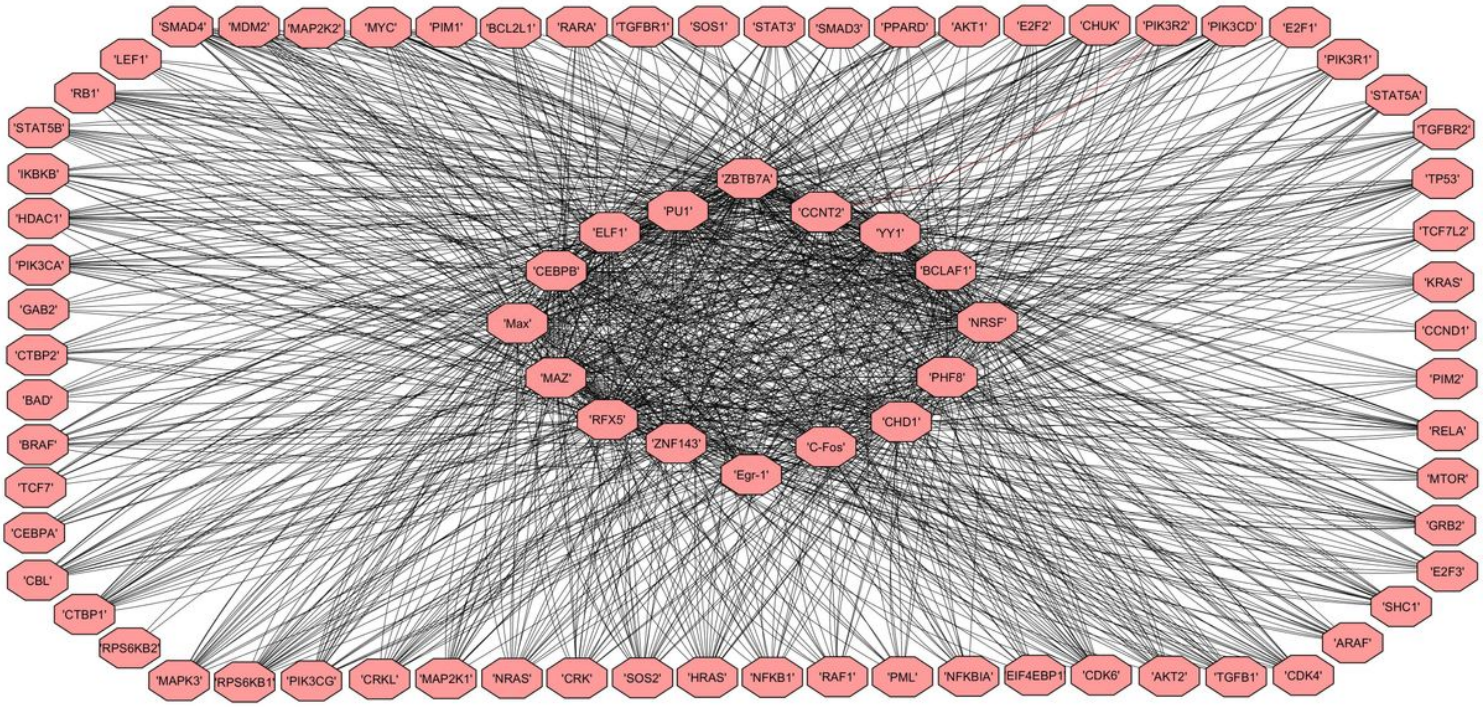

Figure 4

GRN of AML that infer by OIPCQ.

