Reconstructing directed gene regulatory network by only gene expression data

Zhang, Lu; Feng, Xi Kang; Ng, Yen Kaow; Li, Shuai Cheng

Published in:
BMC Genomics

Published: 18/08/2016

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1186/s12864-016-2791-2

Publication details:
Zhang, L., Feng, X. K., Ng, Y. K., & Li, S. C. (2016). Reconstructing directed gene regulatory network by only gene expression data. BMC Genomics, 17, [430]. https://doi.org/10.1186/s12864-016-2791-2

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher’s copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
Abstract

Background: Accurately identifying gene regulatory network is an important task in understanding in vivo biological activities. The inference of such networks is often accomplished through the use of gene expression data. Many methods have been developed to evaluate gene expression dependencies between transcription factor and its target genes, and some methods also eliminate transitive interactions. The regulatory (or edge) direction is undetermined if the target gene is also a transcription factor. Some methods predict the regulatory directions in the gene regulatory networks by locating the eQTL single nucleotide polymorphism, or by observing the gene expression changes when knocking out/down the candidate transcript factors; regrettably, these additional data are usually unavailable, especially for the samples deriving from human tissues.

Results: In this study, we propose the Context Based Dependency Network (CBDN), a method that is able to infer gene regulatory networks with the regulatory directions from gene expression data only. To determine the regulatory direction, CBDN computes the influence of source to target by evaluating the magnitude changes of expression dependencies between the target gene and the others with conditioning on the source gene. CBDN extends the data processing inequality by involving the dependency direction to distinguish between direct and transitive relationship between genes. We also define two types of important regulators which can influence a majority of the genes in the network directly or indirectly. CBDN can detect both of these two types of important regulators by averaging the influence functions of candidate regulator to the other genes. In our experiments with simulated and real data, even with the regulatory direction taken into account, CBDN outperforms the state-of-the-art approaches for inferring gene regulatory network. CBDN identifies the important regulators in the predicted network: 1. TYROBP influences a batch of genes that are related to Alzheimer’s disease; 2. ZNF329 and RB1 significantly regulate those ‘mesenchymal’ gene expression signature genes for brain tumors.

Conclusion: By merely leveraging gene expression data, CBDN can efficiently infer the existence of gene-gene interactions as well as their regulatory directions. The constructed networks are helpful in the identification of important regulators for complex diseases.

Keywords: Gene regulatory network, Regulatory direction, Important regulators, Gene expression
**Background**

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 unprecedented 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 phenomenon, 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 to consider background noise, which naturally accounts for the transitive effects. The method works on the fact that each gene's MI 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).

Another kinds of 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 assessing their co-expressions and the co-occurrence 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 from cMonkey 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 unable to infer regulatory directions.

There have been many attempts at the inference of regulatory directions by introducing external data. The regulatory direction may be determined 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. The genes with cis-eSNPs can be the parent of the genes without cis-eSNPs; 2. The genes without cis-eSNPs cannot be the parent of the genes with cis-eSNPs. These strategies have been very successful [9–11]. However, their applicability is limited by the availability of both SNP and gene 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 asymmetric, so the node with larger influence to the other one 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 gene that influences most of the gene expression signature (GES) genes (e.g. differentially expressed genes) in the network. Carro et al. [15] identified C/EBPβ and STAT3 as important regulators for brain tumor by calculating the overlap between the TF’s targets and ‘mesenchymal’ GES genes based on Fisher’s exact test. TFs were ranked by the number of overlap genes to avoid the influence of the different size of their targets. However, this study only considers the direct influence (Fig. 1(a)) of transcription factors to their target genes, the indirect influence (Fig. 1(b)), through transitive genes, are 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. KDA has been extended to search indirect nodes that are influenced by those regulators, but the influence function is qualitative. It ignores 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 is used to remove transitive interactions in the partial correlation network. Thus each edge predicted by CBDN is both causal and directed, which can be further applied to infer the important regulators quantitatively. The performance of CBDN is compared to a few well-known algorithms, namely ARACNE, CLR, TIGRESS and GENIE3. In the simulation study, CBDN's result is comparable to the best result of these methods in each situation and proves its outstanding ability to predict regulatory direction. For a realistic test, we point out the TYROBP-oriented network which is related to Alzheimer’s disease [17]. In this test, CBDN is superior to other methods in inferring both network structure and regulatory direction. For another real expression data from the patients affected by human brain tumors, CBDN predicts two potential important regulators $ZNF329$ and $RB1$ whose function are associated with brain tumors. All of these results demonstrate the strength of CBDN in the inference of directed GRNs from gene expression data as well as its potential in predicting important regulators.

