PLOS ONE

RESEARCH ARTICLE

Inferring Gene Regulatory Networks Using Conditional Regulation Pattern to Guide Candidate Genes

Fei Xiao, Lin Gao*, Yusen Ye, Yuxuan Hu, Ruijie He

School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi 710071, China

* lgao@mail.xidian.edu.cn

Citation: Xiao F, Gao L, Ye Y, Hu Y, He R (2016) Inferring Gene Regulatory Networks Using Conditional Regulation Pattern to Guide Candidate Genes. PLoS ONE 11(5): e0154953. doi:10.1371/journal.pone.0154953

Editor: Enrique Hernandez-Lemus, National Institute of Genomic Medicine, MEXICO

Received: November 9, 2015

Accepted: April 21, 2016

Published: May 12, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the NSFC (Grant No.61532014 & No.61432010 & No.61402349), and the Fundamental Research Funds for the Central Universities (No. BDZ021404) The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Combining path consistency (PC) algorithms with conditional mutual information (CMI) are widely used in reconstruction of gene regulatory networks. CMI has many advantages over Pearson correlation coefficient in measuring non-linear dependence to infer gene regulatory networks. It can also discriminate the direct regulations from indirect ones. However, it is still a challenge to select the conditional genes in an optimal way, which affects the performance and computation complexity of the PC algorithm. In this study, we develop a novel conditional mutual information-based algorithm, namely RPNI (Regulation Pattern based Network Inference), to infer gene regulatory networks. For conditional gene selection, we define the *co-regulation pattern*, *indirect-regulation pattern* and *mixture-regulation pattern* as three candidate patterns to guide the selection of candidate genes. To demonstrate the potential of our algorithm, we apply it to gene expression data from DREAM challenge. Experimental results show that RPNI outperforms existing conditional mutual information-based methods in both accuracy and time complexity for different sizes of gene samples. Furthermore, the robustness of our algorithm is demonstrated by noisy interference analysis using different types of noise.

Introduction

Inferring gene regulatory networks is a key step in understanding biological processes [1–5]. Microarray techniques generate a large amount of gene expression data, providing a workable data foundation [6]. Many computational methods were developed to infer gene regulatory networks using these high-throughput data [2, 4]. These methods can be divided into two categories: the model-based and the machine learning-based approaches [3].

Model-methods are based mainly on singular value decomposition [7], multiple linear regression [8] and linear programming [9]. In machine learning methods, Bayesian networks, Pearson correlation coefficient, partial correlation coefficients, information theory, and conditional mutual information are applied to measure the regulation strength between genes. Bayesian networks are based on maximizing the scoring function, for the moment, dynamic programming is

the best way to achieve a global optimal structure with 35 nodes [10]. Although Cassio *et al.* [11] proposed a structure constraint method based on Bayesian information criterion (BIC) and Akaike information criterion (AIC), reducing the size limitation to 70 nodes, it remains an open problem due to its local optimum and high computing cost [3, 12, 13]. Pearson correlation coefficient and information theory can reconstruct large-scale networks with limited samples in acceptable time [14, 15]. Compared with Pearson correlation coefficient, mutual information (MI) provides a reasonable gauge to measure non-linear dependence (which commonly exists in biology [16]). Therefore, mutual information is widely applied in inferring gene networks [3, 16–20].

In recent years, conditional mutual information (CMI) has taken the place of MI because MI cannot distinguish the direct interactions from the indirect ones [17–19, 21]. Path consistency (PC) algorithms are an effective strategy to infer a causal network by conditional relation [14, 18, 19, 22]. Combining PC algorithm with CMI and corrected-CMI, PCA-CMI (path consistency algorithm based on conditional mutual information) [18] and CMI2NI (CMI2-based network inference) [17] are proposed to "thin" the edges with independent correlation recursively from zero to high order correlation. Theoretical analysis shows that CMI underestimates the regulatory strength in some cases [23]. CMI2 corrects the underestimation by utilizing interventional probability and KL-divergence (Kullback—Leibler divergence), however, previous methods force to select conditional genes which has exponential complexity w.r.t the data size, so it is still a challenge to select the conditional genes in an optimal way [18], which may affect the performance and sharply reduce the search space [22].

In this work, we aim to define three candidate patterns based on biological processes [24, 25] to guide the selection of candidate genes. A novel algorithm, called RPNI (Regulation Pattern based Network Inference), is developed to infer gene regulatory networks by considering the candidate patterns and PC algorithm based on CMI2 to delete the edges with independent correlation recursively. We also make statistical analysis using different scales of yeast networks. Z-tests show that our defined candidate patterns significantly exist in gene regulatory networks, consistent with the discovered regulation motifs [23, 24]. Our method also greatly reduces the computational complexity. Under the hypothesis of Gaussian distribution of gene expression data, CMI2 can be calculated in a simple form using a covariance matrix of related gene expression data [18]. RPNI follows CMI2's strength to measure the regulatory strength. Moreover, it can accurately predict regulatory networks using limited samples. We apply our algorithm to DREAM data [2, 26, 27], and experimental results show that RPNI outperforms PCA-CMI and CMI2NI in both accuracy and time complexity. Furthermore, the robustness of our algorithm is demonstrated by noisy interference analysis using different types of noise.

Methods

This section includes an introduction to some definitions of information theory, a path consistency algorithm, our defined candidate patterns and the RPNI algorithm for inferring gene regulatory networks.