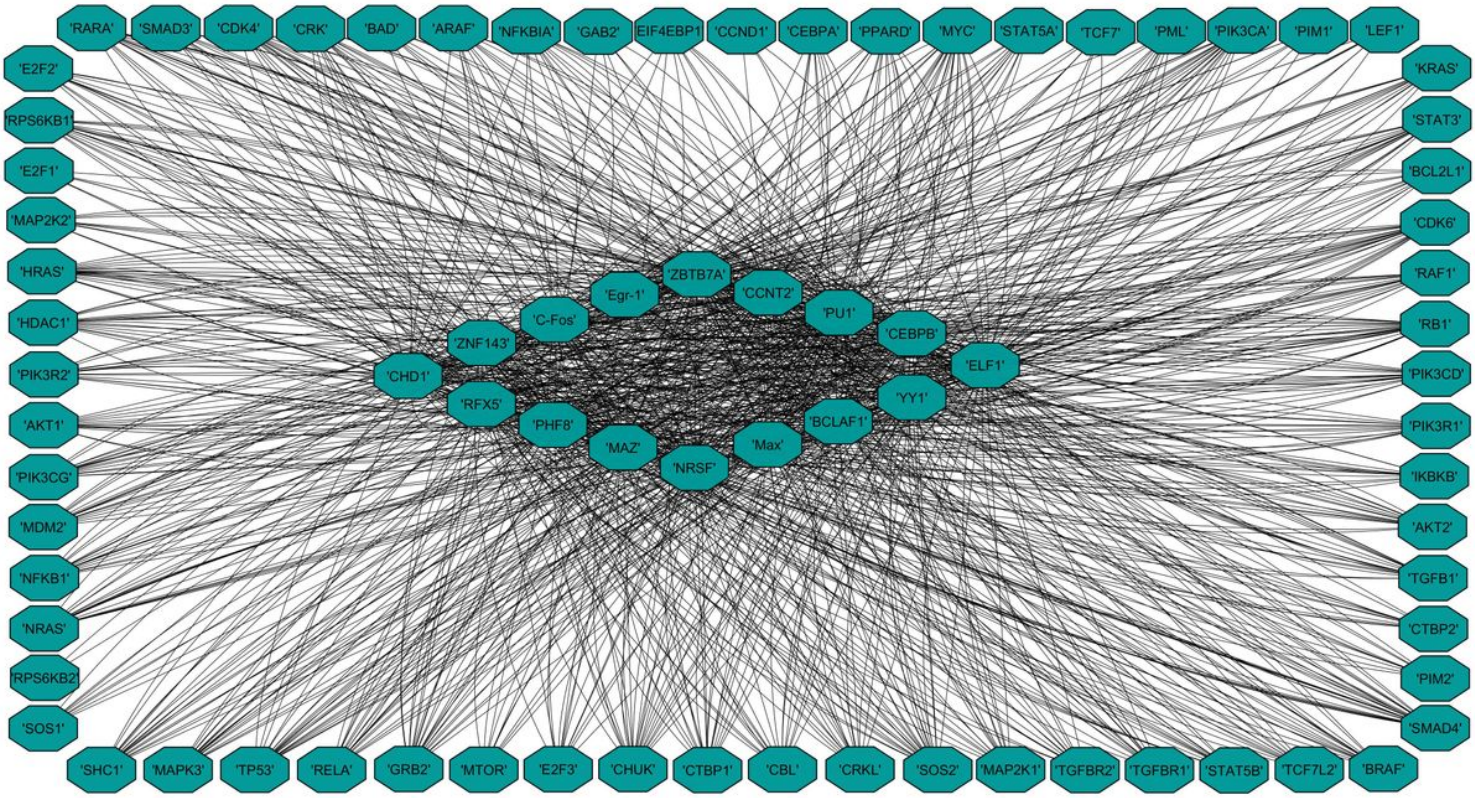

Figure 5

GRN of AML that infer by OIPCQ2.

