


METHODOLOGY ARTICLE

Open Access



A scalable algorithm for structure identification of complex gene regulatory network from temporal expression data

Shupeng Gui¹, Andrew P. Rice², Rui Chen³, Liang Wu⁴, Ji Liu^{1,5} and Hongyu Miao^{4*} 

Abstract

Background: Gene regulatory interactions are of fundamental importance to various biological functions and processes. However, only a few previous computational studies have claimed success in revealing genome-wide regulatory landscapes from temporal gene expression data, especially for complex eukaryotes like human. Moreover, recent work suggests that these methods still suffer from the curse of dimensionality if a network size increases to 100 or higher.

Results: Here we present a novel scalable algorithm for identifying genome-wide gene regulatory network (GRN) structures, and we have verified the algorithm performances by extensive simulation studies based on the DREAM challenge benchmark data. The highlight of our method is that its superior performance does not degenerate even for a network size on the order of 10^4 , and is thus readily applicable to large-scale complex networks. Such a breakthrough is achieved by considering both prior biological knowledge and multiple topological properties (i.e., sparsity and hub gene structure) of complex networks in the regularized formulation. We also validate and illustrate the application of our algorithm in practice using the time-course gene expression data from a study on human respiratory epithelial cells in response to influenza A virus (IAV) infection, as well as the CHIP-seq data from ENCODE on transcription factor (TF) and target gene interactions. An interesting finding, owing to the proposed algorithm, is that the biggest hub structures (e.g., top ten) in the GRN all center at some transcription factors in the context of epithelial cell infection by IAV.

Conclusions: The proposed algorithm is the first scalable method for large complex network structure identification. The GRN structure identified by our algorithm could reveal possible biological links and help researchers to choose which gene functions to investigate in a biological event. The algorithm described in this article is implemented in MATLAB®, and the source code is freely available from <https://github.com/Hongyu-Miao/DML.git>.

Keywords: Gene regulatory network, Hub gene structure, Ultra-high dimensional problem, Decomposable multi-structure identification, Influenza infection

Background

Gene regulatory network (GRN), consisting of multiple regulators and their target molecules, plays critical roles in numerous biological processes by modulating the expression levels of RNAs and proteins [1]. While remarkable successes in dissecting single genes that are responsible for certain biological functions, behavior or diseases have

been achieved over the past few decades, it has been increasingly recognized that elucidating gene functions and interactions in the context of networks becomes more and more important to gain novel insight into mechanisms, effects and interventions of molecular, cellular or organ-level biological processes [2–4]. Clearly, one of the prerequisites for investigators to harvest the benefits of such systematic network approaches is whether the structures of gene regulatory networks can be accurately revealed from experimental data.

*Correspondence: hongyu.miao@uth.tmc.edu

⁴Department of Biostatistics, University of Texas Health Science Center, 77030 Houston, TX, USA

Full list of author information is available at the end of the article

Modern high-throughput experimental technologies such as next generation sequencing [5] can generate time-course data at a much more affordable cost [6], thus provide unprecedented opportunities for researchers to systematically investigate the temporal patterns of gene expressions and infer gene regulatory relationships. However, two well-known major obstacles have significantly hampered our ability to interrogate such data for novel scientific findings. First, limited by resources or technical and ethic issues, the sampling frequency of time-course gene expression profiling data is low (e.g., most of the time-course GEO datasets [7] have less than 6 time points), which renders the sample size far less than the number of unknown parameters in the context of GRN structure identification. Targeting at such scenarios, it is of significant importance to borrow information from additional sources (e.g., previous biological knowledge). Second, considering the fact that for complex eukaryotes like human, the number of protein-coding genes is approximately 19,000 [8] so the problem dimension is *ultra-high* (i.e., tens of thousands or even millions unknown parameters are involved). The development of novel and more efficient algorithms that can scale to such high-dimensional networks is still necessary.

A number of modeling and computational approaches have been developed for gene network structure identification [9], including information theory method (e.g. [10]), clustering method (e.g. [11]), Boolean network [12], Bayesian network (e.g. [13]), state-space model [14], regression model [15], and differential equation model (e.g. [16]). The well-known methods for static expression data include the Pearson correlation coefficient (PCC) and the mutual information (MI) methods [17, 18], the ARACNE algorithm [19] based on the data processing inequality concept [20], the context likelihood of relatedness (CLR) approach [21] that extends the relevance networks method [22], the GENIE3 algorithm [23] that uses tree-based gene ranking and ensemble, and the TIGRESS algorithm [24] that combines the least angle regression with stability selection. However, only a few algorithms were specifically developed for time-course data. Zoppoli et al. [25] proposed to calculate the time-delayed dependencies between gene expressions using mutual information, and developed and tested the TimeDelay-ARACNE algorithm on networks with less than 30 genes. Yang et al. [26] used the S-system (a special form of nonlinear differential equations) to model time-course gene expression data and validated their parameter estimation method using several small systems with six or less variables. Our group introduced the concept of background network to confine the parameter space using prior biological knowledge and built the SITPR pipeline based on a dynamic Bayesian network (DBN) model regularized by L^1 norm [27]. More recently, Huynh-Thu et al. [28] developed the

Jump3 method by combining the on/off stochastic differential equation model with decision trees, where the marginal likelihood of each node was used as the decision function.

The DREAM Challenge [29] makes it feasible to generate authoritative benchmark data for algorithm performance evaluation. To the best knowledge of our authors, none of the algorithms above have been evaluated on DREAM networks of a size greater than 100; actually, their performances all become unsatisfactory at a network size of 100, as verified in several previous studies [17, 18, 27]. Therefore, a breakthrough in algorithm development is necessarily needed to obtain accurate and reliable results for large-scale GRN inference (e.g., a network size $O(10^4)$). The work of Liu et al. [27] is one of the few studies that systematically borrow information from previous biological knowledge to deal with the low sampling frequency and the high dimensionality issues in GRN inference; however, even after the incorporation of prior knowledge, the number of unknown parameters is still on the order of 10^5 so the high dimensionality issue remains. Since GRNs are typical complex networks with structural sparsity [30], it is common to impose sparsity on the inference problem formulation in previous studies [15, 24, 27]; however, it turned out that sparsity alone cannot satisfyingly address the high dimensionality problem [27]. This observation leads us to the hypothesis that considering additional structural or topological properties of large-scale complex networks may result in novel scalable algorithms for GRN inference that are not subject to the curse of dimensionality.

In this study, we adopt the background network approach developed in our previous study [27] for parameter space confinement, and we modify other selected state-of-the-art algorithms to take the advantage of the same background network for fairness of algorithm performance comparison. More importantly, a breakthrough in algorithm performance is achieved by considering both structural sparsity and the existence of hub genes, which is a prominent topological feature of GRNs due to preferential gene attachment [2]. We describe the mathematical formulation and computational steps of the novel algorithm in “Methods” section. We evaluate and compare the performance of our algorithm with other state-of-the-art approaches using DREAM4 benchmark data that are generated from networks of a size up to 2×10^4 , and we apply the proposed method to real temporal gene expression data from an influenza H1N1 virus infection study to illustrate its usage in practice and further validate the proposed algorithm using a large number of known interactions. The related computational settings and results are presented in “Results and discussion” section. Finally, we discuss the advantages and limitations of this work in “Conclusions” section.

Methods

Background regulatory network

For ultra-high dimensional problems considered in this study, it is important to constrain the parameter space based on prior biological information to improve computing efficiency and avoid over-fitting, especially for datasets with limited sample sizes. For this purpose, the concept of background regulatory network was introduced in our previous work [27], and we recently also developed a curated database called RegNetwork for public use [31]. Here we present the key features of the background network and describe how it is employed in this study for GRN inference.