Information theory

With the advantages of measuring non-linear dependence association between two variables and relatively high efficiency, information theory is increasingly used to measure the regulatory strength between genes. The definitions of mutual information (MI) and conditional mutual

information (CMI) are as follows:

$$MI(X, Y) = \iint p(x, y)\log\frac{p(x, y)}{p(x)p(y)}dxdy \quad (1)$$

$$CMI(X, Y|Z) = \iiint p(x, y, z)\log\frac{p(x, y|z)}{p(x|z)p(y|z)} \quad (2)$$

where $p(x,y)$ denotes the joint distribution of X and Y. $p(x)$ and $p(y)$ represent the marginal distribution of x and y, respectively. Since it is widely accepted that gene expression data follow Gaussian distribution [18, 19], formulation of entropy subject to n-dim Gaussian distribution can be easily calculated by a simple equation, where $|C|$ is the determinant of covariance matrix of variables $x_1,x_2,\ldots,x_n$ [28].

$$H(X) = \log(2\pi e)^{\frac{n}{2}}|C|^{-\frac{1}{2}} \quad (3)$$

After mathematical transformation, we can obtain the following equation, guiding us to compute MI and CMI2.

$$MI(X, Y) = \frac{1}{2}\log\frac{|C(X)| \times |C(X)|}{|C(X, Y)|} \quad (4)$$

CMI2 proposed to integrate Kullback—Leibler divergence [28] and interventional probability in order to correct the underestimation of CMI [23],

$$CMI2(X, Y|Z) = \sum_{x,y,z} p(x, y, z)\ln\frac{p(x, y, z)}{p(x, z)\sum_x p(y|x, z)p(x) + p(y, z)\sum_y p(x|z, y)p(y)} \quad (5)$$

With the same hypothesis of Gaussian distribution, CMI2 can be easily calculated. The details of computational process and mathematical proof can be found in Zhang's work [18].

Path consistency algorithms

Path consistency (PC) algorithms are widely used in inferring gene regulatory networks [14, 18, 19]. By removing the most likely uncorrelated edges repeatedly from low to high order dependence correlation until it can't continue, PC-algorithm can construct a high-confidence undirected network [22].

Candidate Pattern

We define the *co-regulation pattern*, *indirect-regulation pattern* and *mix-regulation pattern* to facilitate the selection of candidate genes in inferring gene regulatory networks.

Single-input co-regulation pattern (also denoted as the single input motif) is defined as a pattern in which a set of target genes are regulated by a single gene (Fig 1a), in other words, two or more genes share the same upstream gene in this pattern and guide the deleting of false positive (FP) edges [18]. Single-input co-regulation pattern occurs infrequently in randomized networks ($p<0.01$) and is potentially useful for coordinating a discrete unit of biological function. For example, several genes in the leucine biosynthetic pathway are regulated by the Leu3 transcriptional regulator [23]. In Fig 1a, gene X and gene Y have a common upstream gene A (i.e. gene A regulates gene X and gene Y at the same time). This causes gene X and Y to have higher mutual information (MI), which leads to false positives. Choosing A

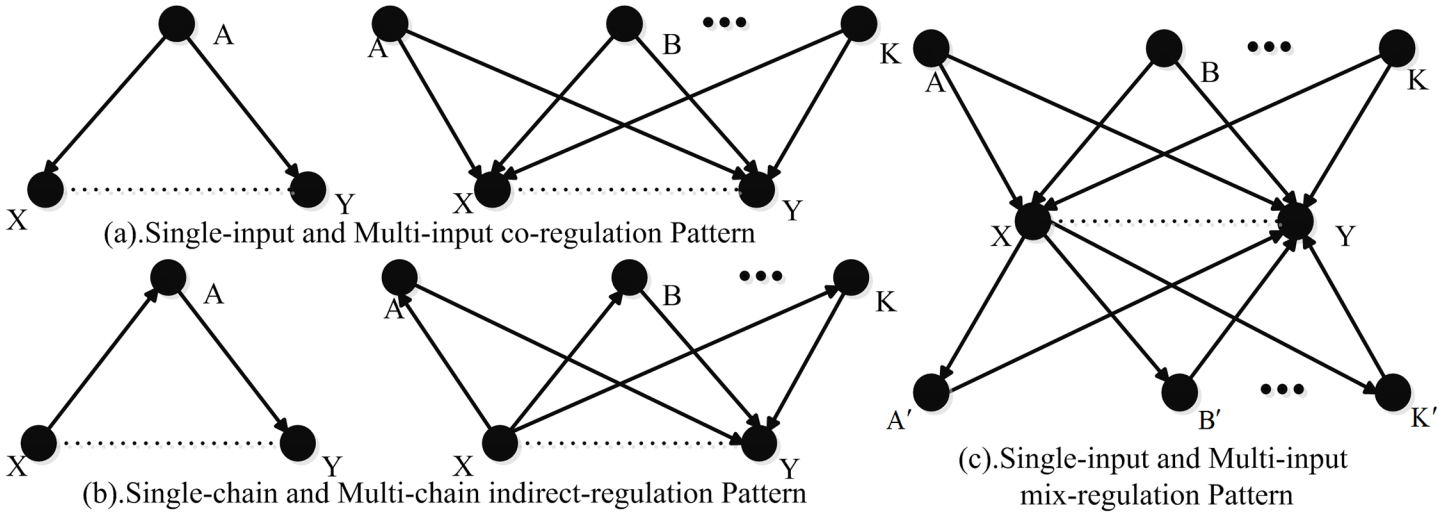

Fig 1. Diagram of candidate patterns. (**a**) *Co-regulation pattern* consists of single-input and multi-input co-regulation. (**b**) *Indirect-regulation pattern* consists of single-chain and multi-chain indirect-regulation. (**c**) *Mix-regulation pattern* includes both co-regulation and indirect-regulation.

doi:10.1371/journal.pone.0154953.g001

as the conditional gene can significantly reduce the MI between X and Y and guide the deleting of false positive (FP) edges [19]. As an extension of co-regulation, we take into account the situation of more than one regulator, whose structure is denoted as the *multi-input co-regulation pattern* [24]. Experiment indicates that the sets of genes regulated by different transcription factors in *E.* coli share much more common genes than expected at random [25] for both cases. In this scenario, their co-regulators are selected as conditional genes. Both single-input co-regulation pattern and multi-input co-regulation pattern are collectively called co-regulation pattern.

