

RESEARCH ARTICLE

Open Access

# Exact reconstruction of gene regulatory networks using compressive sensing

Young Hwan Chang<sup>1</sup>, Joe W Gray<sup>2</sup> and Claire J Tomlin<sup>1,3\*</sup>

## Abstract

**Background:** We consider the problem of reconstructing a gene regulatory network structure from limited time series gene expression data, without any a priori knowledge of connectivity. We assume that the network is sparse, meaning the connectivity among genes is much less than full connectivity. We develop a method for network reconstruction based on compressive sensing, which takes advantage of the network's sparseness.

**Results:** For the case in which all genes are accessible for measurement, and there is no measurement noise, we show that our method can be used to exactly reconstruct the network. For the more general problem, in which hidden genes exist and all measurements are contaminated by noise, we show that our method leads to reliable reconstruction. In both cases, coherence of the model is used to assess the ability to reconstruct the network and to design new experiments. We demonstrate that it is possible to use the coherence distribution to guide biological experiment design effectively. By collecting a more informative dataset, the proposed method helps reduce the cost of experiments. For each problem, a set of numerical examples is presented.

**Conclusions:** The method provides a guarantee on how well the inferred graph structure represents the underlying system, reveals deficiencies in the data and model, and suggests experimental directions to remedy the deficiencies.

**Keywords:** Gene regulatory networks, Inference, Compressive sensing

## Introduction

Mathematical modeling of biological signaling pathways can provide an intuitive understanding of their behavior [1-3]. However, since typically only incomplete knowledge of the network structure exists and the system dynamics is known to be sufficiently complex, the challenge has become to show that the identified networks and corresponding mathematical models are enough to adequately represent the underlying system. In the last years, many data-driven mathematical tools have been developed and applied to reconstruct graph representations of gene regulatory networks (GRNs) from data. These include Bayesian networks, regression, correlation, mutual information and system-based approaches [4-10]. Also, these approaches either focus on static or on time series data. The latter approach has the advantage

of being able to identify dynamic relationships between genes.

However, data-driven reconstruction of the network structure itself remains in general a difficult problem; nonlinearities in the system dynamics and measurement noise make this problem even more challenging. For linear time invariant (LTI) systems, there exist necessary and sufficient conditions for network reconstruction [11]. However, for time-varying or nonlinear systems, there has not been as yet any statistical guarantee on how well the inferred model represents the underlying system [12-15]. The recent work [16] addresses the problem of data-driven network reconstruction, together with measurement noise and unmodelled nonlinear dynamics, yet this work points out that these complications impose a limit on the reconstruction, and with strong nonlinear terms the method fails. Additionally, identifying whether important nodes in the graph structure are missing, how many are missing, and where these nodes are in the interconnection structure remains a challenging problem.

\*Correspondence: tomlin@eecs.berkeley.edu

<sup>1</sup>Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 94720 Berkeley, CA, USA

<sup>3</sup>Faculty Scientist, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

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

Moreover, in order to continue to have an impact in systems biology, identification of the graph topology from data should be able to reveal deficiencies in the model and suggest new experimental directions [17]. For example, Steinke *et al.* [18] presented a Bayesian method for identifying a gene regulatory network from micro-array measurements in perturbation experiments and showed how to use optimal design to minimize the number of measurements.