First, the background regulatory network is comprehensive so it contains not only experimentally-observed regulatory interactions but also physically, chemically or biologically feasible interactions. For this purpose, information from over 20 selected databases or knowledge-bases (e.g., FANTOM [32], TRANSFAC [33], JASPAR [34], KEGG [35], and HPRD [36]) have been collected and integrated. In addition, potential interactions between transcription factors (TFs) and target genes are predicted based on sequence motifs obtained from Ensembl [37]. In this way, the background network also allows the discovery of novel regulatory relationships that have not been officially reported or documented. Second, the background network is not cell type- or condition-specific but it allows the detection of cell type- or condition-specific regulatory relationships. The reason is that under different conditions (e.g., different diseases), only certain regulatory interactions in certain types of cells will be activated in the background network, and an appropriately designed algorithm should be able to detect a reasonable number of activated interactions with the presence of the inactive ones. Third, the background network is different from random networks [38] in terms of certain properties like characteristic path length and node degree distribution [39]; also, its node degrees are found to satisfy the power-law distributions so the background network is scale-free.

Here we use the background regulatory network concept in both the simulation studies and the real data example. That is, our primary task is to identify the activated interactions in the confined parameter space specified by the background network. We recognize that some existing algorithms under comparison start with the full network structure; therefore, for fairness, we make necessary efforts to modify their computer codes to take the background network as the parameter space. Also, it should be mentioned that the same weight is assigned to all the edges in the background network in this study for simplicity. That means, our algorithm will automatically determine the sparse structure and the hub gene structures without informing itself the potential location of

such structures beforehand. The algorithm performance can be further improved if edge weights can be assigned based on evidence strength in, e.g., literature; however, even with a simple equal-weight scheme, our algorithm can already achieve a satisfying performance (see the “Results” section). Finally, for simulation studies, the background network is introduced by adding additional random edges to the DREAM4 network structures; for real data example, the background network for human fetched from the RegNetwork database is used, which contains 23,079 nodes and 372,774 edges.

Model formulation of topological features of GRN

While sparsity of GRNs has been extensively explored in previous studies, the hub gene structure has rarely been addressed in existing formulations of GRN inference. This study suggests that in addition to sparsity, the hub gene structure is also of significant importance to the development of scalable algorithms that are not subject to the curse of dimensionality.

Throughout this paper, the following notations are used:

- n denotes the total number of nodes in a network;
- $\mathbf{x}_t \in \mathbb{R}^n$ denotes the expression level of n genes (and miRNAs) at time t ;
- \odot denotes the Hadamard product (i.e., the product of the entries of two matrices at the same position). For example, $[1, 3] \odot [0, 2] = [0, 6]$.
- $\|\cdot\|_{\mathcal{F}}$ is the Frobenius norm of a matrix M :

$$\|M\|_{\mathcal{F}} = \sqrt{\sum_{i,j} M_{ij}^2}$$
- $\|\cdot\|$ is the Spectral norm (i.e., the maximal singular value of a matrix);
- C_j : j th column of C ;
- C_i : i th row of C .

The dynamic Bayesian network (DBN) model [40] is considered in this study for GRN inference. It has been shown by previous studies [40, 41] that under the normality and Markov assumptions, the DBN model can be reduced to a vector autoregression (VAR) model as follows:

$$\mathbf{x}_{t+1} = P\mathbf{x}_t + \mathbf{w}_t, \quad t = 1, \dots, T-1, \quad (1)$$

where T is the total number of time points, and $P \in \mathbb{R}^{n \times n}$ is the coefficient matrix, characterizing the change from time t to the next time point $t+1$. $P_{ij} \neq 0$ indicates the existence of regulatory relationship between regulator j and target i . Also, for simplicity and robustness, uneven time intervals are treated as of equal length. We consider an equivalent form of Eq. (1):

$$\mathbf{y}_{t+1,t} := \mathbf{x}_{t+1} - \mathbf{x}_t = C\mathbf{x}_t + \mathbf{w}_t, \quad (2)$$

where $C = P - I$. Our goal is to estimate the coefficient matrix C (or equivalently P) given the time-course observations $\{\mathbf{x}_t\}_{t=1}^T$. The objective function can be formulated as follows

$$\min_C \frac{1}{2} \|Y - CX\|_{\mathcal{F}}^2 = \sum_{t=1}^{T-1} \|(\mathbf{x}_{t+1} - \mathbf{x}_t) - C\mathbf{x}_t\|^2, \quad (3)$$

where $Y := [\mathbf{x}_2 - \mathbf{x}_1, \mathbf{x}_3 - \mathbf{x}_2, \dots, \mathbf{x}_T - \mathbf{x}_{T-1}] \in \mathbb{R}^{n \times (T-1)}$ and $X := [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{T-1}] \in \mathbb{R}^{n \times (T-1)}$.

Since the total number of distinct time points are usually small, the solution obtained by directly solving (3) suffers from the overfitting issue. To overcome this problem, the commonly used strategy is to confine the model space using regularizations or constraints; for example, the ℓ_1 norm regularization for imposing model sparsity. However, in addition to sparsity, there exist other prominent topological characteristics for complex networks like GRNs [42]. To illustrate this, we use the Yeast gene network as an example. The top left graph in Fig. 1 shows a typical substructure of the Yeast gene network; the bottom left figure is the corresponding coefficient matrix C , where the blue dots indicate nonzeros (3.6%) and the white areas

indicate zeros (96.4%) in C . Sparsity is thus a clear feature to tell from Fig. 1, and we also consider the following network characteristics in our model formulation.

Hub gene structure separation: As suggested in Fig. 1, there exist a small number of gene nodes with a high degree, called the “hub” genes. If we define the *hub gene structure* as the set of outgoing edges from a hub gene node to its immediate neighbors, we can separate this structure from the original GRN such that matrix C can be decomposed into

$$C = A + B,$$

where matrix A corresponds to the hub gene structure (the upper middle graph in Fig. 1) and matrix B corresponds to the remaining edges (the upper right graph in Fig. 1).

It can be told from the lower middle plot of Fig. 1 that matrix A for hub gene structure has a column group sparse structure (that is, it only contains a few nonzero columns), while the matrix B for the remaining edges shows a more uniform sparsity pattern. Therefore, we consider the $\ell_{2,1}$