Result
CBDN is designed to construct directed regulatory network by only gene expression data. The computation of CBDN consists of three stages: In the first stage, the influence of each gene to the others is calculated to determine the edge direction. This is done through a partial correlation network constructed from the gene expression data; In the second stage, the transitive interactions are removed by DDPI; In the third stage, the important regulators are inferred by ranking the regulators based on their total influences to the GES genes.

Determine the edge direction
CBDN infers the regulatory interaction through the influence function. The influence function of gene $A$ to gene $B$ (denoted as $D(A \rightarrow B)$) is calculated by averaging the Pearson correlation changes between gene $B$ and the other genes in the network, with or without gene $A$.

$$D(A \rightarrow B) = \frac{|\text{Corr}(X_i, X_j)|}{9}$$

Here we give a schematic example based on the simulated GRN structure in Fig. 2(a) to interpret how CBDN determines the edge directionality.

Here, we denote the variable of node $i$ as $X_i$. For instance, the direction between $X_1$ and $X_4$ is determined by comparing $D(X_1 \rightarrow X_4)$ and $D(X_4 \rightarrow X_1)$. $X_4$ merely affects the correlation between $X_1$ and $X_{10}$ (see Methods),

$$D(X_4 \rightarrow X_1) = \frac{|\text{Corr}(X_1, X_{10})|}{9}$$
other variables are not influenced given \(X_4\). When conditioning on \(X_1\), the influences are extended to seven variables \((X_2, X_3, X_5, X_6, X_7, X_8 \text{ and } X_9)\),

\[
D(X_1 \rightarrow X_4) = \frac{\sum_{i=2}^{3,5,6,7,8,9} |\text{Corr}(X_1, X_i)|}{9}
\]  

The upper bound of \(D(X_4 \rightarrow X_1) \leq 1\) is smaller than \(D(X_1 \rightarrow X_4) \leq 7\) in general, so CBDN concludes that \(D(X_4 \rightarrow X_1) \leq D(X_1 \rightarrow X_4)\). The edge direction is from \(X_1\) to \(X_4\).

**Directed data processing inequality**

The influence function described 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. 3 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 \((X_i \rightarrow X_k)\). Data processing inequality (DPI) can be used to remove transitive interactions by assuming the post-processing cannot increase the mutual information. If \(X_i, X_j\) and \(X_k\) form a Markov chain, denoted as \(X_i \rightarrow X_j \rightarrow X_k\),

\[
MI(X_i; X_k) \leq MI(X_i; X_j)
\]  

**Fig. 2** The simulated gene regulatory network structures and edge directions with 10 (a), 20 (b), 50 (c) and 100 (d) nodes

\(D(X_1 \rightarrow X_4)\) = \(\frac{\sum_{i=2}^{3,5,6,7,8,9} |\text{Corr}(X_1, X_i)|}{9}\)
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 triplet of genes. For all the variables in $U$, the influence functions for $X_1$ ($D_1(X_1 \rightarrow X_6)$) and $X_2$ ($D_2(X_2 \rightarrow X_6)$) are

$$D_1(X_1 \rightarrow X_6) = \frac{\sum_{i=3}^{8} \left| \text{Corr}(X_i, X_6) \right|}{6}$$

$$D_2(X_2 \rightarrow X_6) = \frac{\sum_{i=3}^{8} \left| \text{Corr}(X_i, X_6) \right|}{6}$$