Second, *single-chain and multi-chain indirect-regulation pattern* is defined as a pattern in which a gene is both directly or indirectly regulated by two or more genes. As is shown in Fig 1b, target gene Y is both directly and indirectly regulated by gene X and gene A. This structure is also denoted as the *regulator chain motif*. It consists of a chain in which one regulator binds the promoter of a second regulator and the second binds the promoter of a third regulator, and so forth. This network motif is observed frequently in the location data for yeast regulators [25]. As mentioned above, MI cannot distinguish the direct interactions or correlations from indirect ones, which leads to FP [19]. Here, choosing gene A as the candidate gene can guide the deleting of FP edges. Both single-chain and multi-chain indirect-regulation pattern are collectively called indirect-regulation pattern.

Third, in *the mix-regulation pattern* (Fig 1c), gene X and gene Y are affected by both co-regulation and indirect-regulation. It is evident that choosing genes in $\{A,B,\ldots,K\}$ and $\{A',B',\ldots,K'\}$ as the conditional genes can remove the FP edges between gene X and gene Y. The case containing co-regulation pattern and indirect-regulation pattern simultaneously is called mix-regulation pattern.

RPNI

Given an expression dataset with n genes and m samples, we develop a novel algorithm, called RPIN, to infer gene regulatory network. Firstly, we focus on the identification of the three

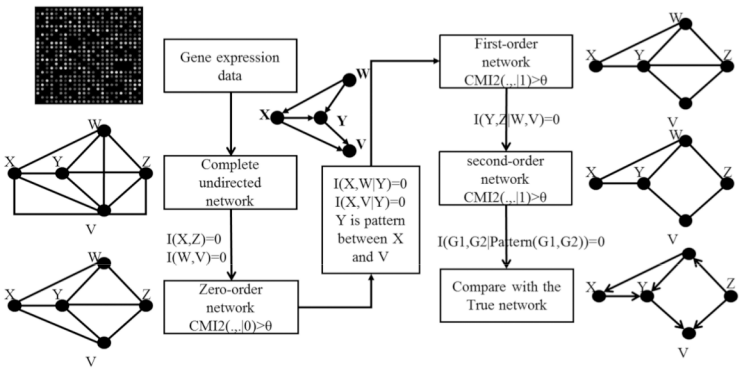

Fig 2. Diagram of RPNI. First, we generate a complete undirected graph. Second, we choose independent edges by MI between any two nodes. If MI is smaller than θ, the corresponding edge will be deleted. Here, $I(X,Z)$ and $I(W,V)$ are equal to zero, so the edges $E(X,Z)$ and $E(W,V)$ are deleted and we obtain the zero—order network. The first-order network is then constructed by deleting $E(X,V)$ because $I(X,V|Y) = 0$ and Y,X,V satisfy the *co-regulation pattern*. Based on n-order network, we construct the n+1-order network by deleting the conditional uncorrelated edges with the evidence of $I(A,B|$choosing any n+1 combination in pattern $(A,B))$. The algorithm terminates after construction of the second-order network because there are no enough regulation pattern genes to compute the third-order conditional mutual information.

doi:10.1371/journal.pone.0154953.g002

patterns. Under the hypothesis of Gaussian distribution, for a perturbed gene, we use z-tests to select differentially expressed genes as its upstream genes. As shown in Fig 1a, in the *co-regulation pattern*, gene A is the co-upstream gene of gene X and Y, in the *indirect-regulation pattern*, gene A is the upstream gene of gene X and downstream gene of gene Y. We thus select gene A is candidate gene of gene X and gene Y. Fig 2 illustrates the flow chart of our RPNI algorithm, using a network consisting of five genes as an example. First, we generate a complete undirected graph with five nodes. Second, we choose independent edges by MI between any two nodes. If MI is smaller than θ, the corresponding edge will be deleted. Here, $I(X,Z)$ and $I(W,V)$ are equal to zero, so the edges $E(X,Z)$ and $E(W,V)$ are deleted and we obtain the zero—order network. The first-order network is then constructed by deleting $E(X,V)$ because $I(X,V|Y) = 0$ and Y,X,V satisfy the *co-regulation pattern*. Based on n-th order network, we construct the n+1-th order network by deleting the conditional uncorrelated edges with the evidence of $I(A,B|$choosing any n+1 combination in pattern $(A,B))$. The algorithm terminates after construction of the second-order network because there are not enough regulation pattern genes to compute the third-order conditional mutual information. The detail procedure of this algorithm is described in Box 1.

Results

Datasets and evaluation metrics

In order to compare our method with CMI and CMI2, we apply these methods to infer gene regulatory networks using the same dataset from DREAM3 challenge and acute myeloid leukemia (AML) based on the Level-3 processed RNA sequencing data of AML patient from TCGA (http://cancergenome.nih.gov/) [29, 30].

The performance of the methods is evaluated using true positive rate (*TPR*), false positive rate (*FPR*), positive predictive value (*PPV*), accuracy (*ACC*) and Matthews coefficient constant

Box 1

Algorithm RPNI

Input:
Gene expression matrix A,
Dependence threshold θ.
Output:
Inferred gene regulatory network G,
Regulatory strength of each edge,
Order of inferred network L.
Step-1. Initialization. Generate a complete connected network G_0. Set $L:=-1$.
Step-2. Choose candidate gene set. $L:=L+1$; For each existing edge (i.e. $G_0(i,j)\neq 0$), select adjacent gene connected with both gene i and j. If their common neighbors form a pattern, add it to candidate gene set. Compute the number of genes (T) in the candidate gene set.
Step-3. Set $G:=G_0$. If $T<L$, stop the algorithm; else, select L genes from these T genes and add each combination into the set $K=\{k_1,k_2,\ldots,k_n\}, n=C_T^L$, compute these $CMI2(i,j|k)$ and choose the maximal one denoted as $CMI2_{\max}(i,j|k)$. If $CMI2_{\max}(i,j|k)<\theta$, set $G(i,j)=0$.
Step-4. If $G=G_0$, stop the algorithm; else, set $G_0=G$ and return to Step-2.

