Reconstructing directed gene regulatory network by only gene expression data

Lu Zhang
Department of Computer Science
City University of Hong Kong
Email: zhanglu295@gmail.com

Xi Kang Feng
Department of Computer Science
City University of Hong Kong
Email: xikanfeng2@gmail.com

Yen Kaow Ng
Department of Computer Science
Faculty of Information and Communication Technology
Universiti Tunku Abdul Rahman, Malaysia
Email: kalngyk@gmail.com

ShuaiCheng Li*
Department of Computer Science
City University of Hong Kong
Email: shuaicli@gmail.com
*Corresponding author

***Abstract*—Accurately identifying gene regulatory network serves an important task in understanding *in vivo* biological activities. The inference of such networks are often accomplished through the use of gene expression data. Some methods further predict the regulatory directions in these networks by using the location of eQTL single nucleotide polymorphisms, or through gene knock out/down experiments; regrettably, these additional data are not always available, especially for the samples deriving from human tissues. In this paper, we propose Context Based Dependency Network (CBDN), a method that is able to infer gene regulatory networks, complete with the regulatory directions, from only gene expression data. CBDN applies directed data processing inequality (DDPI) to distinguish between direct and transitive relationship between genes. In our experiments with simulated and real data, CBDN outperforms the current state-of-the-art approaches. When used to identify important regulators in a network, CBDN 1. correctly identified TYROBP in the network related to Alzheimer's disease; 2. predicted potential important regulators *ZNF329* and *RB1* for brain tumors.**

I. Introduction

Understanding of regulatory mechanisms can help us bridge the gap from genotype to phenotype and enlighten us with more insights on the synthesizing effects of different elements in cells. The advent of high throughput technology provides us an unprecedent opportunity to construct an atlas of these regulatory mechanisms—the gene regulatory network (GRN)—from which one can study important dynamics such as cell proliferation, differentiation, metabolism, and apoptosis.

GRN is often inferred from gene expression data, which is available in abundance from high-throughput microarray and RNA-Seq. Many computational approaches have been developed to infer the dependencies between transcription factor (TF) and its target genes from expression data. The intuitive method is to consider a regulatory dependency as the correlation of the expressions of the TF-target pair, computed through a measure such as mutual information (MI), Pearson correlation, *etc*. However, the correlations captured within the expression data include the effects of intermediary factors; unless taken into account, they will result in the inclusion of transitive edges in the GRN inferred. To overcome this, ARACNE [1], an MI-based method, distinguishes between direct and indirect dependencies by applying data processing inequality. It considers the lowest MI value among any triplet of genes as a transitive edge. CLR (context likelihood of relatedness) [2] presents a framework which considers background noise, which naturally accounts for the transitive effects. The method works on the fact that each gene's MIs or Pearson correlations with other genes follow the Gaussian distribution. This allows the gene-gene correlations to be expressed as Z-scores, thus allowing the comparison of their strengths.

Methods based on regression have also been proposed. They incorporate all the genes in a regression model; one as response variable and the others as regressors. Regression-based methods face two difficulties: 1. most of the regressors are not actually independent, hence potentially resulting in erratic regression coefficients for these variables; 2. The model suffers from severe overfitting which necessitates the use of variable selection strategies. A few successful methods have been reported. TIGRESS [3] treats GRN inference as a sparse regression problem and introduce least angle regression in conjunction with stability selection to choose target genes for each TF. GENIE3 [4] performs variables selection based on an ensemble of regression trees (Random Forests or Extra-Trees).

The other methods are proposed to improve the predicted GRNs by introducing additional information. Considering the heterogeneity of gene expression across different conditions, cMonkey [5] is designed as a bi-clustering algorithm to group genes by calculating the dependency of gene expressions and the co-occurrance of their putative *cis*-acting regulatory motifs. The genes grouped in the same cluster are implied to be regulated by the same regulator. Inferelator [6] is developed to infer the GRN for each gene cluster by regression and $L1-norm$ regularization on gene expression or protein abundance. Recently, Chen *et al.* [7] demonstrated that involving three dimensional chromatin structure with gene expression can improve the GRN reconstruction. While these methods

have relatively good performance in reconstructing GRNs, they are not able to infer regulatory directions.

There have been many attempts at the inference of regulatory directions. The directional information may be obtained from *cis* expression single nucleotide polymorphism data, called *cis*-eSNP. The *cis*-eSNPs are thought of as regulatory anchors by influencing the expression of nearby genes. Zhu *et al.* [8] developed a method called RIMBANET which reconstructs the GRN through a Bayesian network that integrates both gene expression and *cis*-eSNPs. The *cis*-eSNPs determine the regulatory direction with these rules: 1. Genes with *cis*-eSNPs can be the parent of genes without *cis*-eSNPs; 2. Genes without *cis*-eSNPs cannot be the parent of genes with *cis*-eSNPs. This strategy have seen very successful [9], [10], [11]. However, their applicability is limited by the availability of both SNP and expression data.