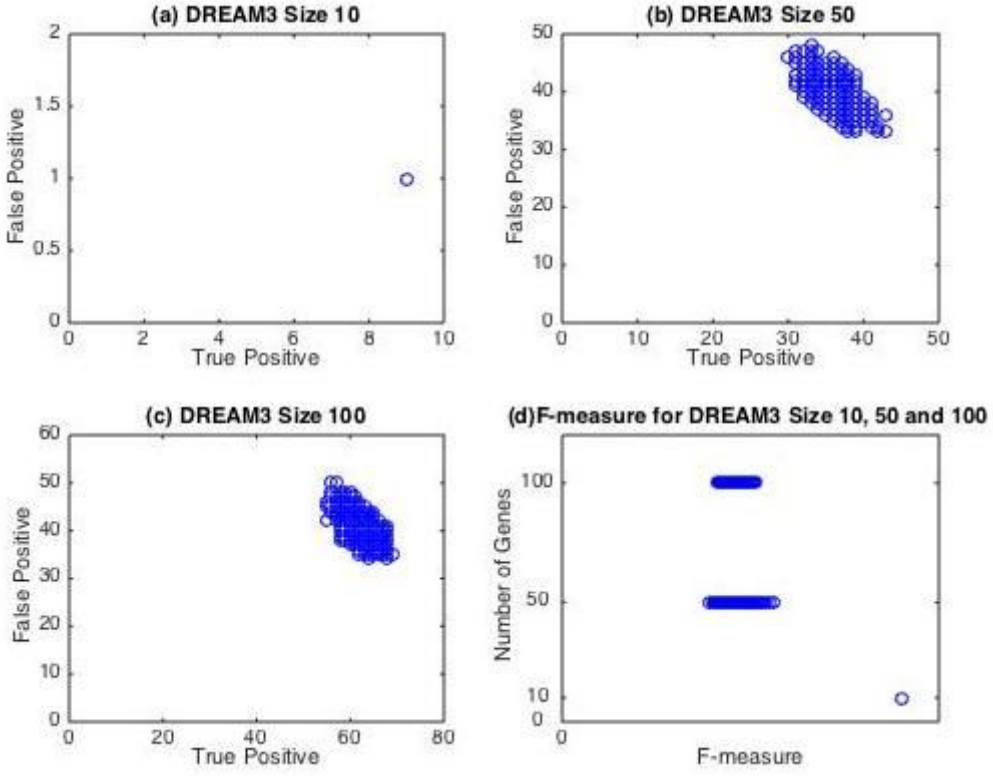

Figure 6

Results of True Positive against False Positive for 1000 randomly-generated permutation for DREAM3 dataset with (a)10 genes, (b) 50 and (c)100 genes. The F-measure values for DREAM3 dataset with 10, 50 and 100 genes is represented in part (d).

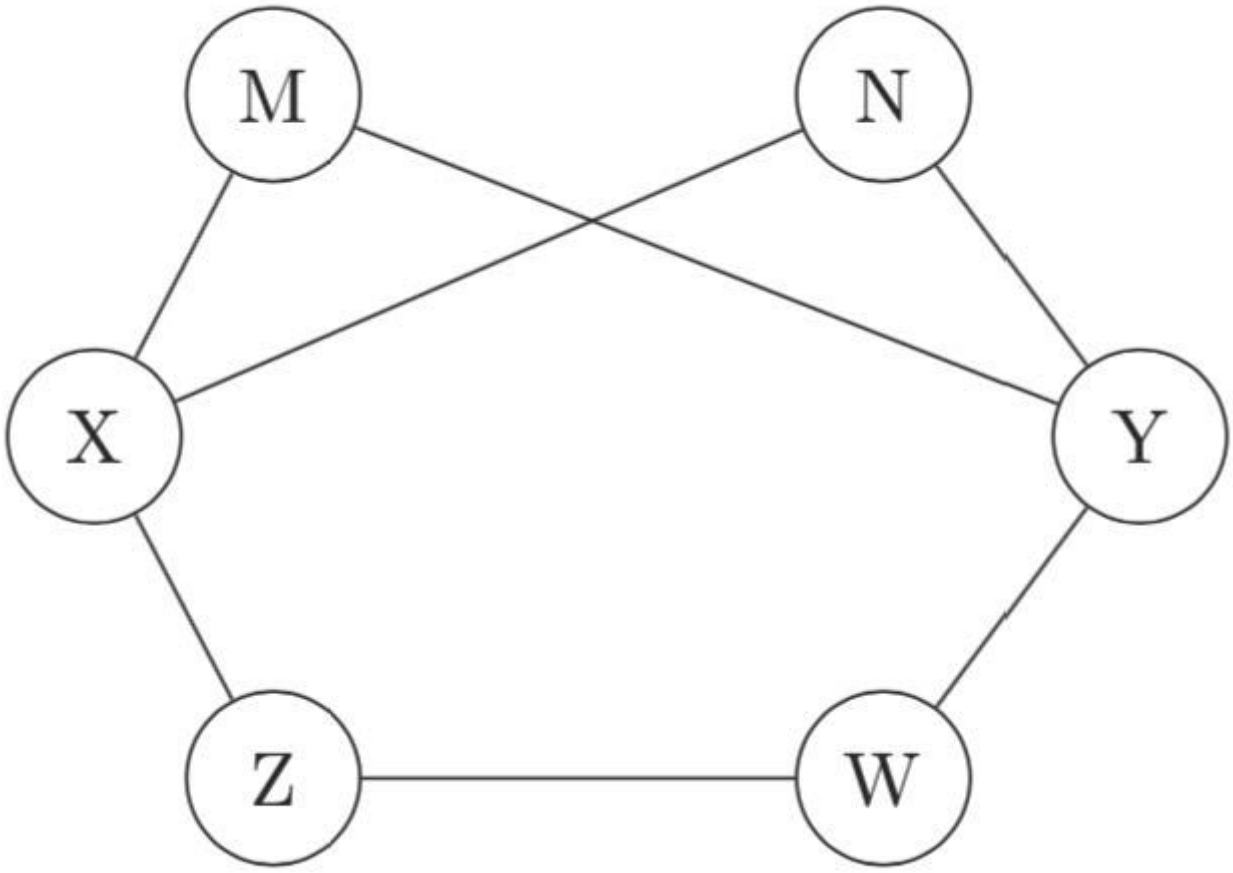

Figure 7

Example of Paths of length 2 and greater than 2 between X and Y .