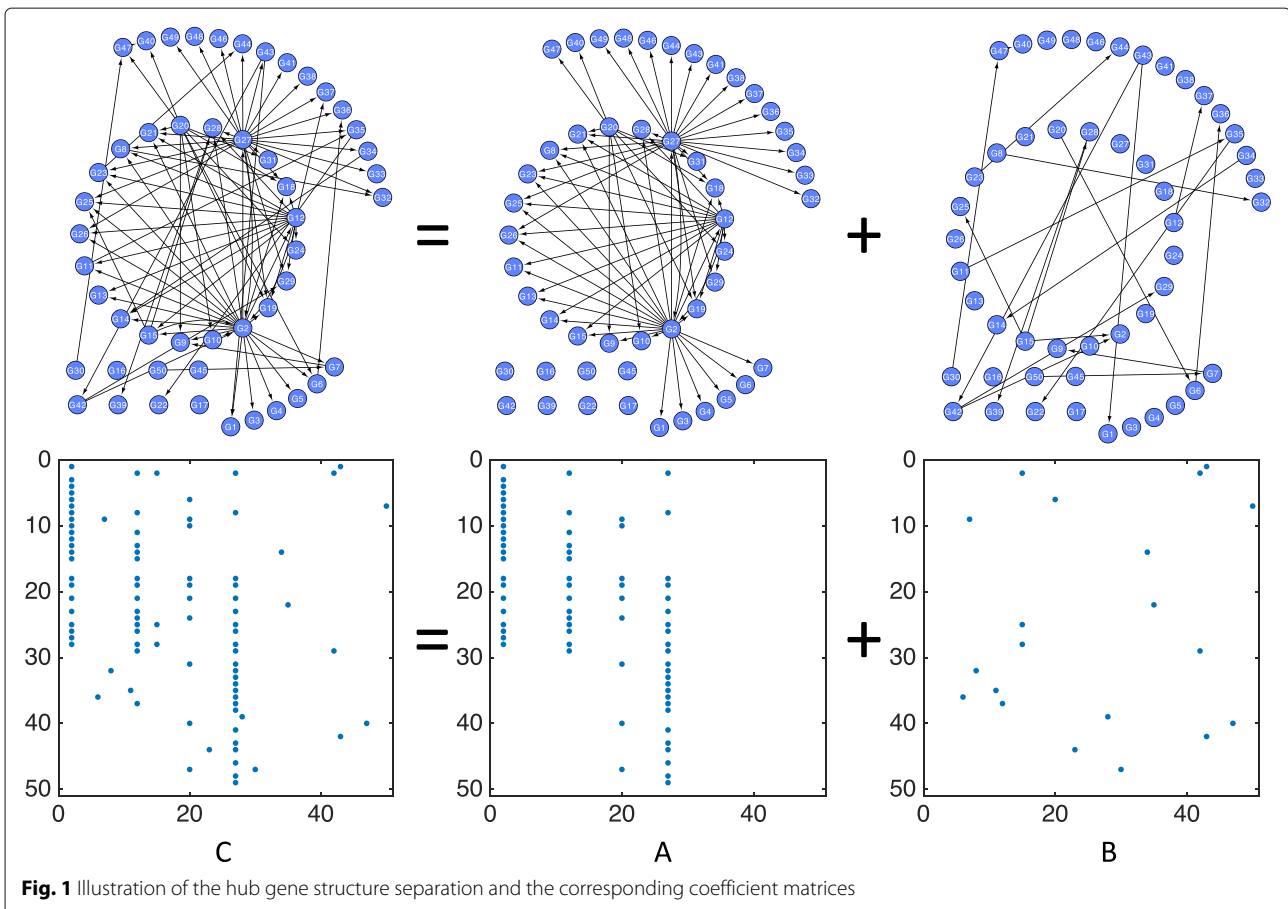


Fig. 1 Illustration of the hub gene structure separation and the corresponding coefficient matrices

norm convex regularization on A to enforce the group sparsity structure [43],

$$\|A\|_{2,1} := \sum_j \left(\sum_i |A_{ij}|^2 \right)^{\frac{1}{2}},$$

and consider the commonly used ℓ_1 regularization [44] on B to enforce the uniform sparsity,

$$\|B\|_1 := \sum_{i,j} |B_{ij}|.$$

Small number of parent nodes: Based on the real GRN structures previously reported (e.g., Yeast), we find that while one gene may directly modulate the expressions of many other genes, a single gene is usually co-regulated by only a few regulators. That is, in a GRN represented by a directed acyclic graph, the number of parent nodes of a node is often small. This observation suggests the non-zero elements in any row of matrix C should be small (see the bottom left of Fig. 1). We thus can impose the following constraint on the rows of matrix C

$$\|C_{i \cdot}\|_0 \leq \xi_i, \quad i = 1, \dots, n,$$

where ξ_i is a predefined upper bound of the number of incoming edges to gene i (e.g., 0.7 times the indegree of a node in the background network).

Background network: The *mask of the background network* can be denoted as $\Omega \in \{0, 1\}^{n \times n}$, where $\Omega_{ij} = 1$ corresponds to the non-existence of any regulatory interaction between genes i and j and $\Omega_{ij} = 0$ corresponds to the existence of a possible interaction between genes i and j that needs to be determined from temporal expression data. Therefore, the background network constraints imposed on matrix C can be denoted as follows:

$$\mathcal{P}_\Omega(C) := C \odot \Omega = 0.$$

Combining all the constraints above, the overall GRN structure identification formulation is given below:

$$\begin{aligned} \min_{A,B,C} \quad & \frac{1}{2} \|(A+B)X - Y\|_{\mathcal{F}}^2 + \alpha \|A\|_{2,1} + \beta \|B\|_1 \\ \text{s.t.} \quad & \|C_{i \cdot}\|_0 \leq \xi_i \\ & \mathcal{P}_\Omega(A) = 0, \mathcal{P}_\Omega(B) = 0 \\ & A + B = C. \end{aligned} \tag{4}$$

where α and β are two penalty coefficients that balance the weights on group sparsity and uniform sparsity.

Computing algorithm

Existing solvers cannot deal with the proposed formulation above, so we introduce an efficient optimization algorithm, called Decomposable Multi-structure Identification (DMI) (Algorithm 1), that can handle a network size greater than $20k$ on an average computer [45].

We first define the Augmented Lagrangian function:

$$\begin{aligned} L(A, B, C, U) = & \frac{1}{2} \|(A+B)X - Y\|_{\mathcal{F}}^2 + \alpha \|A\|_{2,1} + \beta \|B\|_1 \\ & + \frac{\rho}{2} \|A+B-C\|_{\mathcal{F}}^2 + \langle U, A+B-C \rangle \\ & + \mathbb{I}_{\|C_{i \cdot}\|_0 \leq \xi_i}(C) + \mathbb{I}_{\mathcal{P}_\Omega(A)=0}(A) + \mathbb{I}_{\mathcal{P}_\Omega(B)=0}(B), \end{aligned} \tag{5}$$

where ρ is a positive real number, $\mathbb{I}_{\text{condition}}(\cdot)$ denotes the function that gives 0 if the argument variable satisfies the condition and gives positive infinity otherwise, and U is a dual matrix of the same size as C introduced to solve the optimization problem.

Algorithm 1 Decomposable Multi-structure Identification (DMI)

Require: $X, Y, \xi_i, \rho > 0$

Ensure: C

repeat

$[A, B] = \arg \min_{A,B} L(A, B, C, U)$ using Algorithm 2;

$C = \arg \min_C L(A, B, C, U)$ to update C using (8);

Update U by $U \leftarrow U + \rho(A+B-C)$;

until convergence;

return C .

More specifically, we can break the optimization problem into minimization w.r.t. A and B , minimization w.r.t. C , and minimization w.r.t. U as follows.

Minimization w.r.t. A and B . Since some terms in Eq. (5) such as $\mathbb{I}_{\|C_{i \cdot}\|_0 \leq \xi_i}(C)$ contain no A or B , the optimization problem can be simplified:

$$\begin{aligned} \min_{A,B} \quad & \frac{1}{2} \|(A+B)X - Y\|_{\mathcal{F}}^2 + \alpha \|A\|_{2,1} + \beta \|B\|_1 \\ & + \frac{\rho}{2} \|A+B-C\|_{\mathcal{F}}^2 + \langle U, A+B-C \rangle \\ & + \mathbb{I}_{\mathcal{P}_\Omega(A)=0}(A) + \mathbb{I}_{\mathcal{P}_\Omega(B)=0}(B), \end{aligned} \tag{6}$$

which can be solved using the coordinate block descent method [46]. By taking partial derivatives of the objective function, we can derive

$$g = [(A+B)X - Y]X^\top + \rho(A+B-C) + U;$$

and by applying the proximal gradient descent method [47], the following update rules for A and B at each iteration can be obtained

$$\begin{aligned} A_{\cdot i} &= \text{prox}_{\gamma\alpha\|\cdot\|}((A - \gamma g_A)_{\cdot i}) \\ &= \max\left(0, 1 - \frac{\gamma\alpha}{\|(A - \gamma g_A)_{\cdot i}\|}\right)(A - \gamma g_A)_{\cdot i}, \end{aligned}$$

and

$$\begin{aligned}
 B &= \text{prox}_{\gamma\beta\|\cdot\|_1}(B - \gamma g_B) \\
 &= \text{sign}(B - \gamma g_B) \odot \max(0, |B - \gamma g_B| - \gamma\beta),
 \end{aligned}$$

where $\text{sign}(a) = 1$ if $a > 0$, $\text{sign}(a) = 0$ if $a = 0$, and otherwise -1 . Usually we choose $\gamma = \frac{1}{\|X\|^2 + \rho}$ as the safe step length for each update to guarantee the monotonic decreasing of the objective function (6). More details about the accelerated proximal gradient descent method are given in Algorithm 2.