For all the variables in $V$, the influence functions for $X_1$ ($D_2(X_1 \rightarrow X_6)$) and $X_2$ ($D_2(X_2 \rightarrow X_6)$) are

$$D_2(X_1 \rightarrow X_6) = 0$$

$$D_2(X_2 \rightarrow X_6) = \frac{\sum_{i=5}^{7} \left| \text{Corr}(X_i, X_6) \right|}{2}$$

Then we have

$$D_1(X_2 \rightarrow X_6) > D_1(X_1 \rightarrow X_6)$$

$$D_2(X_2 \rightarrow X_6) > D_2(X_1 \rightarrow X_6)$$

$$D(X_2 \rightarrow X_6) = D_1(X_2 \rightarrow X_6) + D_2(X_2 \rightarrow X_6)$$

$$> D_1(X_1 \rightarrow X_6) + D_2(X_1 \rightarrow X_6)$$

$$= D(X_1 \rightarrow X_6)$$

$X_2$ is prefer to be the direct parent of $X_6$ instead of $X_1$ according to Eq. 7. Thus the regulatory structure in Fig. 2(a) should be $X_2 \rightarrow X_6$ rather than $X_1 \rightarrow X_6$.

To account for the influence of noise, we introduce a tolerance parameter $\tau$. 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 $\tau$ implies much more noise exists in the expression data to influence $D(X_i \rightarrow X_k)$ and $D(X_j \rightarrow X_k)$.

**Determine the important regulators**

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, $X_1$ has the largest influence on the other genes in the network and is located on the top level (Methods).

**Simulation**

**Tree structure simulation**

In order to explicitly reflect the nature of directed interactions in the gene regulatory network, we simulate a tree structure in which each node has only one parent (except the root) and is merely regulated by its parent (only one arrow from its parent, shown in Fig. 2). 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 descendents follows bivariate Gaussian distribution with specified covariance and noise. In addition, we mix uniform distributed noise weighted by $\frac{\kappa}{\tau}$ to the simulated expression profiles, where “$\omega$” presents the amount of noise and “$\kappa$” denotes the noise level. We set “$\omega$” to a constant ($\omega = 3$) and change “$\kappa$” from 0 to 2 in the simulations. The expression profiles of 10, 20, 50, 100 nodes are simulated, each of them derived from 1000 samples. The network structure and edge direction are shown in Fig. 2.

**Infer edge direction**

Based on the partial correlation network, CBDN can predict the interaction edge direction by only gene expression data. In the simulation, we calculate the proportion of edges that are assigned the directions correctly to evaluate the CBDN’s performance. Our simulation results demonstrate excellent performance of CBDN in predicting edge direction (Fig. 4). There are 83.3% of the simulations (66/72) where at least 60% of the edges are correctly assigned directions. As the covariance between these nodes increased, the predicted accuracy increases, and reaches optimality when the covariance is above 0.4. The influence of noise is more severe for the networks with small number of nodes (Fig. 4(a), (b) and (f)).
low covariance makes the performance in large networks declined dramatically (Fig. 4(a) and (b)).

**Compare CBDN with other methods**

We evaluate the overall performance of CBDN (including predicted edges and their directions) by comparing it with other famous methods based on a variety of simulated datasets. The true positive rate (TPR) and false positive rate (FPR) are used to plot the receiver operating characteristics (ROC) curve, where $TPR = \frac{TP}{TP + FN}$, $FPR = \frac{FP}{FP + TN}$ (TP: true positive, FN: false negative, FP: false positive). The area under ROC curve (AUC) was applied to evaluate the performance of CBDN. We apply the same tests on four state-of-the-art approaches (ARACNE, CLR, GENIE3 and TIGRESS) for comparison. In Table 1, CBDN's result is the best when no noise exists. Even with small covariance, CBDN correctly revealed the structure and regulatory orientations (Table 1(a)). 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 1). 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 achieves sufficiently good performance in reconstructing the directed GRNs. We also simulate tree structures with 20, 50, 100 nodes, in which CBDN achieves very similar results as the network with 10 nodes simulation (See Tables 2, 3 and 4).