The inference of interaction networks is also actively studied in other fields. Recently, Dror *et al.* [12] proposed the use of a partial correlation network (PCN) to model the interaction network of a stock market. PCN computes the influence function of stock A to B, by averaging the influence of A in the connectivity between B and other stocks. The influence function is not symmetric, so the node with larger influence to the other is assigned as parent. Their framework has been extended to other fields such as immune system [13] and semantic networks [14]. Nevertheless, there is an obvious drawback in using PCNs for the inference of GRNs: PCNs only determine whether one node is at a higher level than the other. They do not distinguish between the direct and transitive interactions.

Another primary goal of GRN analysis is to identify the important regulator in a network. An important regulator is a TF that regulates most of the gene expression signature (GES) genes (e.g. differential expressed genes) in the network. Carro *et al.* [15] identified C/EBPβ and STAT3 as important regulators for brian tumor by calculating the overlap between the TF's targets and 'mesenchymal' GES genes based on Fisher's exact test. TFs are ranked by the number of overlap genes to avoid the influence of the different size of their targets. The TFs in higher eukaryotes are rare, those important genes that can influence the others in the same signalling pathway may be neglected. Zhang *et al.* [16] developed a method called KDA (key driver analysis) to calculate whether the GES genes are enriched in the targets of regulators by searching h-layer neighborhood dynamically or statically with respect to the given directed network. The regulators in KDA are extended to the genes with directed edges (eg. with *cis*-eSNPs) pointing to their downstream genes. In this paper, we define the regulators as KDA proposed. Both of the two aforementioned methods are qualitative, they do not consider the regulatory strength between regulators and their target genes. On the other hand, because the directed network is quantitatively predicted from gene expression data, we cannot regard the interactions as having the same weight.

In this study, we propose a new method, Context Based Dependency Network (CBDN), which introduces the use of an influence function to decide the edge direction. In addition, we show a directed data processing inequality (DDPI), a property of the influence function, which we use to remove transitive interactions in the partial correlation network. Thus each edge predicted by CBDN is both causal and directed. We compare the performance of CBDN to a few well-known algorithms, namely ARACNE, CLR, TIGRESS and GENIE3. In our simulation study, CBDN's result is comparable to the best result of these methods in each situation. For a realistic test, we extracted the TYROBP-oriented network which is related to Alzheimer's disease [17]. In this test, CBDN outperforms the other methods in inferring both network structure and direction. CBDN also successfully infers TYROBP as the important regulator by quantitatively considering TYROBP's influences on the other genes. For another real expression data from the patients affected by human brain tumors, CBDN predicts two potential important regulators ZNF329 and RB1 which have been proved to be associated with brain tumors. These results demonstrates the strength of CBDN in the inference of directed GRNs from gene expression data as well as its potential in predicting important regulators.

II. Methodology

The computation of CBDN consists of three stages:

1. In the first stage, the influence of each gene to the others is calculated. This done through a partial correlation network constructed from the gene expression data;

2. In the second stage, the transitive interactions are removed by DDPI;

3. In the third stage, the important regulators are infered by ranking the regulators based on their total influences to the GES genes.

A. Partial correlation network

In CBDN, a PCN is first constructed to compute the influence of each node to the others. Interaction directions are resolved by choosing the node with a larger influence as the parent. The influence of gene A to gene B is calculated by averaging the difference between the shortest topological paths of gene B to other genes with or without gene A. We assume the input data is an $m \times n$ matrix, $E = (e_{i,j})_{m\times n}$, where each row i (denoted $E_{i,\bullet}$) represents a sample; that is, one expression value per gene; and each column j (denoted $E_{\bullet,j}$) represents the expression values of a gene across all the samples.

The partial correlation between X_i and X_k given X_j is calculated as

$$PC(X_i, X_k|X_j) = \frac{Corr(X_i, X_k) - Corr(X_i, X_j)Corr(X_k, X_j)}{\sqrt{[1 - Corr(X_i, X_j)^2][1 - Corr(X_k, X_j)^2]}} \quad (1)$$

Where $Corr(X_i, X_j)$ is the Pearson correlation between two genes X_i and X_j. The influence of gene X_j for the correlation between X_i and X_k ($k \neq j$) is defined as the difference between $Corr(X_i, X_j)$ and $PC(X_i, X_k|X_j)$,

$$d(X_i, X_k|X_j) = Corr(X_i, X_k) - PC(X_i, X_k|X_j) \quad (2)$$

The influence of gene X_j to X_i, $D(X_j \rightarrow X_i)$ is the average $d(X_i, X_k|X_j)$ across all the gene X_k,

$$D(X_j \rightarrow X_i) = \frac{1}{n-1}\Sigma_{k\neq j}^{n-1}|d(X_i, X_k|X_j)| \quad (3)$$

CBDN assumes no two-gene cyclic regulation in the network, so we remove the interaction $X_i \rightarrow X_j$ if $D(X_i \rightarrow X_j) < D(X_j \rightarrow X_i)$, and vice versa.