(*MCC*) [18]. Their definitions are as follows:

$$TPR=\frac{TP}{(TP+FN)} \tag{6}$$

$$FPR=\frac{FP}{(FP+TN)} \tag{7}$$

$$PPV=\frac{TP}{(TP+FP)} \tag{8}$$

$$ACC=\frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{9}$$

$$MCC=\frac{(TP\times TN-FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{10}$$

where *TP*,*FP*,*TN* and *FN* denote the number of true positives, false positives, true negatives and false negatives, respectively.

We also plot the receiver operating characteristic (ROC) curves and calculate the area under curve (AUC) [18, 31] which is the area under the ROC. Finally, we compare the running time between our method and CMI2 in the same parameter and environment.

Performance on simulation data

We use the gene expression data from DREAM3 challenge, which aims at reconstruction of gene networks from steady state data. There are three sub-challenges corresponding to three

Table 1. Comparison of different methods using networks with sizes 10, 50.

Method	TP	FP	TN	FN	TPR	FPR	PPV	ACC	MCC	AUC
Size 10										
PCA-CMI	9	1	34	1	0.9	0.028	0.9	0.956	0.8714	0.9343
CMI2NI	9	1	34	1	0.9	0.028	0.9	0.956	0.8714	0.9757
RPNI	9	1	34	1	0.9	0.028	0.9	0.956	0.8714	**0.9929**
Size 50										
PCA-CMI	29	34	1114	48	0.377	0.029	0.46	0.933	0.3813	
CMI2NI	32	31	1117	45	0.416	0.027	0.508	0.938	0.427	
RPNI	42	43	1105	35	0.545	0.037	0.494	0.936	**0.4852**	

doi:10.1371/journal.pone.0154953.t001

gene networks with 10 and 50 genes. To validate the performance of our method, we apply it to different sizes of networks and compare the performance between different methods. We choose the null-mutants data which contain the steady state levels for the wild-type and the null-mutant strains for each gene. We test RPNI on Yeast1 gene expression data with 10 genes and 11 samples, and choose 0.03 as the threshold to delete edges. The detailed results are shown in Table 1, and the ROC curves are plotted in Fig 3, which shows that RPNI is superior to both PCA-CMI and CMI2NI.

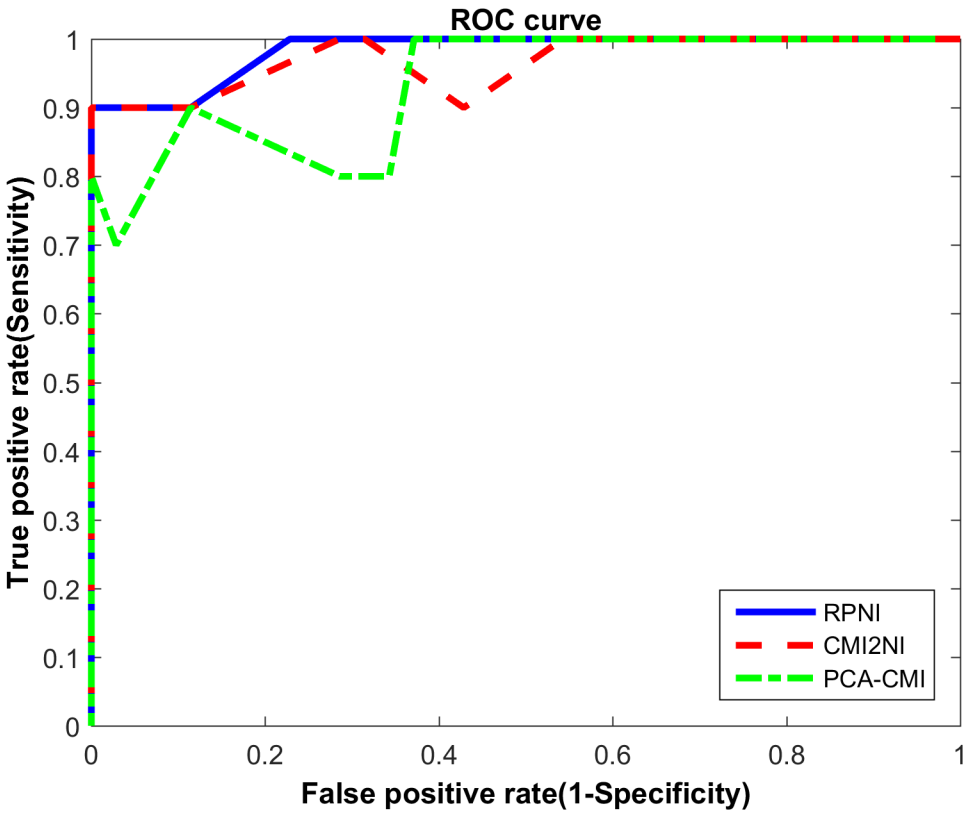

Fig 3. ROC curves generated using several methods including RPNI, CMI2NI, PCA-CMI on DREAM3 challenge Yeast1 dataset in size 10. The blue solid line is the ROC curve of RPNI. The AUC value reaches 0.9929.