Minimization w.r.t. C . Similar to the objective function (6), some terms in Eq. (5) like $\mathbb{I}_{\mathcal{P}_\Omega(A)=0}(A)$ do not contain matrix C . Hence, Eq. (5) can be simplified as follows:

$$\begin{aligned}
 \min_C \quad & \frac{\rho}{2} \|A + B - C\|_{\mathcal{F}}^2 + \langle U, A + B - C \rangle + \mathbb{I}_{\|C_i\|_0 \leq \xi_i}(C), \\
 & \alpha \frac{1}{2} \|A + B - C + \frac{1}{\rho} U\|_{\mathcal{F}}^2 + \mathbb{I}_{\|C_i\|_0 \leq \xi_i}(C).
 \end{aligned} \tag{7}$$

We thus update C by the following rule

$$C = \mathcal{P}_\xi(A + B + \frac{1}{\rho} U), \tag{8}$$

where \mathcal{P}_ξ is a projection of $\|C_i\|_0 \leq \xi_i$ by retaining the largest ξ_i elements in the i -th row of C .

Minimization w.r.t. U . Similar to the derivation of the objective function (7) above, we have

$$\min_U \quad \langle U, A + B - C \rangle. \tag{9}$$

We update U using $U = U + \rho(A + B - C)$ for each iteration.

Algorithmic complexity

Let $|\bar{\Omega}|$ denote the total number of zeros in the background network matrix $\Omega \in \mathbb{R}^{n \times n}$ used in Algorithms 1 and 2. Algorithm 1 is assumed to run J iterations to converge, and for each iteration of Algorithm 1, we run Algorithm 2 for K iterations. Therefore, the complexity of Algorithm 2 is $O(|\bar{\Omega}| \times T \times K)$, where $X, Y \in \mathbb{R}^{n \times (T-1)}$, $T \ll n$ and $|\bar{\Omega}| \ll n^2$. Therefore, the complexity of DMI is $O(|\bar{\Omega}| \times T \times K \times J)$. Note that matrix C is sparse, the complexity of our method mainly depends on the number of non-zero edges in the network, and thus indirectly depends on the number of nodes. On a regular PC (3.0 GHz CPU, 16 GB memory, single thread), the proposed algorithm takes ~ 1 s, $3 \sim 5$ s, $2 \sim 3$ min, and ~ 10 hr to finish 1000 iterations for a network size 10, 100, 1000, and 20,000, respectively. In most of the cases, we found that the DMI algorithm already converged after 100 iterations. See Additional file 1: Text S1 for the system requirements to run the proposed algorithm; also,

the guideline for tuning DMI computing parameters is provided in Additional file 2: Text S2.

Algorithm 2 Accelerated proximal gradient descent method [48]

Require: n (number of genes in network), γ (step length), α, β, C, U

Ensure: A, B

Initialize $t_1 = 1, V_1 = A_1 = \mathbf{0}, W_1 = B_1 = \mathbf{0}, \rho = 10$;

for $k = 1: K$ **do**

$g_k = \mathcal{P}_\Omega([(V_k + W_k)X - Y]X^\top + \rho(V_k + W_k - C) + U)$;

for $i=1$ to n **do**

$A_{k+1,i} = \text{prox}_{\gamma\alpha\|\cdot\|_1}((V_k - \gamma g_k)_{\cdot i})$;

$B_{k+1} = \text{prox}_{\gamma\beta\|\cdot\|_1}(W_k - \gamma g_k)$;

$t_{k+1} = \frac{1}{2} + \frac{1}{2}\sqrt{1 + 4t_k^2}$;

$V_{k+1} = A_{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(A_{k+1} - A_k)$;

$W_{k+1} = B_{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(B_{k+1} - B_k)$;

return A and B .

Results and discussion

In this section, we use both synthetic data and real data for algorithm performance evaluation, comparison and validation.

Synthetic data: evaluation and comparison

Network specification and simulated data

To evaluate and compare the proposed method with other state-of-the-art algorithms, we employ GeneNetWeaver [29, 49], the official DREAM Challenge tool for time-course expression data generation. More specifically, GeneNetWeaver uses the GRNs from Yeast (4441 nodes, 12,873 edges) or E. coli (1565 nodes, 3785 edges) as templates to generate network structures with typical complex network properties, of a network size up to ~ 4400 ; and then ordinary differential equation (ODE) models are built upon the previously generated network structures to produce time-course gene expression data at pre-specified time points.

In our simulation studies, we use GeneNetWeaver to generate the network structures for a given network size $n = 10, 100, 1000$, or $20,000$, and we treat such networks as the ground truth. Let m denote the number of edges in a ground truth network, we then randomly add m additional edges to the ground truth network to generate the background network with $2m$ edges. Consequently, the goal of GRN inference algorithms in the simulation studies is to identify as many ground truth edges as possible from all the background network edges, with a controlled false positive rate. It should be stressed that networks of a size $20,000$ cannot be generated from the Yeast or E. coli templates, so we supply the large-scale GRNs

from the RegNetwork database [31] to GeneNetWeaver as templates (e.g., the human GRN in RegNetwork has $\sim 23,000$ nodes and $\sim 370,000$ edges). For each simulated dataset, one data point is generated at each of 6 distinct time points, respectively, for a pre-specified noise level to match the observation scheme in the real data example in Section Real data: validation and findings.

Fairness of comparison and evaluation metrics

Since the state-of-the-art algorithms under comparison such as ARACNE [19] start with a full matrix C , for fairness of comparison, the background regulatory network knowledge is used to confine the outputs of all other methods (unless it has incorporated such information beforehand). All metrics (e.g., sensitivity, specificity, and AUC) have been calculated based on the confined outputs. There are several reasons why we cannot completely re-implement the competing algorithms so they can also start with the prior network instead of the full network: 1) it demands prohibitive efforts to do so – essentially new algorithms need to be developed for the ultra-high dimensional problem, and it is out of the scope of this study; 2) the current design of the competing algorithms also partially justifies the necessity of appropriately using the prior network, which is one of the advantages of the proposed method.

The computing parameters used by existing algorithms are tuned as suggested in their original manuscripts (see Additional file 3: Table S1). For our DMI algorithm, we set $\alpha = 0.08$, $\beta = 0.16$ and $\rho = 10$ for all the simulated datasets, where α and β are determined using cross-validation.

Five commonly-used criteria are considered to measure algorithm performance, including sensitivity (SN), specificity (SP), accuracy (ACC), Matthews correlation coefficient (MCC), and the Area Under ROC Curve (AUC):

$$SN = \frac{TP}{TP + FN},$$

$$SP = \frac{TN}{TN + FP},$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN},$$

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

where TP and TN denote the true positive and true negative, and FP and FN denote the false positive and false negative, respectively.

Performance evaluation

The first set of experiments are conducted to compare our DMI algorithm with nine representative algorithms, including PCC, ARACNE [19], CLR [21], MINET (the maximum relevance minimum redundancy method) [50],

GENIE3 [23], TimeDelay-ARACNE [25], TIGRESS [24], SITPR [27], and Jump3 [28]. Specifically, SITPR [27] is a multi-step pipeline with the ℓ_1 norm penalty incorporated; to avoid manual intervention needed by certain SITPR steps, we simply compare the constrained LASSO step in SITPR with DMI. In the simulated data, the noise level is fixed at 10%, and the networks size ranges from 10 to 1000 because other competing algorithms cannot handle a network size of 20,000.

For the second set of experiments, the noise level is increased from 10 to 30% to evaluate the robustness of our DMI algorithm against noise. The network size in the corresponding simulated data is 10, 100, 1000 or 20,000.

The two sets of experiment results are summarized in Tables 1 and 2, respectively. From Table 1, it can be clearly told that for network size ranging from 10 to 1000, our method significantly outperform other state-of-the-art methods; e.g., the AUCs of our method remain 0.8 or higher for all cases, but the AUCs of all other methods are 0.6 or less and they drop to $0.36 \sim 0.50$ for $n = 1,000$. From Table 2, one can tell that our DMI algorithm is robust against noise because its performance measured by AUC is still greater than 0.73 as the noise level increases to 30% for a network size of 20,000. Additional experiment results can be found in Additional file 4: Table S2.