B. Directed data processing inequality

The partial correlation network constructed above only determines whether one gene is the parent or child of another gene; it does not provide the regulatory relationship. As an example, the partial correlation network in Fig. 1 identifies X_i as the parent of X_k, but does not distinguish whether a transitive relation ($X_i \rightarrow X_j \rightarrow X_k$) exists or not ($X_i \rightarrow X_k$). Here, Data Processing Inequality (DPI) can be used to show that post-processing cannot increase the mutual information. If X_i, X_j and X_k form a Markov chain, denoted as $X_i \rightarrow X_j \rightarrow X_k$

$$MI(X_i; X_k) \leq MI(X_i; X_j) \quad (4)$$

which shows that the mutual information between the genes with transitive interaction cannot be greater than direct interaction. This observation has been used in ARACNE to remove transitive interactions for every triplet of genes. Considering the edge direction and the nature of influence function, we propose a directed data processing inequality to show that the influence of a TF which interacts transitively with the target genes cannot be greater than that of a TF which interacts directly, that is

$$D(X_i \rightarrow X_k) \leq D(X_j \rightarrow X_k) \quad (5)$$

The proof is quite straightforward. Except X_i, X_j and X_k, other genes can be separated into two categories: non-descendents of X_i ($U = \{X_m ... X_n\}$) and descendents of X_i ($V = \{X_p ... X_a\}$). For the elements in U,

$$D_1(X_i \rightarrow X_k) = \frac{1}{|U|}\Sigma_{t\neq i}^{|U|}|d(X_k, X_t|X_i)| \quad (6)$$

$$D_1(X_j \rightarrow X_k) = \frac{1}{|U|}\Sigma_{t\neq j}^{|U|}|d(X_k, X_t|X_j)| \quad (7)$$

Based on Eq. 2, X_k is conditionally independent with the elements in U given X_i or X_j, thus we have $PC(X_k, X_t|X_j) = PC(X_k, X_t|X_i) = 0$, $|d(X_k, X_t|X_i)| = |d(X_k, X_t|X_j)| = |Corr(X_k, X_t)|, \forall t \in U$. For the genes in U, X_i and X_j have the same influence to X_k, $D_1(X_i \rightarrow X_k) = D_1(X_j \rightarrow X_k)$

For the elements in V

$$D_2(X_i \rightarrow X_k) = \frac{1}{|V|}\Sigma_{t\neq i}^{|V|}|d(X_k, X_t|X_i)| \quad (8)$$

$$D_2(X_j \rightarrow X_k) = \frac{1}{|V|}\Sigma_{t\neq j}^{|V|}|d(X_k, X_t|X_j)| \quad (9)$$

Because X_k is the direct descendent of X_j, X_k is independent with other genes in V given X_j ($PC(X_k, X_t|X_j) = 0, d(X_k, X_t|X_j) = |Corr(X_k, X_t)| \geq 0, \forall t \in V$). The correlations between X_k and the other genes in V do not change when given X_i, so $|d(X_k, X_t|X_i)| = 0, \forall t \in V$. We conclude that $D_2(X_i \rightarrow X_k) = 0$ and $D_2(X_j \rightarrow X_k) \geq 0$

$$\begin{aligned} D(X_i \rightarrow X_k) &= D_1(X_i \rightarrow X_k) + D_2(X_i \rightarrow X_k) \\ &\leq D_1(X_j \rightarrow X_k) + D_2(X_j \rightarrow X_k) \\ &= D(X_j \rightarrow X_k) \end{aligned} \quad (10)$$

To account for the influence of noise, we introduce a tolerant parameter τ. A transitive relationship $X_j \rightarrow X_k$ is removed when $D(X_i \rightarrow X_k) - D(X_j \rightarrow X_k) > \tau$. Otherwise, $X_i \rightarrow X_k$ is removed. Large τ implies much more noise exists in the expression data to influence $D(X_i \rightarrow X_k)$ and $D(X_j \rightarrow X_k)$.

C. Determine the important regulators

We propose a new method to identify the important regulators in a quantitative way. Assume the genes with gene expression signature (GES) (eg. different expressed genes) are $X_{s1}, X_{s2}, ..., X_{sn}$, the total influence value (TIV) of TF X_i is $TIV(X_i) = \Sigma_{t=1}^{n} D(X_i \rightarrow X_{st})$. Regulators are ranked by their TIVs.

D. A schematic example

Here we give a schematic example based on the simulated GRN structure in Fig. 2 to interpret how CBDN determines the edge directionality and how to find the important regulator. For instance, the direction between gene 1 and gene 4 is determined by comparing $D(X_1 \rightarrow X_4)$ and $D(X_4 \rightarrow X_1)$. Gene 4 merely affects the correlation between gene 1 and gene 10,

$$\begin{aligned} PC(X_1, X_{10}|X_4) &= 0, \\ d(X_1, X_{10}|X_4) &= Corr(X_1, X_{10}) \\ D(X_4 \rightarrow X_1) &= \frac{|Corr(X_1, X_{10})|}{9} \end{aligned} \quad (11)$$

the correlation between gene 1 and other genes are not influenced given gene 4. When conditioning on gene 1, the influences are extended to seven genes (gene 2,3,5,6,7,8,9),

$$\begin{aligned} PC(X_4, X_{2,3,5,6,7,8,9}|X_1) &= 0, \\ d(X_4, X_{2,3,5,6,7,8,9}|X_1) &= Corr(X_1, X_{2,3,5,6,7,8,9}), \\ D(X_1 \rightarrow X_4) &= \frac{\Sigma_i^{2,3,5,6,7,8,9}|Corr(X_1, X_i)|}{9} \end{aligned} \quad (12)$$