**Infer important regulators**

From the network structure for simulation (Fig. 2), the confirmed important regulator is node 1, which is the
Table 1 Simulation result for 10 nodes tree by comparing CBDN with other methods by AUC

| Covariance | ARACNE | CLR | GENIE3 | TIGRESS | CBDN |
|------------|--------|-----|--------|---------|------|
| (a) Simulation without any noise |          |     |        |         |      |
| 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    |
| (b) Simulation with 1/3 random noise |          |     |        |         |      |
| 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    | 1      | 0.9985 | 0.8338 |
| (c) Simulation with 2/3 random noise |          |     |        |         |      |
| 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 |

Real data

For this test, we download the processed expression data from GEO [18] (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 to background noise. We impute these missing values with the average positive expression values across all the samples of 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 [19]. 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 structure and direction of the regulatory network (Fig. 6). From the TYROBP regulatory network, we choose 47 GES genes, the expressions of which are altered when TYROBP is 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 the input for ARACNE, CLR, GENIE3, TIGRESS, and CBDN. The results are compared with the true network structure and edge directions from mouse embryonic stem cells experiment. Figure 7 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 assume all the genes parent of all the other nodes in the network. Here, we calculate the proportion of those nodes in the network, whose total influence value \( TIV \) (Methods) is smaller than the \( TIV \) for node 1, to evaluate the inference ability of CBDN. From Fig. 5(a) and (b), we see that smaller networks are in general inferred more accurately, while the effects of noise is unpredictable. For example, for the 50 nodes network in Fig. 5(a), the case with 2/3 noise applied is better predicted than the cases with smaller noise. The important regulator prediction is unstable and unbelievable in the network with weak correlation. The proportion tends to one when the covariance is larger than 0.6 and the nodes in the network are larger than 20 (Fig. 5(d), (e) and (f)), which suggest that the inference is quite reliable for above medium covariance.

Table 2 Simulation result for 20 nodes tree by comparing CBDN with other methods by AUC

| Covariance | ARACNE | CLR | GENIE3 | TIGRESS | CBDN |
|------------|--------|-----|--------|---------|------|
| (a) Simulation without any noise |          |     |        |         |      |
| 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    |
| (b) Simulation with 1/3 random noise |          |     |        |         |      |
| 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    |
| (c) Simulation with 2/3 random noise |          |     |        |         |      |
| 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 3 Simulation result for 50 nodes tree by comparing CBDN with other methods by AUC

| Covariance   | ARACNE  | CLR      | GENIE3   | TIGRESS  | CBDN   |
|--------------|---------|----------|----------|----------|--------|
| (a) Simulation without any noise |         |          |          |          |        |
| 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   | 0.9092 |
| 0.8          | 1       | 1        | 1        | 0.8470   | 1      |
| (b) Simulation with 1/3 random noise |         |          |          |          |        |
| 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       | 1        | 0.8738   | 0.9688   |        |
| 0.6          | 1       | 1        | 0.9032   | 1        |        |
| 0.8          | 1       | 0.9994   | 0.9998   | 0.9300   | 1      |
| (c) Simulation with 2/3 random noise |         |          |          |          |        |
| 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 4 Simulation result for 100 nodes tree by comparing CBDN with other methods by AUC

| Covariance   | ARACNE  | CLR      | GENIE3   | TIGRESS  | CBDN   |
|--------------|---------|----------|----------|----------|--------|
| (a) Simulation without any noise |         |          |          |          |        |
| 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      |
| (b) Simulation with 1/3 random noise |         |          |          |          |        |
| 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 |
| (c) Simulation with 2/3 random noise |         |          |          |          |        |
| 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 |