Since biological regulatory networks are known to be sparse [19-23], meaning that most genes interact with only a small number of genes compared with the total number in the network, many methods [12-15,24-27] take advantage of the sparsity. The methods typically use  $l_1$ -norm optimization, which leads to a sparse representation of the network and improves the ability to find the actual network structure. Even though many methods [14,15] show that the reconstruction results are fairly good, the methods cannot guarantee *exact recovery*. This stems from the fact that the so-called *incoherence* condition is typically not satisfied for the matrix  $\Omega$  in a linear measurement model  $\mathbf{Y} = \Omega\mathbf{q}$ , where  $\mathbf{q}$  is the signal to be reconstructed,  $\mathbf{Y}$  is the measurement, and  $\Omega$  is known as the sensing matrix. In this paper, we construct  $\mathbf{Y}$  and  $\Omega$  from time series protein or gene expression data with candidate basis functions or kinetic features, and  $\mathbf{q}$  reflects the (unknown) underlying GRN structure to be reconstructed. Roughly, incoherence is a measure of the correlation between columns of the sensing matrix. Since the incoherence condition of the sensing matrix provides a metric of performance, this is one of the motivating factors for the use of compressive sensing (CS) [28] in GRN reconstruction. CS is a signal processing technique for efficiently acquiring and reconstructing a signal by taking advantage of the signal's sparsity and allowing the entire signal to be determined from relatively few measurements under a certain condition, i.e., it requires that the incoherence condition to be satisfied. In the Human Epidermal Growth Factor Receptor2 (HER2) positive breast cancer signaling pathway that we studied in [29,30], time series data sets consist of only 8 time point measurements of 20 protein signals, and we would like to use this limited data to identify a graph structure which could have  $20 \times 20$  or 400 edges.

In this paper, we consider reconstruction of biochemical reaction networks (protein-protein interaction) or GRNs (including mRNA, transcription factors). Biological processes in cells are properly performed by gene regulation, signal transduction and interactions between proteins and thus, the network structure that we consider is an abstraction of the system's biochemical reaction dynamics, describing the manifold ways in which one substance affects all the others to which it is connected. Thus, we are interested in directed graph representations

of signaling pathways and propose models of biochemical reaction networks or GRNs as a set of basis functions or features that best explain a set of time series observations such as protein expression or gene expression time series data. We develop a new algorithm for GRN reconstruction based on CS. First, we focus on sparse graph structures using limited time series data with all nodes accessible and no measurement noise. We test the network reconstruction algorithm on a simulated biological pathway in which the structure is known *a priori*. We explain the limitation of the proposed method's performance when the dataset naturally has high coherence and propose a way to overcome this limitation by designing additional effective experiments. Thus, one big part of the whole study is the method for proposing new experiments. As systems biology matures, experimental study design becomes more important in systems biology rather than simply building computational models of experimental data. Thus, the proposed method can be useful for experimental design. By revealing deficiencies in the data and model, the method can suggest experimental directions to remedy the deficiencies. By doing this, we can minimize the number of experiments and reduce the cost of experiments.

Next, the proposed algorithm is extended to a more general problem: we consider partially corrupted data with data inconsistencies due to model mismatch, and measurement noise affecting all the data. Typically, data inconsistencies may result from missing nodes in the model; or in some cases arbitrary data corruption may result from human errors such as mislabeling or the improper use of markers or antibodies. The question is whether one can still recover the graph structure reliably under these conditions. Inspired by a robust error correction method [31], the exact recovery of the graph structure can be guaranteed under suitable conditions on the node dynamics, provided that hidden nodes can affect relatively few nodes in the graph structure. Also, a set of numerical examples is provided to demonstrate the method, including some from an RPPA (Reverse Phase Protein Array) dataset [32] collected from HER2 positive breast cancer cell lines.

In this paper, the main contributions are the following.

- The CS framework uses the coherence of the sensing matrix as a performance index, which allows us to assess and optimize mathematically network reconstruction.
- Coherence also provides a guideline for optimizing experiment design for network reconstruction.
- By utilizing an error correction method in conjunction with the CS framework, network reconstruction may be performed even when there are hidden nodes and measurement noise.

## Background

### Overview: compressive sensing

Consider measurements  $\mathbf{Y} \in \mathbb{R}^m$  of a signal  $\mathbf{q} \in \mathbb{R}^n$ :

$$\mathbf{Y} = \mathbf{\Omega} \mathbf{q} \tag{1}$$

where  $\mathbf{\Omega} \in \mathbb{R}^{m \times n}$  is called the sensing matrix.