The upper bound of $D(X_4 \rightarrow X_1)$ ($D(X_4 \rightarrow X_1) \leq 1$) is smaller than $D(X_1 \rightarrow X_4)$ ($D(X_1 \rightarrow X_4) \leq 7$) by random, so in general $D(X_4 \rightarrow X_1) \leq D(X_1 \rightarrow X_4)$. The edge direction is from gene 1 to gene 4.

The important regulator identified by CBDN is not required to regulate most of the GES genes. Instead, it should have large influence on them, which guarantees such regulator is always on the top level. In this example, gene 1 has the largest influence on the other genes in the network and is located on the top level.

III. Results

A. Simulation

In order to explicitly reflect the nature of directed interactions in the gene regulatory network, we simulated a tree structure in which each node has only one parent (except the root) and is merely regulated by its parent. In other words, the expression profiles of the descendents are only determined by their parents. The expression profiles for each node were sampled from Gaussian distribution. The joint distribution of the parent and one of its descendent follows bivariate Gaussian distribution with specified covariance and noise. The true positive rate (TPR) and false positive rate (FPR) are used to plot the receiver operating characteristics (ROC) curve, where $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+FN}$, TP:true positive, FN:false negative,FP:false positive. The area under ROC curve (AUC) was applied to evaluate the performance of CBDN. In addition, we incorporated uniform distributed noise weighted by $\frac{\omega}{\kappa}$ to the simulated expression profiles, where "ω" presents the amount of noise and "κ" denotes the noise level. We set "ω" to a constant (i.e. 3) and changes "κ" from 0 to 2 in the simulations. We performed the same tests on four state-of-the-art approaches (ARACNE, CLR, GENIE3 and TIGRESS) for comparison. We simulated the expression profiles of 10 genes, each of them derived from 1000 samples as shown in Fig. 2. CBDN's result is the best when no noise exists. Even with small covariance, CBDN correctly revealed the structure and regulatory orientations (TABLE I). When noise is introduced, CBDN's result remains comparable with the best result in each situation. CBDN worked well in general under medium covariance; large or small covariance make it difficult to distinguish direct and transitive interactions, especially when a large amount of noise is introduced (TABLE II-III). However, our comparison is very conservative here, since the performance of CBDN is evaluated by considering both structure and direction, while the other four methods are evaluated only on the inferred structures. Nevertheless, CBDN achieved sufficiently good performance in reconstructing the directed GRNs. We also simulated tree structures with 20, 50 ,100 genes, in which CBDN obtained very similar results with the 10 genes simulation (See **Appendix**).

B. Real data

For this test, we downloaded the processed expression data from GEO (GSE44770), which is from dorsolateral prefrontal cortex of human brains. The expression data include 230 tissues from the individuals with or without Alzheimer's disease. The negative expression values are considered missing values because of their low intensities compared with background noise. We refilled the missing values according to the average positive expression values in the same gene. Using gene expression and *cis*-eSNPs data, Zhang *et al.* [17] had earlier found the disease-related network to be regulated by TYROBP. In addition, loss-of-function-mutations were recognized in TYROBP in Finnish and Japanese patients affected by presenile dementia with bone cysts [18]. Zhang *et al.* also overexpressed either full-length or a truncated version of TYROBP in microglia cells from mouse embryonic stem cells to confirm the tructure and direction of the regulatory network (Fig. 3). From the TYROBP regulatory network, we chose 47 GES genes, the expressions of which were altered when TYROBP was overexpressed and captured by microarray data, multiple probes designed for the same gene are combined by averaging their expression values.

This dataset is then used as input with ARACNE, CLR, GENIE3, TIGRESS, and CBDN. The results are compared with the true network structure and edge directions from mouse embryonic stem cells experiment. Fig. 4 demonstrates the AUC scores for the five methods. CBDN achieves the best performance, which is 2% higher than the second best result from GENIE3. To evaluate the capability of CBDN in predicting the regulatory direction and important regulator, we assumed all the genes to be potential regulators and ranked them based on TIV. If one gene is assessed as an regulators, other genes are assumed to be GES genes. Fig. 5 lists the top 10 genes with the largest TIV, only the values of TYROBP and SLC7A7 are above eight, the validate important regulator TYROBP is ranked at the top. SLC7A7 regulates eleven GES genes (HCLS1, IL10RA, RNASE6, GIMAP2, RGS1, TNFRSF1B, IL18, SFT2D2, KCNE3, LHFPL2 and MAF) and may be another candidate regulator and required to be validated in the future.