to be potential regulators and ranked them based on TIV. If one gene is assessed as a regulators, other genes are assumed to be GES genes. Figure 8 lists the top 10 genes with the largest TIV, only the values of TYROBP and SLC7A7 are above 8, 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 download the expression data for brain tumors (GSE19114) and pre-process them as for Alzheimer’s disease. Eventually, we choose 132 ‘mesenchymal’ gene expression signature (MGES) genes and 883 TFs from Supplementary Tables 1 and 2 from the original paper [15]. Both MGES genes and TFs are combined together to calculate TIV for each TFs, because we are also required to consider the regulatory relationships between TFs. We are unable to identify the two key regulators (STAT3 and C/EBPβ) described in the original papers from the top TIV ranked TFs (Fig. 9), 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 family has been proved to be associated with brain tumor. Zhao et al. [20] identified ZNF325 as a transcription repressor in MAPK/ERK signaling pathway. Recently, Das et al. [21] made a comprehensive review to clarify the relationship between MAPK/ERK signaling pathway and brain tumors and how can one 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 [22]. Mathivanan et al. found loss of heterozygosity and deregulated expression of RB1 in human brain tumors [23].

Discussion
In this paper, we propose a new computational method called Context Based Dependency Network (CBDN), which constructs directed GRNs from only gene expression data. This provides 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. Although gene expression data can reflect the
gene-gene interactions in GRN, there are still three limitations that must be addressed. First, the transcription factors prefer to act together as a protein complex rather than individually. The protein complex may be blocked or inactivated, for reasons such as incorrect folding, being restricted in the nucleus or 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-dependent. Because the time delay exists between transcription and translation, high mRNA expression level does not imply a simultaneous high in protein abundance. Third, even when TFs are bound to their target genes, they may demonstrate different effects because of their three dimensional distances and histone modification.

The probes with low fluorescence signals are impossible to be distinguished from background noise. CBDN treats 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 [24], the reconstructed GRN seems stable and consistence. In the future, some noise filtering methods should be incorporated in CBDN such as described in [25, 26].

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 edge direction. For CBDN, both of the interactions and directions are taken into consideration for evaluating its performance. Even though only 2% of AUC is improved in TYROBP oriented GRN inference, the result is more powerful and useful since they incorporate edge directions. The performance of CBDN is significantly better

Fig. 6 The network structure for the TYROBP oriented regulatory network for Alzheimer’s disease

Fig. 7 The performance of different methods for predicting TYROBP oriented regulatory network

Fig. 8 The top ten genes with the largest TIV values for Alzheimer’s disease
than other methods in some situations such as Table 1(c) 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 gene expression data alone, as it identifies 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 may be neglected by the previous approaches.

This paper also contributes 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 bidirected interactions which are quite common in nature. By incorporating external data, we hope to use it to tackle the situations where more than one TFs co-regulate a gene simultaneously.

**Conclusion**

The reconstruction of gene regulatory network has been actively researched in the past decade, many methods have been designed to achieve this using only high-throughput gene expression data. However, the edge direction is usually unknown and seems hard to be determined by only gene expression data. Even when the directions can be affirmed, the available approaches is unable to remove transitive interactions from directed network. Here, we propose a novel method CBDN, which can reconstruct direct gene regulatory network by only gene expression data. CBDN first constructs an asymmetric partial correlation network to determine the two influence functions for each pair of genes and determine the edge direction between them. DDPI extends data processing inequality applied in directed network to remove transitive interactions. By aggregating the influence function to all the nodes in the network, the total influence value is calculated to assess whether the node is an important regulator. For both simulation and real data test, CBDN demonstrated superior performance compared to other available methods in reconstructing directed gene regulatory network. It also successfully identified the important regulators for Alzheimer’s disease and brain tumors.

**Methods**

**Partial correlation network**

In CBDN, a partial correlation network 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, $E = (e_{ij})_{m \times n}$, 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

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

(8)

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 $PC(X_i, X_k | X_j)$,

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

(9)

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 $X_k$,

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

(10)

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.