One key question [33] is how many measurements  $m$  are needed to exactly recover the original signal  $\mathbf{q}$  from  $\mathbf{\Omega}$ :

- If  $m > n$  and  $\mathbf{\Omega}$  is a full rank matrix, then the problem is overdetermined. If  $m = n$  and  $\mathbf{\Omega}$  is a full rank matrix, the problem is determined and may be solved uniquely for  $\mathbf{q}$ .
- If  $m < n$ , the problem is underdetermined even if  $\mathbf{\Omega}$  has full rank. We can restrict  $\mathbf{q} \in \mathbb{R}^n$  to the subspace which satisfies  $\mathbf{Y} = \mathbf{\Omega} \mathbf{q}$ . However,  $\mathbf{q}$  cannot be determined uniquely.

For the underdetermined case, the least squares solution  $\mathbf{q}^* = \arg \min_{\mathbf{q}} \|\mathbf{q}\|_{l_2} = \mathbf{\Omega}^*(\mathbf{\Omega} \mathbf{\Omega}^*)^{-1} \mathbf{Y}$  is typically used as the “best guess” in many applications. However, if  $\mathbf{q}$  is known to be sparse, meaning that many of its components are zero, one might expect that fewer than  $n$  measurements are needed to recover it. It is thus of interest to obtain a good estimator for underdetermined problems in which  $\mathbf{q}$  is assumed to be  $s$ -sparse, meaning that  $s$  of the elements of  $\mathbf{q}$  are nonzero. In principle, the theory of compressive sensing (CS) asserts that the number of measurements needed to recover  $\mathbf{q} \in \mathbb{R}^n$  is a number proportional to the compressed size of the signal ( $s$ ), rather than the uncompressed size ( $n$ ) [34]. To be able to recover  $\mathbf{q}$ , CS relies on two properties: *sparsity*, which pertains to the signals of interest, and *incoherence*, which pertains to the sensing matrix.

**Proposition 1.** [33]. *Suppose that any  $2s$  columns of an  $m \times n$  matrix  $\mathbf{\Omega}$  are linearly independent (this is a reasonable assumption if  $m \geq 2s$ ). Then, any  $s$ -sparse signal  $\mathbf{q} \in \mathbb{R}^n$  can be reconstructed uniquely from  $\mathbf{\Omega} \mathbf{q}$ .*

The proof [33] of the above proposition also shows how to reconstruct an  $s$ -sparse signal  $\mathbf{q} \in \mathbb{R}^n$  from the measurement  $\mathbf{Y} = \mathbf{\Omega} \mathbf{q}$  where  $\mathbf{q}$  is the unique sparsest solution to  $\mathbf{Y} = \mathbf{\Omega} \mathbf{q}$ :

$$\mathbf{q}^* = \arg \min_{\mathbf{q}} \|\mathbf{q}\|_{l_0} \text{ subject to } \mathbf{Y} = \mathbf{\Omega} \mathbf{q} \tag{2}$$

and  $\|\mathbf{q}\|_{l_0} := \sum_{i=1}^n I(\mathbf{q}_i \neq 0)$  is the cardinality of  $\mathbf{q}$ . However, the  $l_0$ -minimization is computationally intractable (NP-hard in general). Recent breakthroughs enable approximating  $l_0$ -optimization by using  $l_1$ -minimization which is a convex optimization problem

and can be solved in a simple but effective way by linear programming:

$$\mathbf{q}^* = \arg \min_{\mathbf{q}} \|\mathbf{q}\|_{l_1} \text{ subject to } \mathbf{Y} = \mathbf{\Omega} \mathbf{q} \tag{3}$$

The  $l_1$ -minimization in Equation (3) requires mild over-sampling, more specifically,  $m \geq c \cdot \mu(\Phi, \Psi)^2 \cdot s \log n$  for some positive constant  $c$  where  $\mathbf{\Omega}$  can be decomposed as a product of a sparsity basis  $\Psi$ , and an orthogonal measurement system  $\Phi$  [28]. We have  $m$  measurement in the  $\Phi$  domain (*sensing modality*) under bases  $\Psi$  (*signal model*) and  $\mu$  represents the coherence defined as follows:

**Definition 1.** [28]. *The coherence of a matrix  $\mathbf{\Omega}$  is given by*

$$\mu(\mathbf{\Omega}) = \max_{j < k} \frac{|\langle \mathbf{\Omega}_j, \mathbf{\Omega}_k \rangle|}{\|\mathbf{\Omega}_j\|_2 \|\mathbf{\Omega}_k\|_2}$$

where  $\mathbf{\Omega}_j$  and  $\mathbf{\Omega}_k$  denote columns of  $\mathbf{\Omega}$ . In plain English, the coherence measures that the largest correlation between any two columns in  $\mathbf{\Omega}$ .