For another experiment, we downloaded the expression data for brain tumors (GSE19114) and pre-processed them as for Alzheimer's disease. Eventually, we selected 132 'mesenchymal' gene expression signature (MGES) genes and 883 TFs from Supplementary Table 1 and 2 of the original paper [15]. Both MGES genes and TFs are combined together to calculate TIV for each TFs, because we need to consider the regulatory relationships between TFs. We are unable to identify the two key regulators (*STAT3* and *C/EBP*β) described in the original papers from the top TIV ranked TFs (Fig. 6), because we adopt different definitions and inherent characteristics of important regulators. The top two TFs, *ZNF329* and *RB1* with TIVs exceed 120, are selected as new candidate important regulators. The relationship between *ZNF329* and brain tumors is still unclear, but zinc finger protein has been proved to be associated with brain tumor. Zhao *et al.* [19] identified *ZNF325* is a transcription repressor in MAPK/ERK signaling pathway. Recently, Das *et al.* [20] made a comprehensive review to clarify the relationship between MAPK/ERK signaling pathway and brain tumors and how can inhibit this pathway to treat paediatric brain tumors. *RB1* gene is the most important cell cycle regulatory genes and the first reported human tumor suppressor gene. It has been identified to be related with a variety of human cancers including brain tumors [21]. Mathivanan *et al.* found loss of heterozygosity and deregulated expression of *RB1* in human brain tumors [22].

IV. Conclusion

In this paper, we propose a new computational method called Context Based Dependency Network (CBDN), which

TABLE I
SIMULATION RESULT FOR 10 NODES TREE WITHOUT NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.8367	0.8009	0.8765	0.8157	0.8750
0.2	1	1	1	0.8410	1
0.4	1	1	1	0.8502	1
0.6	1	1	1	0.8272	1
0.8	1	1	1	1	1

TABLE II
SIMULATION RESULT FOR 10 NODES TREE WITH 1/3 NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.6304	0.6358	0.5879	0.8107	0.8571
0.2	0.9192	0.9846	0.9884	0.8162	1
0.4	1	1	1	0.8327	1
0.6	1	1	1	0.8557	1
0.8	1	1	0.9985	0.8338	1

TABLE III
SIMULATION RESULT FOR 10 NODES TREE 2/3 NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.6904	0.6172	0.6813	0.6241	0.8571
0.2	0.6889	0.8086	0.8480	0.8309	1
0.4	0.9531	0.9599	0.9437	0.8428	1
0.6	1	1	0.9931	0.8424	0.8750
0.8	0.9333	0.9907	0.9807	0.8058	0.8750

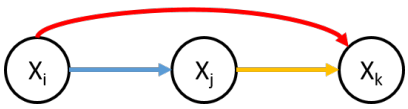

Fig. 1. The diagram for how removing transitive connections based on DDPI. We assume X_i regulates X_j(blue arrow), DDPI is used to determine whether X_i directly regulate X_k (red arrow) or not (yellow arrow).

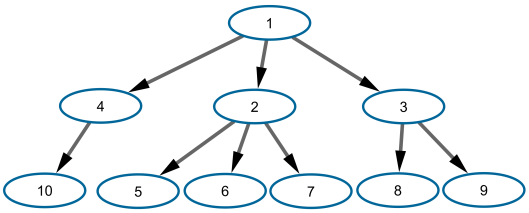

Fig. 2. The network structure and regulatory direction of each edge for simulation study.

constructs directed GRNs from only expression data. This gives us an opportunity to gain deeper insights from the readily available gene expression data that we have accumulated for years in databases such as GEO and Arrayexpress. Although gene expression data can reflect the gene-gene interactions in GRN, there are still three limitations must be noticed here. First, the transcription factors prefer act as protein complex rather than individually. The protein complex may be blocked

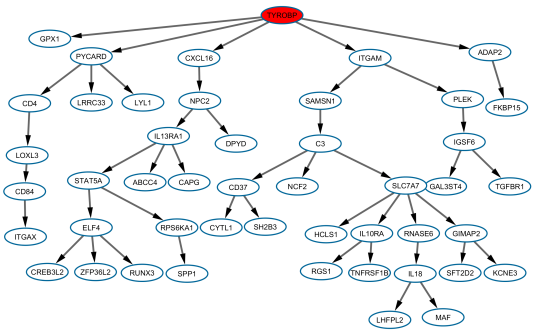

Fig. 3. The network structure for the TYROBP oriented regulatory network for Alzheimer's disease.

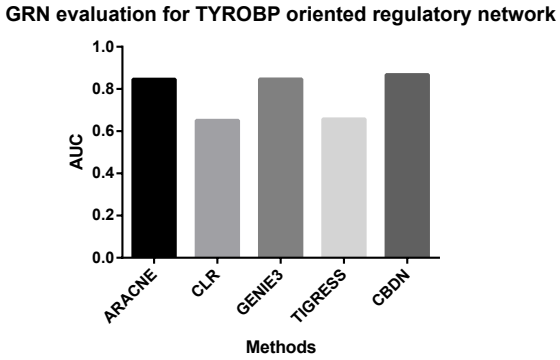

Fig. 4. The performance of different methods for TYROBP oriented regulatory network.