**Proof for directed data processing inequality**

In the directed GRN, we assume three genes ($X_i$, $X_j$ and $X_k$) form a Markov chain ($X_i \rightarrow X_j \rightarrow X_k$), the other genes are separated into two categories: non-descendants...
of $X_i$ ($U = \{X_m \cdots X_n\}$) and descendents of $X_i$ ($V = \{X_p \cdots X_a\}$). For the elements in $U$,

$$D_1(X_i \rightarrow X_k) = \frac{1}{|U|} \sum_{t \neq i} |d(X_k, X_t | X_i)|$$  \hspace{1cm} (11)

$$D_1(X_j \rightarrow X_k) = \frac{1}{|U|} \sum_{t \neq j} |d(X_k, X_t | X_j)|$$  \hspace{1cm} (12)

Based on Eq. 9, $X_k$ is conditionally independent with other genes in $U$ given $X_i$ or $X_j$, thus we have $PC(X_k,X_i | X_j) = 0, |d(X_k, X_t | X_i)| = |d(X_k, X_t | X_j)| \forall t \in U$. 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 $V$

$$D_2(X_i \rightarrow X_k) = \frac{1}{|V|} \sum_{j \neq i} |d(X_k, X_t | X_i)|$$  \hspace{1cm} (13)

$$D_2(X_j \rightarrow X_k) = \frac{1}{|V|} \sum_{j \neq j} |d(X_k, X_t | X_j)|$$  \hspace{1cm} (14)

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

$$D(X_i \rightarrow X_k) = D_1(X_i \rightarrow X_k) + D_2(X_i \rightarrow X_k)$$

$$\leq D_1(X_j \rightarrow X_k) + D_2(X_j \rightarrow X_k)$$  \hspace{1cm} (15)

$$= D(X_j \rightarrow X_k)$$

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. differentially expressed genes) are $X_{t1}, X_{t2}, \ldots, X_{tm}$, the total influence value ($TIV$) of gene $X_i$ is $TIV(X_i) = \sum_{t=1}^{m} D(X_i \rightarrow X_{st})$. Regulators are ranked by their $TIV$s.

Declarations

Publication of this article was funded by RGF Grant No. 9041901 (CityU 118413). This article has been published as part of BMC Genomics Volume 17 Suppl 4 2016: Selected articles from the IEEE International Conference on Bioinformatics and Biomedicine 2015: genomics. The full contents of the supplement are available online at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4885317&tool=pmcentrez&rendertype=abstract.

Authors' contributions

SL supervised the work and together with LZ, developed CBDN and procedure of experiment. LZ implemented the CBDN method in matlab. LZ, XK did the experiments on simulation and real data. LZ, SL and YK wrote the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

1Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. 2Faculty of Information and Communication Technology, University Tunku Abdul Rahman, Kampar, Perak, Malaysia.

Published: 18 August 2016

References

1. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Fava R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7 Suppl 1:7.

2. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):18.

3. Haury AC, Mordelet F, Vera-Licona P, Vert JP. TIGRESS: trustful inference of gene regulation using stability selection. BMC Syst Biol. 2012;6:145.

4. Huynh-Thu VA, Iritani A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS ONE. 2010;5(9):12776;1–10.

5. Reiss DJ, Baliga NS. Bonneau R. Integrated clustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics. 2006;7:280.

6. Bonneau R, Reiss DJ, Sharrow P, Facciotti M, Hood L, Baliga NS, Thorsson V. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol. 2006;7(5):36.

7. Chen H, Chen J, Muir LA, Ronquist S, Meixner W, Ljungman M, Ried T, Smale S, Rajapakse I. Functional organization of the human 4D Nucleome. Proc Natl Acad Sci U S A. 2015;112(26):8002–7.

8. Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, Berger JP, Wu MS, Thompson J, Sachs AB, Schadt EE. An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet Genome Res. 2004;105(1-2):363–74.

9. Zhu J, Zhang B, Smith EN, Drees B, BREM RB, Kruglyak L, Bumgarner RE, Schadt EE. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet. 2006;38(7):854–61.