Several theoretical results [35] ensure that the  $l_1$ -minimization guarantees exact recovery whenever the sensing matrix  $\mathbf{\Omega}$  is sufficiently *incoherent*. For example, we can say that  $\mathbf{\Omega}$  is *incoherent* if  $\mu$  is small. Coherence is a key property in the compressed sensing framework because if two columns are closely correlated, it would be very difficult to distinguish whether the influence from the components in the sparse signal comes from one or the other (recall that the measurement  $\mathbf{Y} (= \mathbf{\Omega} \mathbf{q})$  is a linear combination of each columns of the sensing matrix  $\mathbf{\Omega}$  with the components in  $\mathbf{q}$  as coefficients). Also, numerical experiments suggest that in practice, most  $s$ -sparse signals are in fact recovered exactly once  $m \geq 4s$  [28]. Here, “exact recovery” means that we find the sparsest solution ( $\mathbf{q}$ ) such that  $\mathbf{Y} = \mathbf{\Omega} \mathbf{q}$ .

Therefore, if the sensing matrix  $\mathbf{\Omega}$  satisfies the incoherence condition ( $m \geq c \cdot \mu(\Phi, \Psi)^2 \cdot s \log n$ , or in practice,  $m \geq 4s$ ), a sufficiently sparse signal  $\mathbf{q}$  can be exactly recovered from the limited dataset without any prior knowledge of the number of nonzero elements, their locations, and their values. On the other hand, if the condition is not satisfied, exact recovery cannot be guaranteed [17,34]. However, it is possible to use the property of coherence to guide biological experiment design, basically to collect a more informative dataset. As we will discuss in this paper, this can be done by inhibiting or stimulating certain genes to manipulate the gene expression.

### CS can help reconstruct GRNs

In graph theory, a digraph can be represented by  $G = (V, E)$  where  $V$  and  $E$  represent nodes and edges respectively. For GRNs, each node represents a gene and each

edge represents an influence map which models how genes affect each other. For example, the interactions could be how genes inhibit or stimulate each other. Since the connectivities of GRNs are typically unknown, often the best we can do is to select a set of possible candidate functions encoding possible unknown connectivities between genes.

In this section, we formulate a data-driven network identification problem into CS framework: first, we define a dynamical model of gene regulatory network. Then, we encode system dynamics into the sensing matrix ( $\Omega$ ) and denote unknown connectivities between genes by  $\mathbf{q}$ , a signal to be recovered.

### A dynamical model of gene regulatory networks (GRNs)

We consider a dynamical system described by:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{u} + \mathbf{g}(\mathbf{x}_h) + \mathbf{w} \quad (4)$$

where  $\mathbf{x} \in \mathbb{R}^n$  denotes the concentrations of the rate-limiting species which can be measured in experiments;  $\dot{\mathbf{x}} = [\dot{x}_1 \ \dot{x}_2 \ \dots \ \dot{x}_n]^\top \in \mathbb{R}^n$  is a vector whose elements are the change in concentrations of the  $n$  species over time;  $\mathbf{f}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^n$  represents biochemical reactions, such as those governed by mass action kinetics, Michaelis-Menten, or Hill kinetics. Thus,  $\mathbf{f}(\cdot)$  can include functions of known form such as product of monomials, monotonically increasing or decreasing Hill functions, simple linear terms, and constant terms [24].  $\mathbf{u} \in \mathbb{R}^n$  denotes the control input which could represent inhibitions and stimulations; For example, if we consider protein-protein interaction,  $\mathbf{u}$  represents drug-induced perturbation such as inhibition and stimulation. For a GRN, we consider the use of siRNA as a control input to the system for gene knockdown;  $\mathbf{g}(\cdot) : \mathbb{R}^{n_h} \rightarrow \mathbb{R}^n$  represents influence from hidden nodes  $\mathbf{x}_h \in \mathbb{R}^{n_h}$ , which cannot be measured in experiments;  $n_h$  is the number of hidden nodes and unknown; and  $\mathbf{w} \in \mathbb{R}^n$  represents energy-bounded process noise or measurement noise. Here  $\mathbf{x}$ ,  $\dot{\mathbf{x}}$  and  $\mathbf{u}$  is assumed to be known where  $\mathbf{x}$  can be measured in experiments and  $\dot{\mathbf{x}}$  may not be measured directly in experiments but we could calculate these quantities by interpolating  $\mathbf{x}$  and using numerical derivatives. In [36], the authors pointed out that although data on time derivative can be difficult to obtain especially in the presence of noise, it is possible to estimate the gene expressions relatively accurately by repeating measurement with careful instrumentation and statistics [37].