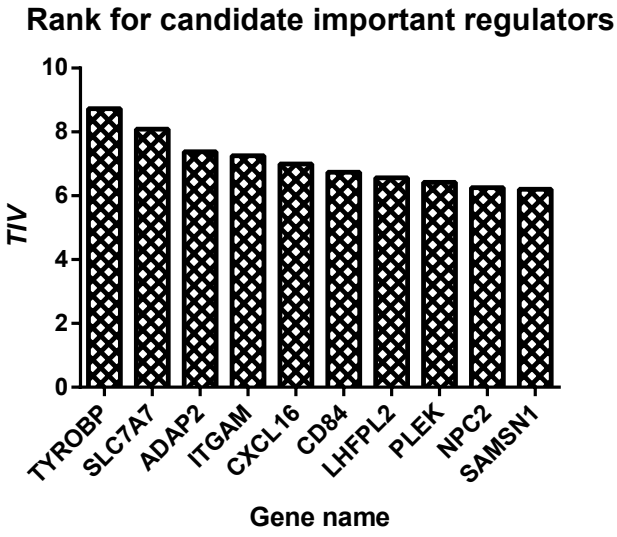

Fig. 5. The top ten genes with the largest TIV values for Alzheimer's disease.

or inactivated, for the reasons such as incorrectly folding, restricted in the nucleus, inactivated by the phosphorylation or other modifications, *etc.*, even if its transcribed mRNA has high expression level. Second, the expression of TF and TF binding are time dependency. Because the time delay exists between transcription and translation, the high mRNA expression level does not imply the protein abundance is also high simultaneously. Third, even those TFs physically bind to the target genes, different effects may be appeared because of their three dimensional distances and the histone modification presents.

The probes with low florescence signals are impossible to be distinguished from background noise, CBDN serves them as missing values and imputes them by the average value of the other samples. We have further tested other gene expression imputation methods such as the `impute` package from Bioconductor or BPCA [23], the reconstructed GRN seems stable and consistence. In the future, some noise filtering method should be incorporated in CBDN such as described in [24] and [25].

The performances of CBDN are underestimated for both simulated and real expression data. Except CBDN, the true positive results are defined as the interactions exist in both predictions and ground truth, which neglect the direction information. For CBDN, both of the interactions and directions are taken into consideration for evaluating its performance. Even through only 2% of AUC is improved in TYROBP oriented GRN inference, the result is more powerful and useful with the direction for each regulatory interaction. The performance of CBDN significantly better than other methods in some situations such as TABLE III with covariance$= 0.1$, but most of the time CBDN is only slightly better or comparable with other methods.

We believe that CBDN will be invaluable to biomedical studies by transcriptome sequencing, where there is a need for the identification of important regulators. Such studies used to be limited by the availability of SNP data to anchor regulatory directions. However, CBDN may be able to infer such important regulators from expression data alone, as it demonstrated by identifying the important regulator TYROBP in Alzheimer's disease. Because CBDN uses new concept of important regulators, it can also help us get new findings which are ignored by the approaches before.

This paper also contribute to mathematics in the form of an inequality for directed data processing (DDPI) which naturally extends the data processing inequality for mutual information. DDPI is applied to remove transitive interactions in CBDN.

In the future CBDN should be extended to predict bi-directed interactions which are quite common in nature and tackle with the TFs co-regulating a gene simultaneously by incorporating external data.

REFERENCES

[1] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7 Suppl 1, p. S7, 2006.

[2] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biol.*, vol. 5, no. 1, p. e8, Jan 2007.

[3] A. C. Haury, F. Mordelet, P. Vera-Licona, and J. P. Vert, "TIGRESS: Trustful Inference of Gene REgulation using Stability Selection," *BMC Syst Biol*, vol. 6, p. 145, 2012.

[4] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PLoS ONE*, vol. 5, no. 9, 2010.

[5] D. J. Reiss, N. S. Baliga, and R. Bonneau, "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks," *BMC Bioinformatics*, vol. 7, p. 280, 2006.

[6] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson, "The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo," *Genome Biol.*, vol. 7, no. 5, p. R36, 2006.

[7] H. Chen, J. Chen, L. A. Muir, S. Ronquist, W. Meixner, M. Ljungman, T. Ried, S. Smale, and I. Rajapakse, "Functional organization of the human 4D Nucleome," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 112, no. 26, pp. 8002–8007, Jun 2015.

[8] J. Zhu, P. Y. Lum, J. Lamb, D. GuhaThakurta, S. W. Edwards, R. Thieringer, J. P. Berger, M. S. Wu, J. Thompson, A. B. Sachs, and E. E. Schadt, "An integrative genomics approach to the reconstruction of gene networks in segregating populations," *Cytogenet. Genome Res.*, vol. 105, no. 2-4, pp. 363–374, 2004.

[9] J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt, "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks," *Nat. Genet.*, vol. 40, no. 7, pp. 854–861, Jul 2008.