10. Yang X, Zhang B, Molony C, Chadin E, Hako K, Zhu I, Gaedigk A, Suver C, Zhong H, Leeder JS, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich RG, Slater JG, Schadt EE, Kasarskis A, Lum PY. Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. Genome Res. 2010;20(8):1202–36.

11. Wang IM, Zhang B, Yang X, Zhu J, Stephanantis S, Zhang C, Meng Q, Peters M, He Y, Ni C, Slizpet D, Crackower MA, Houyhahar T, Tan CM, Asante-Appiah E, O'Neill L, Loo MJ, Thieringer R, Yuan J, Chu LS, Lum PY, Lamb J, Boie Y, Wilkinson HA, Schadt EE, DAI H, Roberts C. Systems analysis of eleven rodent disease models reveals an inflammatory signature and key drivers. Mol Syst Biol. 2012;8:594.

12. Kenett DY, Tumminoelli M, Madi A, Guer-Gershgroen G, Mantegna RN, Ben-Jacob E. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. PLoS ONE. 2010;5(12):15032.

13. Madi A, Kenett DY, Bransburg-Zabary S, Merbl Y, Quintana FJ, Boccaletti S, Tauber AI, Cohen IR, Ben-Jacob E. Analyses of antigen dependency networks unveil immune system reorganization between birth and adulthood. Chaos. 2011;21(1):016109.

14. Kenett YN, Kenett DY, Ben-Jacob E, Faust M. Global and local features of semantic networks: evidence from the Hebrew mental lexicon. PLoS ONE. 2011;6(8):23912.

15. Carro MS, Lim WK, Alvarez-MJ, Bolivar MJ, Zhao X, Snyder EY, Kulman EP, Anne SL, Doetsch F, Colman H, Lasorella A, Aldape K, Califano A, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7 Suppl 1:7.

16. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee-J, Podtelezhnikov AA, Zhang C, Xie T, Tran L, Dobrin R, Fluder E, Clurman B, Melquist S, Narayanan M, Suver C, Shah H, Mahajan M, Gillis T, Myers J, MacDonald ME, Lamb JR, Bennett DA, Molony C, Stone DJ, Gudnason V, Myers AJ, Schadt EE, Neuman H, Zhu J, Esmisson V. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell. 2013;153(3):707–20.

17. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillips KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO.
19. Paloneva J, Kehtila M, Wu J, Salminen A, Bohling T, Ruotsalainen V, Hakola P, Bakker AB, Phillips JH, Pekkarinen P, Lanier LL, Timonen T, Peltonen L. Loss-of-function mutations in TYROBP (DAP12) result in a presenile dementia with bone cysts. Nat Genet. 2000;25(3):357–61.

20. Zhao Y, Zhou L, Liu B, Deng Y, Wang Y, Wang Y, Huang W, Yuan W, Wang Z, Zhu C, Liu M, Wu X, Li Y. 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. 2006;346(4):1191–1199.

21. Paramita Das DJG. Treating Pediatric Brain Tumors by Inhibiting the RAS-ERK Signaling Pathway: A Review. J Pediatric Oncol. 2015;3(1):1–7.

22. Mathivanan J, K R, Gope ML, Gope R. Possible role of the tumor suppressor gene retinoblastoma (rb1) in human brain tumor development. Ann Neurosci. 2007;14(3):72–82.

23. Mathivanan J, Rohini K, Gope ML, Anandh B, Gope R. Altered structure and deregulated expression of the tumor suppressor gene retinoblastoma (RB1) in human brain tumors. Mol Cell Biochem. 2007;302(1-2):67–77.

24. Oba S, Sato MA, Takemasa I, Monden M, Matsubara K, Ishii S. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics. 2003;19(16):2088–96.

25. Aris VM, Cody MJ, Cheng J, Dermody JJ, Soteropoulos P, Recce M, Tolia PP. Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer. BMC Bioinformatics. 2004;5:185.

26. Ziesel A, Amir A, Kostler WJ, Domany E. Intensity dependent estimation of noise in microarrays improves detection of differentially expressed genes. BMC Bioinformatics. 2010;11:400.