doi:10.1371/journal.pone.0154953.g003

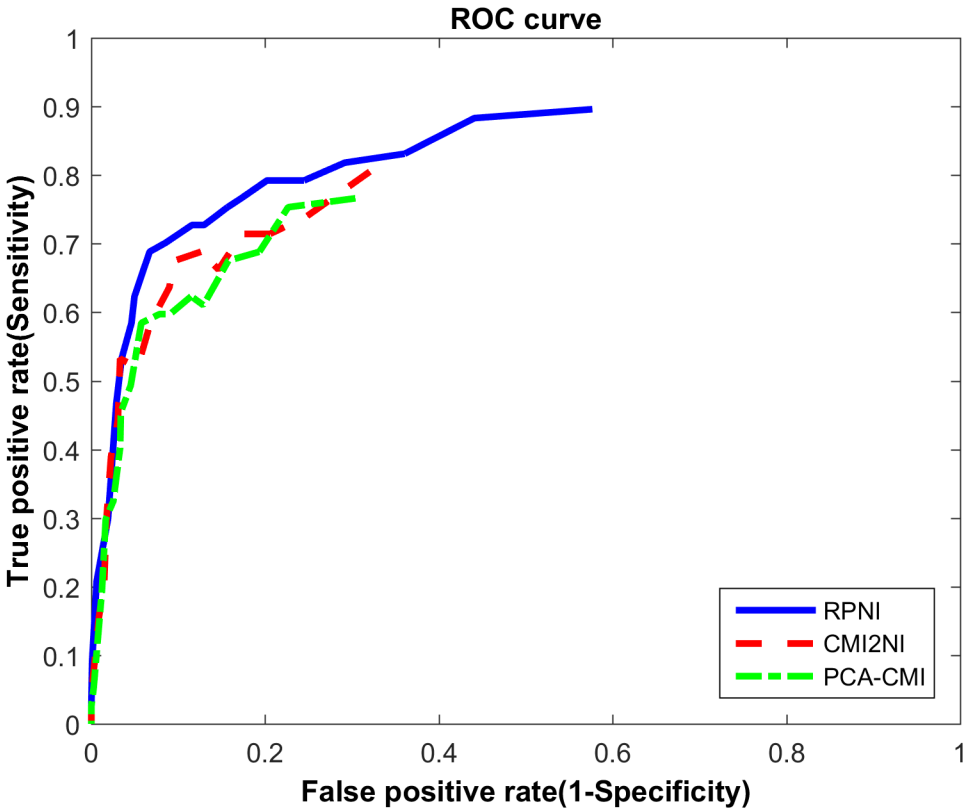

Fig 4. ROC curves of several methods on DREAM3 challenge Yeast1 dataset in size 50.

doi:10.1371/journal.pone.0154953.g004

Second, we test RPNI on Yeast1 gene expression data with 50 genes and 51samples. We choose the same threshold of 0.05 as in Zhang's work [17] to delete edges. The detailed results are listed in Table 1.

The time complexity of our algorithm for a graph G is bounded by the largest degree in G. Let k be the maximal degree of any vertex and let n be the number of vertices. Then, in the worst case, the $T(n)$ of CMI2 required by the algorithm is bounded by $\frac{n^2(n-1)^{k-1}}{(k-1)!}$. The closer to 0 the threshold, the closer to $\frac{n^2(n-1)^{k-1}}{(k-1)!}$ the calculation counts [22]. So we cannot compare the complete ROC curves because when the threshold is too small, the time complexity will reach $O\left(\frac{50^2(49)^{47}}{(47)!} * t\right)$. Assuming one second can accomplish 100000 counts of CMI2 (actually it is 1000 times per second on Intel i5-3470 3.20GHz). The computation time of this algorithm is 8.4E10 years. Based on the above discussion, we conclude that it is meaningless when threshold is too small, leading to deleting few edges. Fig 4 shows the ROC curve with the parameter ranging from 0.001 to infinity. ROC curve shows our method outperforms other methods [19] in accuracy.

Case study: cancer regulatory networks

As a case study, we construct a gene regulatory network for cancer which can provide a global view of disease-causing gene regulations [32]. We thus use RPNI to build a gene regulatory network for acute myeloid leukemia (AML) based on the Level-3 processed RNA sequencing data

of AML patient from TCGA (http://cancergenome.nih.gov/) [29, 30]. The RPKM value is used as the gene expression level.

We constructed an AML-specific regulatory network with RPNI considering the 81 cancer genes involved in a network built by RACER [33]. Our reconstructed network consists of 16 regulators, 65 target genes, and 151 regulatory links, showed in Fig 5, among which 33 regulatory links have also been inferred by RACER [32].

In order to show the superiority of RPNI, the same gene expression data are used as input for CMI2NI to infer a regulatory network. The comparative results show that our algorithm generates 16 differential regulation links compared with CMI2NI and that the network generated using CMI2NI includes 14 differential regulation relationships. In order to verify the effectiveness of these differential regulating relationships, we hereby verify how many genes in these two target genes sets are involved in the pathways related to the AML. By analyzing the two differential target gene sets using cancer gene annotation system CaGe (http://mgrc.kribb.re.kr/cage/), we noticed that 8 in 9 target genes with RPNI and 4 in 7 target genes with CMI2NI are related to the AML. Apparently, the target gene set inferred by RPNI is more significantly enriched for AML cancer pathways than that inferred by CMI2NI. Our method has a significantly better