Real data: validation and findings

Data description and previous work

The real data example for illustrating the use of DMI in practice is from the recent study of Loveday et al. [51], where human A549 cells were infected with influenza H1N1 virus (A/Mexico/InDRE4487/2009). Illumina HumanHT-12 v3 BeadChips and Febit miRBase 14 Geniom miRNA Biochips were employed to measure the mRNA and miRNA samples, respectively. Six replicates of the expression levels were collected at each of six time points (0, 4, 8, 24, 48, and 72 h post infection), but one sample at hour 4 was excluded due to degradation. The within-sample normalized dataset is available from NCBI GEO (GSE36553 and GSE36461). We further conducted between-sample normalization on the data across the six time points.

The original analyses in [51] did not use all the time points; instead, a small number of regulatory relationships between functional miRNAs and gene targets were inferred based on data on hr 0 and hr 8 only. The genome-wide regulatory landscape has not been revealed based on the entire dataset although Liu et al. [27] made an attempt to investigate the subnetwork structures of a size from 2 to 300. In this study, 1,572 genes and 14 non-coding RNAs were identified to be differentially expressed (Fig. 2a) using the FPCA method [52] for a false discovery rate controlled at 0.05. Using the human background network in the RegNetwork database, the DMI algorithm

Table 1 Performance evaluation of DMI and other competing algorithms on a network size 10, 100, or 1,000 at a 10% noise level, based on the arithmetic average over 10 simulation runs

Methods	Size	SN	SP	ACC	MCC	AUC
DMI	10	0.7600±0.2119	0.9143±0.1380	0.8235±0.1301	0.6837±0.2180	0.8371±0.1187
	100	0.7321±0.1818	0.8807±0.0401	0.8069±0.0825	0.6264±0.1460	0.8064±0.0831
	1000	0.8119±0.0472	0.9023±0.0147	0.8564±0.0245	0.7171±0.0460	0.8571±0.0242
CLR	10	0.5900±0.1776	0.4143±0.3125	0.5176±0.1331	0.0043±0.3125	0.5457±0.1331
	100	0.4854±0.0200	0.4916±0.0400	0.4885±0.0202	-0.0231±0.04	0.4768±0.0202
	1000	0.5063±0.0134	0.4908±0.0269	0.4987±0.0135	-0.0029±0.0269	0.4292±0.0135
PCC	10	0.5600±0.0938	0.3714±0.1698	0.4824±0.0719	-0.0686±0.1698	0.5386±0.0719
	100	0.5244±0.0268	0.5301±0.0536	0.5273±0.0268	0.0545±0.0536	0.5148±0.0268
	1000	0.5061±0.0172	0.4907±0.0344	0.4985±0.0173	-0.0032±0.0344	0.4290±0.0173
MINET	10	0.5500±0.0834	0.5571±0.1745	0.5529±0.0915	0.1076±0.1745	0.5871±0.0915
	100	0.3378±0.0319	0.6988±0.0351	0.5194±0.0183	0.0391±0.0351	0.5073±0.0183
	1000	0.2464±0.0164	0.7801±0.0196	0.5092±0.0096	0.0313±0.0196	0.4224±0.0096
TIGRESS	10	0.5700±0.0949	0.3857±0.1355	0.4941±0.1116	-0.0443±0.2304	0.5171±0.0865
	100	0.3154±0.0239	0.6904±0.0405	0.5040±0.0276	0.0067±0.0595	0.4898±0.0281
	1000	0.0062±0.0007	0.9941±0.0014	0.4926±0.0006	0.0024±0.0077	0.3630±0.0008
ARACNE	10	0.2800±0.1200	0.7286±0.1966	0.4647±0.0962	0.0074±0.1966	0.6043±0.0962
	100	0.0598±0.0325	0.9514±0.0357	0.5083±0.0103	0.0225±0.0357	0.4860±0.0103
	1000	0.0188±0.0094	0.9820±0.0246	0.4930±0.0040	0.0023±0.0246	0.3670±0.0040
TimeDelay-ARACNE	10	0.0500±0.1214	0.9571±0.2029	0.4235±0.0744	0.0222±0.2029	0.6386±0.0744
	100	0.0118±0.0250	0.9912±0.0596	0.5044±0.0073	0.0000±0.0596	0.4793±0.0073
	1000	0.0034±0.0028	0.9968±0.0179	0.4924±0.0015	0.0011±0.0179	0.3622±0.0015
GENIE3	10	0.5600±0.0138	0.3714±0.0419	0.4824±0.0098	-0.0686±0.0419	0.5400±0.0098
	100	0.5093±0.0009	0.5153±0.0035	0.5123±0.0010	0.0246±0.0035	0.5007±0.0010
	1000	0.5075±0.0001	0.4922±0.0003	0.5000±0.0001	-0.0003±0.0003	0.4304±0.0001
Jump3	10	0.6400±0.1095	0.4857±0.2048	0.5765±0.1094	0.1257±0.2048	0.6000±0.1094
	100	0.2606±0.0431	0.8390±0.0474	0.5515±0.0245	0.1217±0.0474	0.5471±0.0245
	1000	0.0384±0.0074	0.9843±0.0192	0.5040±0.0035	0.0694±0.0192	0.3823±0.0035
SITPR	10	0.4900±0.1197	0.7143±0.2020	0.5824±0.0757	0.2172±0.1819	0.6021±0.0862
	100	0.1610±0.0794	0.8253±0.0656	0.4952±0.0255	-0.0212±0.0733	0.4931±0.0257
	1000	0.2382±0.0509	0.7599±0.0381	0.4950±0.0118	-0.0032±0.0279	0.4990±0.0114

was applied ($\alpha = 0.08$ and $\beta = 0.22$) to identify 1926 regulatory interactions between the differentially expressed genes or miRNAs, as shown in Fig. 2b.

Network analysis

We conducted comprehensive network analyses on the topological structure and found that the A549 GRN in response to influenza H1N1 infection is a typical complex network, evidenced by, e.g., its power law degree distribution and a large clustering coefficient (see Additional file 5: Figure S1 and Additional file 6: Table S3 for details). A close examination of the inferred GRN by the DMI algorithm suggests consistency between the findings in this study and those in literature. For instance,

'CEBPB' (CCAAT/Enhancer-Binding Protein Beta) is a transcription factor (TF) that plays an important role in immune and inflammatory responses [53], and the subnetwork structure centered at 'CEBPB' has been previously reported in [27]. Our study also revealed a similar subnetwork structure around 'CEBPB' (e.g., interactions between 'CEBPB' and 'NCS1', 'ISG20', or 'ABCG1'. See Additional file 7: Figure S2); however, some experimentally-verified interactions (e.g., between 'CEBPB' and 'NOLC1' in HPRD) were successfully identified in this study, but were missing in the work of Liu et al. [27]. More interestingly, we found a number of hub gene structures, e.g., centered at the Activating Transcription Factor genes ('ATF2' and 'ATF4') or the E2F Transcription

Table 2 Evaluation of DMI at a noise level from 10 to 30% for a network size 10, 100, 1,000 or 20,000, based on the arithmetic average over 10 simulation runs