Since we do not know whether important nodes in the gene regulatory network are missing, how many missing nodes there are, and how they affect system dynamics (i.e.,  $\mathbf{x}_h$ ,  $n_h$  and  $\mathbf{g}(\mathbf{x}_h)$ ), we denote a vector of (unknown)

influence from hidden nodes' dynamics by  $\mathbf{v}(\triangleq \mathbf{g}(\mathbf{x}_h))$ ; Also, without loss of generality, since we know  $\mathbf{u}$ , we define  $\mathbf{y}$  as follows:

$$\mathbf{y} \triangleq \dot{\mathbf{x}} - \mathbf{u} = \mathbf{f}(\mathbf{x}) + \mathbf{v} + \mathbf{w} \quad (5)$$

### Formulating a dynamical system as a GRN

The nonlinear function  $\mathbf{f}(\mathbf{x})$  can be decomposed into a linear sum of scalar basis functions,  $f_{b,i}(\mathbf{x}) \in \mathbb{R}$ , where we select the set of possible candidate basis functions:

$$\mathbf{f}(\mathbf{x}) = \sum_{i=1}^N f_{b,i}(\mathbf{x}) \begin{bmatrix} q_{i1} \\ q_{i2} \\ \dots \\ q_{in} \end{bmatrix} \quad (6)$$

where  $N$  is the number of possible candidate basis functions and  $q_{ij}$  are unknown parameters which reflect underlying structure, i.e., influence of  $f_{b,i}(\mathbf{x})$  on the  $j$ -th state ( $\dot{x}_j$ ). Typically, we may choose a larger set than necessary, and allow the CS method to indicate the importance of each function, as we shall describe. Thus the Equation (5) can be written as follows:

$$\mathbf{y} = \mathbf{S}_q \mathbf{F}_b(\mathbf{x}) + \mathbf{v} + \mathbf{w} \quad (7)$$

where  $\mathbf{S}_q = \begin{bmatrix} q_{11} & \dots & q_{N1} \\ \dots & \dots & \dots \\ q_{1n} & \dots & q_{Nn} \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1^\top \\ \dots \\ \mathbf{q}_N^\top \end{bmatrix} \in \mathbb{R}^{n \times N}$  reflects the (unknown) underlying GRN structure,  $\mathbf{q}_i = [q_{i1} \ q_{i2} \ \dots \ q_{in}]^\top \in \mathbb{R}^n$  and  $\mathbf{F}_b(\mathbf{x}) = [f_{b,1}(\mathbf{x}), \dots, f_{b,N}(\mathbf{x})]^\top \in \mathbb{R}^N$  is the vector field which includes possible candidate basis functions. In this way, any biochemical reactions can be represented by a linear map  $\mathbf{S}_q$  and a function  $\mathbf{F}_b(\mathbf{x})$  where  $\mathbf{S}_q$  encodes the underlying graph structure and  $\mathbf{F}_b(\mathbf{x})$  includes all possible candidate functions that could be included in the biochemical reactions.

In practice, we can construct  $\mathbf{F}_b(\mathbf{x})$  by selecting the most commonly used candidate basis functions to model GRNs or protein-protein interaction, for example, all monomials, binomials, other combinations and Hill function.