[10] X. Yang, B. Zhang, C. Molony, E. Chudin, K. Hao, J. Zhu, A. Gaedigk, C. Suver, H. Zhong, J. S. Leeder, F. P. Guengerich, S. C. Strom, E. Schuetz, T. H. Rushmore, R. G. Ulrich, J. G. Slatter, E. E. Schadt, A. Kasarskis, and P. Y. Lum, "Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver," *Genome Res.*, vol. 20, no. 8, pp. 1020–1036, Aug 2010.

[11] I. M. Wang, B. Zhang, X. Yang, J. Zhu, S. Stepaniants, C. Zhang, Q. Meng, M. Peters, Y. He, C. Ni, D. Slipetz, M. A. Crackower, H. Houshyar, C. M. Tan, E. Asante-Appiah, G. O'Neill, M. J. Luo, R. Thieringer, J. Yuan, C. S. Chiu, P. Y. Lum, J. Lamb, Y. Boie, H. A. Wilkinson, E. E. Schadt, H. Dai, and C. Roberts, "Systems analysis of eleven rodent disease models reveals an inflammatome signature and key drivers," *Mol. Syst. Biol.*, vol. 8, p. 594, 2012.

[12] D. Y. Kenett, M. Tumminello, A. Madi, G. Gur-Gershgoren, R. N. Mantegna, and E. Ben-Jacob, "Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market," *PLoS ONE*, vol. 5, no. 12, p. e15032, 2010.

[13] A. Madi, D. Y. Kenett, S. Bransburg-Zabary, Y. Merbl, F. J. Quintana, S. Boccaletti, A. I. Tauber, I. R. Cohen, and E. Ben-Jacob, "Analyses of antigen dependency networks unveil immune system reorganization between birth and adulthood," *Chaos*, vol. 21, no. 1, p. 016109, Mar 2011.

[14] Y. N. Kenett, D. Y. Kenett, E. Ben-Jacob, and M. Faust, "Global and local features of semantic networks: evidence from the Hebrew mental lexicon," *PLoS ONE*, vol. 6, no. 8, p. e23912, 2011.

[15] M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, A. Lasorella, K. Aldape, A. Califano, and A. Iavarone, "The transcriptional network for mesenchymal transformation of brain tumours," *Nature*, vol. 463, no. 7279, pp. 318–325, Jan 2010.

[16] B. Zhang and J. Zhu, "Identification of key causal regulators in gene networks," in *(Proceedings of the World Congress on Engineering 2013)*, London, U.K, july 2013.

[17] B. Zhang, C. Gaiteri, L. G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, E. Fluder, B. Clurman, S. Melquist, M. Narayanan, C. Suver, H. Shah, M. Mahajan, T. Gillis, J. Mysore, M. E. MacDonald, J. R. Lamb, D. A. Bennett, C. Molony, D. J. Stone, V. Gudnason, A. J. Myers, E. E. Schadt, H. Neumann, J. Zhu, and V. Emilsson, "Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease," *Cell*, vol. 153, no. 3, pp. 707–720, Apr 2013.

[18] J. Paloneva, M. Kestila, J. Wu, A. Salminen, T. Bohling, V. Ruotsalainen, P. Hakola, A. B. Bakker, J. H. Phillips, P. Pekkarinen, L. L. Lanier,

T. Timonen, and L. Peltonen, "Loss-of-function mutations in TYROBP (DAP12) result in a presenile dementia with bone cysts," *Nat. Genet.*, vol. 25, no. 3, pp. 357–361, Jul 2000.

[19] Y. Zhao, L. Zhou, B. Liu, Y. Deng, Y. Wang, Y. Wang, W. Huang, W. Yuan, Z. Wang, C. Zhu, M. Liu, X. Wu, and Y. Li, "ZNF325, a novel human zinc finger protein with a RBaK-like RB-binding domain, inhibits AP-1- and SRE-mediated transcriptional activity," *Biochem. Biophys. Res. Commun.*, vol. 346, no. 4, pp. 1191–1199, Aug 2006.

[20] D. J. G. Paramita Das, "Treating Pediatric Brain Tumors by Inhibiting the RAS-ERK Signaling Pathway: A Review." *Journal of Pediatric Oncology*, vol. 3, no. 1, 2015.

[21] M. L. G. R. G. Mathivanan Jothi, Rohini Keshava, "Possible Role of The Tumor Suppressor Gene Retinoblastoma (Rb1) In Human Brain Tumor Development," *Annals of Neurosciences*, vol. 14, no. 3, 2007.

[22] J. Mathivanan, K. Rohini, M. L. Gope, B. Anandh, and R. Gope, "Altered structure and deregulated expression of the tumor suppressor gene retinoblastoma (RB1) in human brain tumors," *Mol. Cell. Biochem.*, vol. 302, no. 1-2, pp. 67–77, Aug 2007.

[23] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, Nov 2003.

[24] V. M. Aris, M. J. Cody, J. Cheng, J. J. Dermody, P. Soteropoulos, M. Recce, and P. P. Tolias, "Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer," *BMC Bioinformatics*, vol. 5, p. 185, Nov 2004.