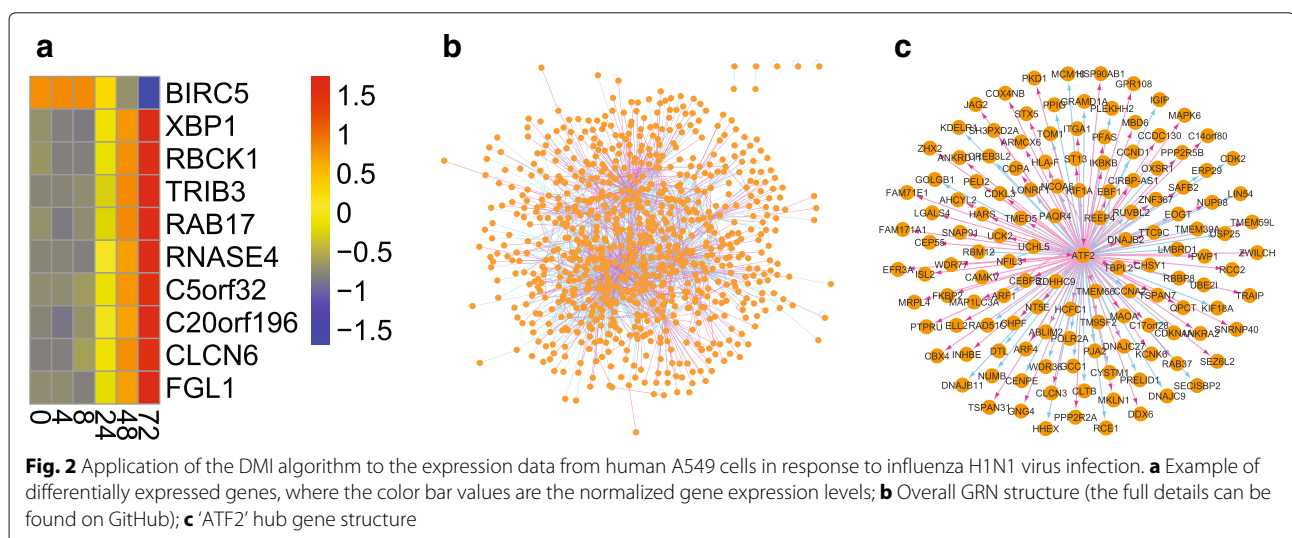
Noise level	Size	SN	SP	ACC	MCC	AUC
10% noise	10	0.9333 ± 0.0573	0.9000 ± 0.0860	0.9200 ± 0.0688	0.8333 ± 0.1434	0.9166 ± 0.0717
	100	0.7687 ± 0.0379	0.8253 ± 0.0407	0.7984 ± 0.0283	0.5965 ± 0.0571	0.7970 ± 0.0282
	1000	0.8273 ± 0.0244	0.8305 ± 0.0131	0.8289 ± 0.0182	0.6579 ± 0.0363	0.8289 ± 0.0182
	20k	0.7748 ± 0.0041	0.7753 ± 0.0041	0.7751 ± 0.0041	0.5502 ± 0.0083	0.7751 ± 0.0041
20% noise	10	0.8000 ± 0.1365	0.7000 ± 0.2048	0.7600 ± 0.1639	0.5000 ± 0.3414	0.7500 ± 0.1707
	100	0.7312 ± 0.0440	0.7569 ± 0.0398	0.7447 ± 0.0418	0.4881 ± 0.0838	0.7440 ± 0.0419
	1000	0.8056 ± 0.0133	0.8006 ± 0.0137	0.8031 ± 0.0135	0.6063 ± 0.0271	0.8031 ± 0.0135
	20k	0.7452 ± 0.0045	0.7457 ± 0.0045	0.7454 ± 0.0045	0.4909 ± 0.0091	0.7454 ± 0.0045
30% noise	10	0.7333 ± 0.1194	0.6000 ± 0.1791	0.6800 ± 0.1433	0.3333 ± 0.2986	0.6667 ± 0.1493
	100	0.7005 ± 0.0320	0.7291 ± 0.0289	0.7155 ± 0.0304	0.4297 ± 0.0610	0.7148 ± 0.0305
	1000	0.7829 ± 0.0080	0.7774 ± 0.0082	0.7802 ± 0.0081	0.5603 ± 0.0162	0.7801 ± 0.0081
	20k	0.7339 ± 0.0048	0.7344 ± 0.0048	0.7342 ± 0.0048	0.4684 ± 0.0096	0.7342 ± 0.0048

Factor genes ('E2F6' and 'E2F7'). More specifically, 'ATF2' encodes a DNA binding protein of the leucine zipper family that can perform distinct functions via different mechanisms. For example, it can form heterodimers with c-Jun, both of which are activating protein-1 (AP-1) family transcription factors. AP-1 is known to activate antiviral cytokines in cells as a defensive response against infection caused by influenza virus and other viruses [54, 55]. Therefore, our results reflect the known biological events. Also, our results showed that 'ATF2' interacts with more than 130 other genes, including 'CEBPB' (Fig. 2c). Although it has been previously reported that influenza A viral RNA molecules can indirectly activate transcription factors like ATF2 [56], the importance of ATF2 in the context of influenza infection has not been fully appreciated by previous experimental studies, considering the

large number of target genes associated with ATF2 in the hub gene structure. The proposed DMI algorithm thus provides us not only an opportunity to investigate the genome-wide regulatory landscapes but also a way to identify interesting subnetworks like hub gene structures. We defer the discussions on hub structures to the next section.

Validation and findings

Although the ENCODE database [57] is not used when constructing the background network, 102 TFs, 723 miRNAs and 13,607 target genes in the background network can also be found in ENCODE. Therefore, we used the TF-gene interactions verified by the ENCODE CHIP-seq data as the ground truth for validating our analysis results, as suggested in [28]. It is impossible to list all the hub



structures here, so we selected the 5 biggest hubs from the GRN structure identified from real data by DMI if the centers of these hubs (i.e., hub genes) are also included in ENCODE. The first interesting finding we noticed is that all the five hub nodes are transcription factors, i.e., 'ATF2', 'CEBPB', 'CUX1', 'E2F6' and 'EBF1'. Since DMI is designed to identify genuine regulator-target interactions, such an observation is expected. Second, 97, 52, 40, 85, and 47 TF-gene interactions were found in ENCODE for the five TFs above, respectively. That is, a large set of known interactions (321 in total) were considered here for result validation. Third, according to Table 1, the top three algorithms are DMI, GENIE3 and SITPR (ranked based on their AUCs for a network size 1,000) so we compared their performances here. As shown in Table 3, DMI's performance is superior or comparable to that of GENIE3 and SITPR for all the five sets of known interactions. For instance, out of 97 ENCODE interactions associated with 'ATF2', DMI identified 74 interactions and thus achieved a precision ($\frac{\Delta \# \text{ of predicted interactions}}{\# \text{ of known interactions}}$) of 0.7629, which is the same as that of GENIE3; however, for 'E2F6', DMI identified 79 ENCODE interactions while GENIE3 only identified 52. We thus believe the proposed algorithm can more accurately identify the activated interactions than other competing algorithms.

Motivated by the finding above, We further examined the top 10 hub structures and found that transcription factors are the center nodes of all such hubs; i.e., 'ATF2', 'E2F6', 'E2F7', 'ATF4', 'CUX1', 'CEBPB', 'EBF1', 'XBPI1', 'NFIL3', and 'FOXO4'. Therefore, besides the roles of 'ATF2' discussed in the previous section, it is also interesting to understand the roles of the rest hub TFs in the context of virus infection. For example, 'CEBPB' also has a leucine zipper structure to interact with DNA. It is capable of forming homodimers or heterodimers with ATF4, and previous studies showed that 'CEBPB' is essential in immune responses due to its involvement in inducing macrophagic functions during inflammation [53, 58]. The work by Granberg et al. [59] demonstrated that 'CEBPB' has antiviral properties in the presence of adenovirus infection. However, further investigations are required to verify whether 'CEBPB' retains these properties facing H1N1 infection. 'CUX1' is a homeodomain family DNA-binding protein, and it has been shown to regulate cell cycle, and previously identified as a tumor

suppressor [60]. One study reported that 'CUX1' is capable of repressing human cytomegalovirus infection by regulating the viral major immediate early (MIE) gene expression [61]. 'E2F6' is a member of the well-known E2F transcription factor family that regulates cell cycle. Overexpression of 'E2F6' has been shown to stall cell cycle in S-phase by interacting with E2F target genes and reducing their expressions. The study by Zhang et al. [62] shows that 'E2F6' has abnormal expression in Epstein-Barr virus (EBV)-associated nasopharyngeal carcinoma. Early B cell factor 1 ('EBF1') is a transcription factor that is critical for specifying B cell identity and lineage maintenance [63]. And in turn, B cells produce antibodies that can help fight-off viral infection. Although there is not much known link between 'EBF1' and influenza viral infection, one study claimed that 'EBF1' has the function to regulate Epstein-Barr virus latent membrane protein 1 ('LMP1') expression so that tumorigenesis can be prevented [64]. The roles of 'XBPI1', 'NFIL3', and 'FOXO4' have also been reported before (see, e.g., [65]), but their contributions to immune responses against influenza virus infection need to be elucidated by further experiments. In short, one can hypothesize that these hub TFs may show the same antiviral activities facing H1N1 as they do for other viruses, and the GRN identified by our DMI algorithm could be a useful guide to biomedical researchers when they are choosing which gene functions in a biological process to validate by bench work.