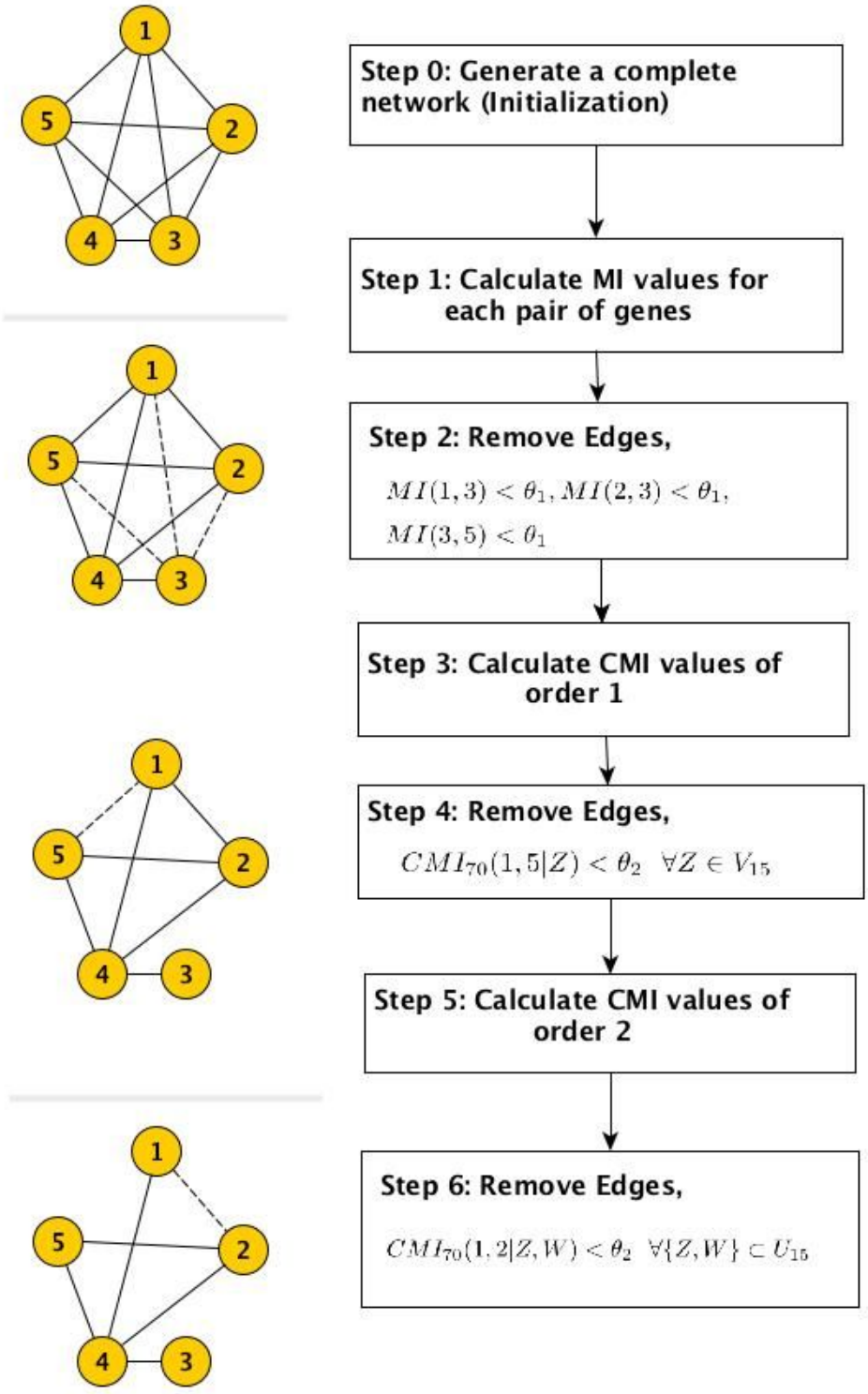

Figure 8

An example of the OIPCQ algorithm. MI and CMI denote the mutual information and conditional mutual information. CMI70(X; Y jZ;W) indicates the 70th percentile of the CMI values. VXY = ADJ(X) T ADJ(Y) and UXY = ADJ(X) S ADJ(Y).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Sfile.pdf