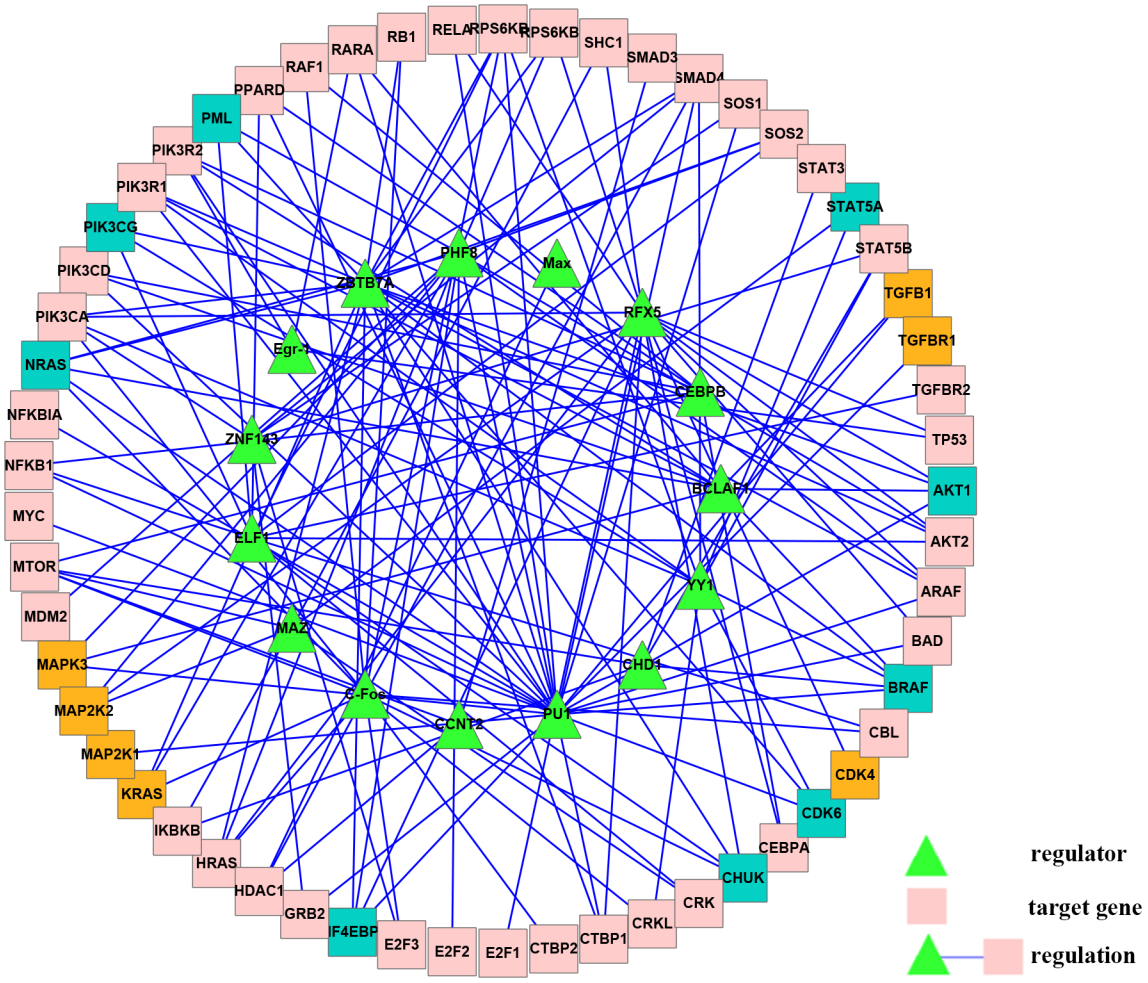

Fig 5. AML-specific gene regulatory network reconstructed by CMI2NI and RPNI. The common target genes are colored in pink. The differential target genes inferred by RPNI and CMI2NI are colored in blue and yellow, respectively.

doi:10.1371/journal.pone.0154953.g005

Table 2. Enrichment analysis results using RPNI algorithm.

No.	Base Pathways[a)]	Pathway (Database)[b)]	Genes in pathway	Genes overlapped	p-value[c)]	q-value[d)]
1	ALL	ACUTE MYELOID LEUKEMIA (KEGG)	60	8	2.04E-16	1.70E-13
2	ALL	CHRONIC MYELOID LEUKEMIA (KEGG)	73	7	4.31E-13	1.79E-10
3	ALL	PATHWAYS IN CANCER (KEGG)	328	8	1.73E-10	4.80E-08
4	ALL	ERBB SIGNALING PATHWAY (KEGG)	87	6	2.95E-10	6.14E-08
5	ALL	NON SMALL CELL LUNG CANCER (KEGG)	54	5	3.24E-09	5.40E-07
6	ALL	GLIOMA (KEGG)	65	5	8.35E-09	1.16E-06
7	ALL	PANCREATIC CANCER (KEGG)	70	5	1.22E-08	1.36E-06
8	ALL	MELANOMA (KEGG)	71	5	1.31E-08	1.36E-06
9	ALL	PROSTATE CANCER (KEGG)	89	5	4.08E-08	3.77E-06
10	ALL	INSULIN SIGNALING PATHWAY (KEGG)	137	5	3.46E-07	2.60E-05

[a)] Pathway set used for the test. CGC: 146 Cancer Gene Censers gene-based pathways, CGI: 179 Cancer Gene

[b)] Index gene-based pathways, and ALL: All 833 pathways from BioCarta/KEGG/Reactome databases.

[c)] p-value from Fisher's exact test for the overlapping gene.

[d)] q-value for false discovery rate control.

doi:10.1371/journal.pone.0154953.t002

p-value (p-value = 2.1742e-16) than CMI2NI's p-value (p-value = 1.9473e-07), statistical test using the method of fisher exact test (the detailed results are listed in Table 2).

Robustness study

We use different types of noise to demonstrate the robustness of our algorithm. First, we use measurement errors, which follows the Gaussian distribution. Considering that different genes follow different Gaussian distribution, we add a noise of Gaussian distribution to the k-th gene whose mean is 0 and variance is $\frac{\sigma(k)}{2}$ ($\sigma(k)$ is the k-th gene's variance). For each parameter, we repeat this procedure ten times and compute the mean of *FPR* and the median of *TPR* as the label of x and y, respectively. For showing the result more comprehensive, we add the box and whisker chart for each point to indicate the *TPR* range for each *FPR*. Fig 6a shows that our approach outperforms CMI2NI in robustness. Moreover, we find that our method has smaller variance in *TPR*, demonstrating its good robustness in noise. Second, outlier noise is also considered here, which often leads to recording errors or instrument errors. We replace one-tenth original expression data with noise data following Gaussian distribution with the mean and variance of all expression data. As analyzed above, we plot the box and whisker chart in the ROC curve in Fig 6b. Simulation result shows that the performance of both methods are significantly decreased, nevertheless, our result is still ahead of CMI2NI in this case. Finally, we conduct a perturbation analysis. We choose one-tenth expression data and perturb their positions randomly. This procedure is also repeated 10 times. The ROC curve reveals our method is superior to CMI2NI using perturbation data.