We also found that a number of big hub structures centered at non-coding RNAs such as 'hsa-miR-30a', 'hsa-miR-30c-5p', 'hsa-miR-30d-tp', 'hsa-miR-543', and 'hsa-miR-26a'. Even though studies on microRNAs are rapidly expanding, functions of these dynamic molecules are hard to enumerate. Therefore, discussing the relationship between these microRNAs and influenza is beyond the scope of this article.

Conclusions

We proposed a novel scalable algorithm called DMI for large-scale GRN inference from temporal gene expression data in this study. Extensive simulation studies and real data application suggest the superiority of the DMI algorithm over other state-of-the-art approaches, and the success mainly relies on the incorporation of multiple topological characteristics of GRNs like sparsity and

Table 3 Prediction performance based on ENCODE data

	ATF2		CEBPB		CUX1		E2F6		EBF1	
	#TP	Precision	#TP	Precision	#TP	Precision	#TP	Precision	#TP	Precision
DMI	74	0.7629	49	0.9423	24	0.6000	79	0.9294	42	0.8936
GENIE3	74	0.7629	38	0.7308	22	0.5500	52	0.6118	35	0.7447
SITPR	37	0.3814	32	0.6154	24	0.6000	50	0.5882	35	0.7447

hub gene structures into our model formulation. Using the gene expression data generated from IAV infected human respiratory epithelial cells, our algorithm found that the top ten hub structure are all TFs. This finding is biologically reasonable, considering TFs' central roles in regulating gene expression. Some of these TFs have known association with virus infection but not with IAV. Therefore, the GRN produced by our algorithm could reveal previously unidentified biological links, and guide researchers to choose target genes to investigate.

We also recognize several limitations of the current study: 1) DMI's performance can be further improved if the background network edges can be appropriately weighted; 2) Prediction of previously unseen interactions heavily depends on the background network preparation so further efforts need to be invested to the development of public knowledgebases like RegNetwork; 3) This work mainly focuses on solving the ultra-high dimensional inference problems, and data heterogeneity and integration issues have not been tackled; 4) Our real data analysis is based on only one data set, and more data is needed to establish more solid GRN. We believe that it is necessary to address the aforementioned limitations in separate articles, and this study provides a solid basis for such future investigations.

Additional files

Additional file 1: Text S1. System requirements for running DMI. (PDF 52.0 kb)

Additional file 2: Text S2. Parameter tuning guideline for DMI. (PDF 64.4 kb)

Additional file 3: Table S1. Computational settings for the competing methods under comparison. (PDF 56.5 kb)

Additional file 4: Table S2. Additional experiment results of algorithm performance evaluation of DMI on a real background network with 2768 nodes. We retrieved the UCSC background network for human from the RegNetwork database (<http://www.regnetworkweb.org/>), and then selected the highly confident interactions as the activated edges (RegNetwork provides the tool to select edges of different levels of confidence). The number of such activated edges is about 1/6 of the total number of edges. In this way, we avoid introducing or removing any random edge into the background network. (PDF 59.7 kb)

Additional file 5: Figure S1. In-degree distribution of the A549 GRN. The power-law model fitting result is labelled in red. (PDF 65.6 kb)

Additional file 6: Table S3. Basic network statistics of the A549 GRN. (PDF 138 kb)

Additional file 7: Figure S2. Subnetwork structure centered at 'CEBPB'. The subnetwork structure is identified using the DMI algorithm based on the time-course expression data from human A549 cells in response to IAV infection. Blue arrows stand for negative regulatory relationships, and yellow arrows stand for positive regulatory relationships. (PDF 130 kb)

Abbreviations

CLR: Context likelihood of relatedness; DMI: Decomposable multi-structure identification; GRN: Gene regulatory network; IAV: Influenza A virus; MI: Mutual information; ODE: Ordinary differential equation; PCC: Pearson correlation coefficient; SITPR: Systematic identification of transcriptional and post-transcriptional regulation; TF: Transcription factor

Acknowledgements

We thank Dr. Zhiping Liu (Shandong University, China) and Dr. Hulin Wu (UTSPH) for developing the RegNetwork database.

Funding

This work was partially supported by NIH/NEI grants R01EY022356, R01EY018571, and R01EY020540 (RC), NSF grant CNS-1548078 (JL), and NSF grant DMS-1620957 (HM).

Availability of data and materials

All the data sets used in this study are from public databases, including RegNetwork, GEO, and ENCODE.

Authors' contributions

SG implemented the algorithm, conducted experiments, and wrote the manuscript. AR, RC and LW contributed to result interpretation and biological findings. JL and HM proposed the idea, oversaw the study and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

¹Department of Computer Science, University of Rochester, 14620 Rochester, NY, USA. ²Department of Molecular Virology and Microbiology, Baylor College of Medicine, 77030 Houston, TX, USA. ³Department of Molecular and Human Genetics, Baylor College of Medicine, 77030 Houston, TX, USA. ⁴Department of Biostatistics, University of Texas Health Science Center, 77030 Houston, TX, USA. ⁵Goergen Institute for Data Science, University of Rochester, 14620 Rochester, NY, USA.

Received: 22 October 2016 Accepted: 20 January 2017

Published online: 31 January 2017

References

- Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol.* 2008;9(10):770–80.
- Barabasi AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nat Rev Genet.* 2004;5:101–13.
- Barabasi AL, Culbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet.* 2011;12:56–68.
- Kidd BA, Peters LA, Schadt EE, Dudley JT. Unifying immunology with informatics and multiscale biology. *Nat Immunol.* 2014;15(2):118–27.
- Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet.* 2010;11(1):31–46.
- Chen R, Snyder M. Promise of personalized omics to precision medicine. *Wiley Interdiscip Rev Syst Biol Med.* 2013;5(1):73–82. doi:10.1002/wsbm.1198.
- Edgar R, Domrachev M, Lash AE. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30(1):207–10.
- Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum Mol Genet.* 2014. doi:10.1093/hmg/ddu309. <http://hmg.oxfordjournals.org/content/early/2014/07/01/hmg.ddu309.full.pdf+html>.
- Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Consortium D, Kellis M, Collins JJ, Stolovitzky G. Wisdom of crowds for robust gene network inference. *Nat Methods.* 2012;9(8):796–804. doi:10.1038/nmeth.2016.
- Opgen-Rhein R, Strimmer K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst Biol.* 2007;1:37. doi:10.1186/1752-0509-1-37.
- Langfelder P, Horvath S. Wgcna: an R package for weighted correlation network analysis. *BMC Bioinforma.* 2008;9(1):559.

12. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002;18(2):261–74.
13. Hartemink A. In: Do K-A, Muller P, Vannucci M, editors, editors. Bayesian networks and informative priors: Transcriptional regulatory network models. Cambridge: Cambridge University Press; 2006, pp. 401–24.
14. Hirose O, Yoshida R, Imoto S, Yamaguchi R, Higuchi T, Charnock-Jones DS, Print C, Miyano S. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*. 2008;24(7):932–42. doi:10.1093/bioinformatics/btm639.
15. Yeung MK, Tegner J, Collins JJ. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A*. 2002;99(9):6163–8. doi:10.1073/pnas.092576199.
16. De Jong H. Modeling and simulation of genetic regulatory systems: A literature review. *J Comput Biol*. 2002;9(1):67–103.
17. Bansal M, Belcastro V, Ambesi-Impombato A, di Bernardo D. How to infer gene networks from expression profiles. *Mol Syst Biol*. 2007;3:78. doi:10.1038/msb4100120.
18. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. *Proc Natl Acad Sci U S A*. 2010;107(14):6286–91. doi:10.1073/pnas.0913357107.
19. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinforma*. 2006;7(Suppl 1):7.
20. Cover TM, Thomas JA. *Elements of Information Theory*. New York: Wiley; 1991.
21. 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):8.
22. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*. 2000;5:418–29.
23. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*. 2010;9(5). doi:10.1371/journal.pone.0012776.
24. Haury AC, Mordelet F, Vera-Licona P, Vert JP. Tigrass: Trustful inference of gene regulation using stability selection. *BMC Syst Biol*. 2012;6:145. doi:10.1186/1752-0509-6-145.
25. Zoppoli P, Morganello S, Ceccarelli M. Timedelay-aracne: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinforma*. 2010;11(1):1–15. doi:10.1186/1471-2105-11-154.
26. Yang X, Dent JE, Nardini C. An s-system parameter estimation method (sperm) for biological networks. *J Comput Biol*. 2012;19(2):175–87.
27. Liu ZP, Wu H, Zhu J, Miao H. Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza virus infection. *BMC Bioinforma*. 2014;15(1):1.
28. Huynh-Thu VA, Sanguinetti G. Combining tree-based and dynamical systems for the inference of gene regulatory networks. *Bioinformatics*. 2015;31(10):1614–22. doi:10.1093/bioinformatics/btu863.
29. Schaffter T, Marbach D, Floreano D. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*. 2011;27(16):2263–70. doi:10.1093/bioinformatics/btr373.
30. Leclerc RD. Survival of the sparsest: robust gene networks are parsimonious. *Mol Syst Biol*. 2008;4(1):213. doi:10.1038/msb.2008.52.
31. Liu ZP, Wu C, Miao H, Wu H. Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*. 2015;2015:095.
32. Ravasi T, Suzuki H, Cannistraci CV, et al. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*. 2010;140(5):744–52.
33. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*. 2006;34(Database issue):108–10.
34. Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. Jasp, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*. 2008;36(Database issue):102–6.
35. Kanehisa M, Goto S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27–30.
36. Mishra GR, Suresh M, Kumaran K, et al. Human protein reference database–2006 update. *Nucleic Acids Res*. 2006;34(Database issue):411–4.
37. Flicek P, Amode MR, Barrell D, et al. Ensembl 2012. *Nucleic Acids Res*. 2012;40(Database issue):84–90. doi:10.1093/nar/gkr991.
38. Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):50912. doi:7898.
39. Albert R. Scale-free networks in cell biology. *J Cell Sci*. 2005;118(Pt 21):4947–57. doi:10.1242/jcs.02714.
40. 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. <http://bioinformatics.oxfordjournals.org/content/21/1/71.full.pdf+html>.
41. Kim SY, Imoto S, Miyano S. Inferring gene networks from time series microarray data using dynamic bayesian networks. *Brief Bioinform*. 2003;4(3):228–35. doi:10.1093/bib/4.3.228. <http://bib.oxfordjournals.org/content/4/3/228.full.pdf+html>.
42. Costa LF, Rodrigues FA, Travieso G, Boas PRV. Characterization of complex networks: A survey of measurements. *Adv Phys*. 2007;56(1):167–242. doi:10.1080/00018730601170527.
43. Huang J, Zhang T, et al. The benefit of group sparsity. *Ann Stat*. 2010;38(4):1978–2004.
44. Candès E, Romberg J. l1-magic: Recovery of sparse signals via convex programming. 2005;4:46. www.acm.caltech.edu/l1magic/downloads/l1magic.pdf.
45. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations Trends® Mach Learn*. 2011;3(1):1–122.
46. Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *J Optim Theory Appl*. 2001;109(3):475–94.
47. Nesterov Y. Smooth minimization of non-smooth functions. *Math Program*. 2005;103(1):127–52.
48. Nesterov Y. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Math Doklady*. 1983;27(2):372–6.
49. 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. WingX.
50. Meyer PE, Lafitte F, Bontempi G. minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinforma*. 2008;9(1):1–10. doi:10.1186/1471-2105-9-461.
51. Loveday EK, Svinti V, Diederich S, Pasick J, Jean F. Temporal- and strain-specific host microRNA molecular signatures associated with swine-origin h1n1 and avian-origin h7n7 influenza a virus infection. *J Virol*. 2012;86(11):6109–122. doi:10.1128/JVI.06892-11.
52. Wu S, Wu H. More powerful significant testing for time course gene expression data using functional principal component analysis approaches. *BMC Bioinforma*. 2013;14(1):1–13. doi:10.1186/1471-2105-14-6.
53. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. Genecards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*. 1998;14(8):656–64.
54. Wang X, Li M, Zheng H, Muster T, Palese P, Beg AA, Garcia-Sastre A. Influenza a virus ns1 protein prevents activation of $\text{nf-}\kappa\text{b}$ and induction of alpha/beta interferon. *J Virol*. 2000;74(24):11566–73.
55. Ludwig S, Ehrhardt C, Neumeier ER, Kracht M, Rapp UR, Pleschka S. Influenza virus-induced ap-1-dependent gene expression requires activation of the jnk signaling pathway. *J Biol Chem*. 2001;276(14):10990–8.
56. Kochs G, Garcia-Sastre A, Martínez-Sobrido L. Multiple anti-interferon actions of the influenza a virus ns1 protein. *J Virol*. 2007;81(13):7011–21.
57. Gerstein MB, Kundaje A, Hariharan M, et al. Architecture of the human regulatory network derived from encode data. *Nature*. 2012;489(7414):91–100. doi:10.1038/nature11245.
58. Poli V. The role of c/ebp isoforms in the control of inflammatory and native immunity functions. *J Biol Chem*. 1998;273(45):29279–82.
59. Granberg F, Svensson C, Pettersson U, Zhao H. Adenovirus-induced alterations in host cell gene expression prior to the onset of viral gene expression. *Virology*. 2006;353(1):1–5.

60. Ramdzan ZM, Nepveu A. Cux1, a haploinsufficient tumour suppressor gene overexpressed in advanced cancers. *Nat Rev Cancer*. 2014;14(10):673–82.
61. Stern JL, Cao JZ, Xu J, Mocarski ES, Slobedman B. Repression of human cytomegalovirus major immediate early gene expression by the cellular transcription factor ccaat displacement protein. *Virology*. 2008;378(2):214–25.
62. Zhang W, Zeng Z, Zhou Y, Xiong W, Fan S, Xiao L, Huang D, Li Z, Li D, Wu M, et al. Identification of aberrant cell cycle regulation in epstein-barr virus-associated nasopharyngeal carcinoma by cdna microarray and gene set enrichment analysis. *Acta Biochim Biophys Sin*. 2009;41(5):414–428.
63. Nechanitzky R, Akbas D, Scherer S, Györy I, Hoyle T, Ramamoorthy S, Diefenbach A, Grosschedl R. Transcription factor ebf1 is essential for the maintenance of b cell identity and prevention of alternative fates in committed cells. *Nat Immunol*. 2013;14(8):867–75.
64. Arvey A, Tempera I, Tsai K, Chen HS, Tikhmyanova N, Klichinsky M, Leslie C, Lieberman PM. An atlas of the epstein-barr virus transcriptome and epigenome reveals host-virus regulatory interactions. *Cell Host Microbe*. 2012;12(2):233–45.
65. Xu W, Domingues R, Fonseca-Pereira D, Ferreira M, Ribeiro H, Lopez-Lastra S, Motomura Y, Moreira-Santos L, Bihl F, Braud V, Kee B, Brady H, Coles M, Vosshenrich C, Kubo M, Di Santo J, Veiga-Fernandes H. NFIL3 orchestrates the emergence of common helper innate lymphoid cell precursors. *Cell Rep*. 2015;10(12):2043–54. doi:10.1016/j.celrep.2015.02.057.

Submit your next manuscript to BioMed Central
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