[25] A. Zeisel, A. Amir, W. J. Kostler, and E. Domany, "Intensity dependent estimation of noise in microarrays improves detection of differentially expressed genes," *BMC Bioinformatics*, vol. 11, p. 400, 2010.

V. Appendix

TABLE I
SIMULATION RESULT FOR 20 NODES TREE WITHOUT NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.8775	0.9332	0.9747	0.7916	0.9306
0.2	0.9961	0.9963	0.9985	0.8034	1
0.4	1	1	1	0.8245	1
0.6	1	1	1	0.7975	1
0.8	1	1	1	0.8015	1

TABLE II
SIMULATION RESULT FOR 20 NODES TREE WITH 1/3 NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.7261	0.8864	0.8369	0.7812	0.8269
0.2	0.9166	0.9836	0.9877	0.7940	0.9286
0.4	1	1	1	0.8249	1
0.6	1	1	1	0.7845	1
0.8	1	1	0.9996	0.8387	1

TABLE III
SIMULATION RESULT FOR 20 NODES TREE 2/3 NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.6364	0.5499	0.5748	0.5848	0.7500
0.2	0.7797	0.8680	0.9146	0.7735	0.8462
0.4	0.9825	0.9905	0.9988	0.8126	1
0.6	0.9977	1	0.9994	0.8465	0.9000
0.8	0.8804	0.9920	0.9911	0.8146	1

TABLE IV
SIMULATION RESULT FOR 50 NODES TREE WITHOUT NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.7643	0.8991	0.9225	0.8562	0.8646
0.2	0.9988	0.9997	0.9999	0.8352	0.9762
0.4	1	1	1	0.8448	0.9286
0.6	1	1	1	0.8483	0.9902
0.8	1	1	1	0.8470	1

TABLE V
SIMULATION RESULT FOR 50 NODES TREE WITH 1/3 NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.7018	0.7831	0.8208	0.8151	0.7561
0.2	0.9617	0.9936	0.9985	0.8409	0.9748
0.4	1	0.9999	1	0.8738	0.9688
0.6	1	1	1	0.9032	1
0.8	1	0.9994	0.9998	0.9300	1

TABLE VI
SIMULATION RESULT FOR 50 NODES TREE 2/3 NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.6266	0.5486	0.6385	0.6712	0.7561
0.2	0.6196	0.7746	0.8675	0.8139	0.9625
0.4	0.9893	0.9967	0.9991	0.8673	0.8600
0.6	0.9948	0.9982	0.9982	0.8828	0.9697
0.8	0.9286	0.9943	0.9942	0.9043	1

TABLE VII
SIMULATION RESULT FOR 100 NODES TREE WITHOUT NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.7445	0.8674	0.9388	0.8394	0.9804
0.2	0.9976	0.9995	1	0.8632	0.9231
0.4	1	1	1	0.8676	0.9792
0.6	1	1	1	0.8872	1
0.8	1	1	0.8426	0.9018	1

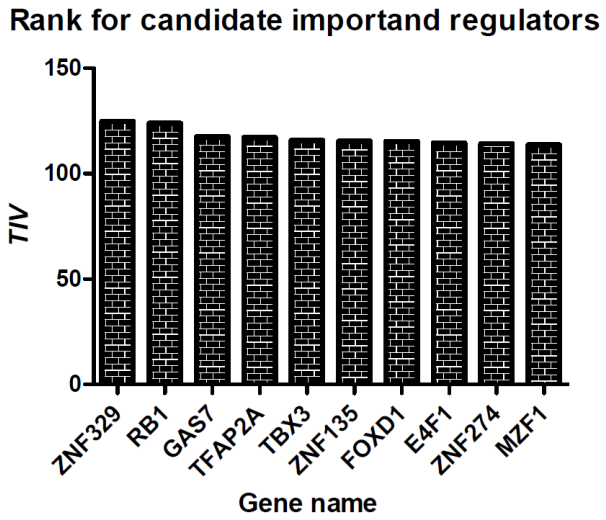

Fig. 6. The top ten genes with the largest TIV values for brain tumors.

TABLE VIII
SIMULATION RESULT FOR 100 NODES TREE WITH 1/3 NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.6929	0.7572	0.8303	0.7765	0.8333
0.2	0.9561	0.9915	0.9992	0.8615	0.9894
0.4	1	1	1	0.8745	0.9875
0.6	1	1	1	0.9071	0.9905
0.8	1	0.9992	1	0.9511	0.9965

TABLE IX
SIMULATION RESULT FOR 100 NODES TREE 2/3 NOISE

Covariance	ARACNE	CLR	GENIE3	TIGRESS	CBDN
0.1	0.4874	0.6362	0.6480	0.6547	0.9756
0.2	0.7527	0.8294	0.8867	0.8169	0.9794
0.4	0.9737	0.9871	0.9976	0.8843	0.9938
0.6	0.9990	0.9996	0.9998	0.9237	0.9907
0.8	0.9520	0.9973	0.9979	0.9123	0.9965