Discussion

Information theory-based methods show a strong ability to measure non-linear dependence that exists commonly in biology. PC algorithm is an effective strategy to "thin" the inferred graph by removing edges from zero order to higher order conditional independent relations. Due to these advantages, PCA-CMI [18] and CMI2NI [17], combining PC algorithm with CMI and CMI2, show a good performance. However, both CMI and CMI2 have not yet solved

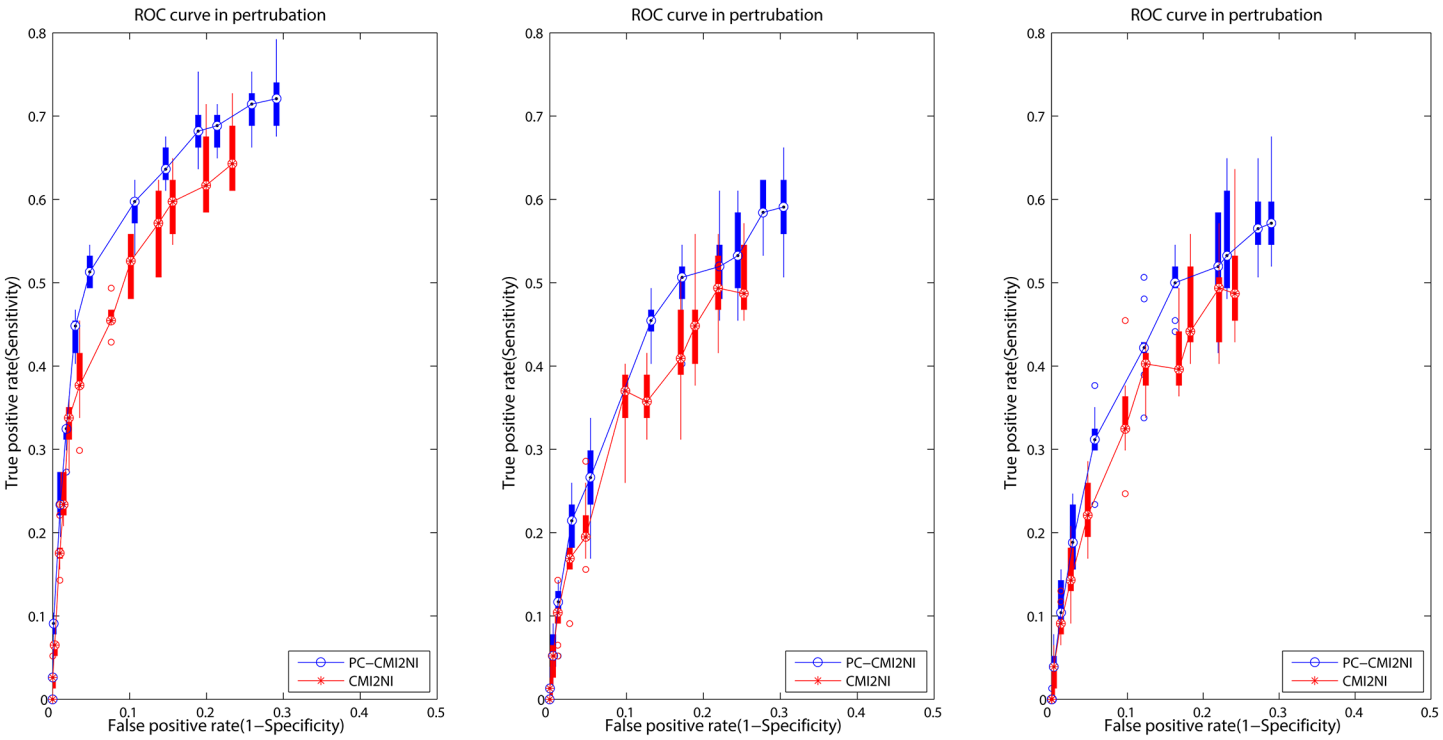

Fig 6. ROC curves of two methods of CMI2NI and RPNI using different types of noise. (A) ROC curves of two methods with noise. (B) ROC curves of two methods with outliers. (C) ROC curves of two methods with perturbation.

doi:10.1371/journal.pone.0154953.g006

the challenge of how to select the conditional genes in an optimal way. In this paper, we propose three candidate patterns, namely *co-regulation pattern*, *indirect-regulation pattern* and *mix-regulation pattern*, to guide the choice of candidate genes. Choosing reasonable conditional genes may improve the performance of PC algorithm. Actually, not limiting candidate genes will lead to deleting some true positive edges for random noise, which is a key barrier in improving the accuracy of regulatory network inference. On the basis of CMI2, we propose a novel network inference algorithm, namely RPNI, to infer gene regulatory networks. Selecting candidate gene set sharply reduces the search space in PC algorithm simultaneously. Experimental results show that RPNI outperforms the state-of-art approaches in both accuracy and time complexity.

Despite the advantages of RPNI, there exist several promising directions to further improve its performance. First, RPNI cannot infer the direction of edges in the network. Combining Bayesian network model with RPNI may overcome this weakness. Second, choosing a biological significance pattern will improve the precision of inferred regulatory networks.

Supporting Information

S1 File. The sample data set used in the paper.
(TSV)

S2 File. The benchmark for the sample data set.
(TXT)

Acknowledgments

We thank Dr. Xiujun Zhang, Yuxuan Hu, and Peizhuo Wang for their helpful advice and discussions.

Author Contributions

Conceived and designed the experiments: FX YSY. Performed the experiments: FX YSY. Analyzed the data: FX YSY. Contributed reagents/materials/analysis tools: FX YSY. Wrote the paper: FX YSY YXH LG. Paper polishing: LG YXH. Equation editing: RJH FX YSY.

References

1. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004; 306 (5696):636–40. doi: 10.1126/science.1105136 PMID: 15499007.
2. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. P Natl Acad Sci USA. 2010; 107(14):6286–91. doi: 10.1073/pnas.0913357107 WOS:000276374400031.
3. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol. 2007; 3. Artn 78 doi: 10.1038/Msb4100120 WOS:000244571300005. PMID: 17299415
4. De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nat Rev Microbiol. 2010; 8(10):717–29. doi: 10.1038/Nrmicro2419 WOS:000281908700011. PMID: 20805835
5. Chai LE, Loh SK, Low ST, Mohamad MS, Deris S, Zakaria Z. A review on the computational approaches for gene regulatory network construction. Computers in biology and medicine. 2014; 48:55–65. doi: 10.1016/j.compbiomed.2014.02.011 PMID: 24637147
6. Consortium M, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature biotechnology. 2006; 24(9):1151–61. doi: 10.1038/nbt1239 PMID: 16964229; PubMed Central PMCID: PMC3272078.
7. Yeung MKS, Tegner J, Collins JJ. Reverse engineering gene networks using singular value decomposition and robust regression. P Natl Acad Sci USA. 2002; 99(9):6163–8. doi: 10.1073/pnas.092576199 WOS:000175377800076.
8. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science. 2003; 301(5629):102–5. doi: 10.1126/science.1081900 WOS:000183914700043. PMID: 12843395
9. Wang Y, Joshi T, Zhang XS, Xu D, Chen LN. Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics. 2006; 22(19):2413–20. doi: 10.1093/bioinformatics/btl396 WOS:000241271000014. PMID: 16864593
10. Koivisto M, Sood K. Exact Bayesian structure discovery in Bayesian networks. J Mach Learn Res. 2004; 5:549–73. WOS:000236327500005.
11. De Campos CP, Ji Q. Efficient structure learning of Bayesian networks using constraints. The Journal of Machine Learning Research. 2011; 12:663–89.
12. de Campos CP, Ji Q. Efficient Structure Learning of Bayesian Networks using Constraints. J Mach Learn Res. 2011; 12:663–89. WOS:000289635000001.
13. Zou M, Conzen SD. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics. 2005; 21(1):71–9. doi: 10.1093/bioinformatics/bth463 PMID: 15308537.
14. de la Fuente A, Bing N, Hoeschele I, Mendes P. Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics. 2004; 20(18):3565–74. doi: 10.1093/bioinformatics/bth445 WOS:000225786600031. PMID: 15284096
15. Meyer PE, Lafitte F, Bontempi G. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. Bmc Bioinformatics. 2008; 9. Artn 461 doi: 10.1186/1471-2105-9-461 WOS:000262998200001. PMID: 18959772
16. Brunel H, Gallardo-Chacon JJ, Buil A, Vallverdu M, Soria JM, Caminal P, et al. MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. Bioinformatics. 2010; 26(15):1811–8. doi: 10.1093/bioinformatics/btq273 WOS:000280263400001. PMID: 20562420

17. Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics. 2002; 18:S231–S40. WOS:000178836800032. PMID: 12386007

18. Zhang X, Zhao J, Hao JK, Zhao XM, Chen L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. Nucleic acids research. 2015; 43(5):e31. doi: 10.1093/nar/gku1315 PMID: 25539927; PubMed Central PMCID: PMC4357691.

19. Zhang X, Zhao XM, He K, Lu L, Cao Y, Liu J, et al. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. Bioinformatics. 2012; 28(1):98–104. doi: 10.1093/bioinformatics/btr626 PMID: 22088843.

20. Zhao WT, Serpedin E, Dougherty ER. Inferring connectivity of genetic regulatory networks using information-theoretic criteria. Ieee Acm T Comput Bi. 2008; 5(2):262–74. doi: 10.1109/Tcbb.2007.1067 WOS:000255443900010.

21. Frenzel S, Pompe B. Partial mutual information for coupling analysis of multivariate time series. Phys Rev Lett. 2007; 99(20). Artn 204101 doi: 10.1103/Physrevlett.99.204101 WOS:000251003600023. PMID: 18233144

22. Spirtes P, Glymour CN, Scheines R. Causation, prediction, and search. 2nd ed. Cambridge, Mass.: MIT Press; 2000. xxi, 543 p. p.

23. Janzing D, Balduzzi D, Grosse-Wentrup M, Scholkopf B. Quantifying Causal Influences. Ann Stat. 2013; 41(5):2324–58. doi: 10.1214/13-Aos1145 WOS:000327746100003.

24. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science. 2002; 298(5594):799–804. doi: 10.1126/science.1075090 PMID: 12399584.

25. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nature genetics. 2002; 31(1):64–8. doi: 10.1038/ng881 PMID: 11967538.

26. Marbach D, Schaffter T, Mattiussi C, Floreano D. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. J Comput Biol. 2009; 16(2):229–39. doi: 10.1089/cmb.2008.09TT PMID: 19183003.

27. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, et al. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PloS one. 2010; 5(2):e9202. doi: 10.1371/journal.pone.0009202 PMID: 20186320; PubMed Central PMCID: PMC2826397.

28. Shannon CE. The mathematical theory of communication. 1963. MD computing: computers in medical practice. 1997; 14(4):306–17. PMID: 9230594.

29. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455(7216):1061–8. doi: 10.1038/nature07385 PMID: 18772890; PubMed Central PMCID: PMC2671642.

30. Cancer Genome Atlas Research N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. The New England journal of medicine. 2013; 368(22):2059–74. doi: 10.1056/NEJMoa1301689 PMID: 23634996; PubMed Central PMCID: PMC3767041.

31. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006; 27(8):861–74.

32. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nature genetics. 2001; 29(2):153–9. doi: 10.1038/ng724 PMID: 11547334.

33. Li Y, Liang M, Zhang Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. PLoS computational biology. 2014; 10(10):e1003908. doi: 10.1371/journal.pcbi.1003908 PMID: 25340776; PubMed Central PMCID: PMC4207489.