**Example 1.** Consider the simple nonlinear ordinary differential equations (ODEs):

$$\begin{aligned} y_1 &= \dot{x}_1 = \gamma_1 x_1 + k_1 \frac{x_2^{n_{act}}}{1 + x_2^{n_{act}}} \\ y_2 &= \dot{x}_2 = \gamma_2 x_2 + k_2 \frac{1}{1 + x_3^{n_{inh}}} \\ y_3 &= \dot{x}_3 = \gamma_3 x_3 + k_3 \frac{1}{1 + x_1^{n_{inh}}} \end{aligned} \quad (8)$$

where  $x_i$  denotes the concentration of the  $i$ -th species,  $\dot{x}_i$  is the change in concentration of the  $i$ -th species, the  $\gamma_i$  denotes protein decay rate, the  $k_i$  denotes the maximum promoter/inhibitor strength. Here, there is no input ( $\mathbf{u} = \mathbf{0}$ ), no hidden node ( $\mathbf{v} = \mathbf{0}$ ) and no process noise ( $\mathbf{w} = \mathbf{0}$ ). Also,  $n_{act}$  represents positively cooperative binding (activation) and  $n_{inh}$  represents negative cooperative binding (inhibition). The set of ODEs corresponds to a topology where gene 1 is activated by gene 2, gene 2 is inhibited by gene 3, and gene 3 is inhibited by gene 1 as shown in Figure 1(A). We can write Equation (8) as follows:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \underbrace{\begin{bmatrix} \gamma_1 & 0 & 0 & 0 & k_1 & 0 \\ 0 & \gamma_2 & 0 & 0 & 0 & k_2 \\ 0 & 0 & \gamma_3 & k_3 & 0 & 0 \end{bmatrix}}_{\mathbf{S}_q} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \frac{1}{1+x_1^{n_{inh}}} \\ \frac{x_2^{n_{act}}}{1+x_2^{n_{act}}} \\ \frac{1}{1+x_3^{n_{inh}}} \end{bmatrix} = \mathbf{S}_q \mathbf{F}_b(\mathbf{x})$$

where  $\mathbf{S}_q \in \mathbb{R}^{3 \times 6}$  represents the influence map.

We can also consider a version of Equation (8) in which there exists a hidden node ( $x_h$ ) affecting ( $x_1$ ) as shown in Figure 1(B), as well as process noise.

$$\mathbf{y} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \mathbf{S}_q \mathbf{F}_b(\mathbf{x}) + \mathbf{v} + \mathbf{w}$$

where  $\mathbf{v} = [v_1, 0, 0]^T$  and  $\mathbf{w} = [w_1, w_2, w_3]^T$ .

### Formulating GRN into the CS framework

Suppose the time series data are sampled from a real experimental system at discrete time points  $t_k$ . By taking the transpose of both sides of Equation (7), we obtain

$$\mathbf{y}(t_k)^T = \mathbf{F}_b^T(\mathbf{x}(t_k)) \mathbf{q} + \mathbf{v}(t_k)^T + \mathbf{w}(t_k)^T \tag{9}$$

where  $\mathbf{q} = \mathbf{S}_q^T = [\mathbf{q}_1 \ \mathbf{q}_2 \ \dots \ \mathbf{q}_n] \in \mathbb{R}^{N \times n}$ . Assuming that  $M$  successive data points are sampled, then define:

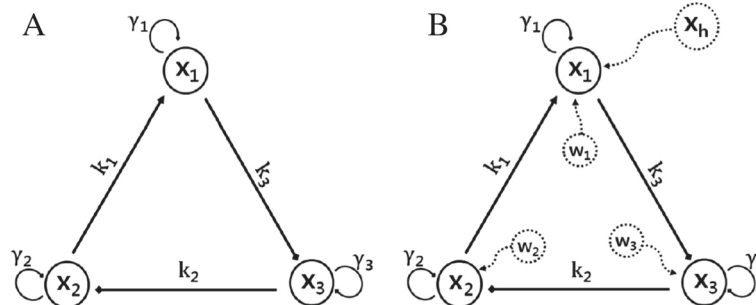
$$\begin{aligned} \mathbf{y} \triangleq [\mathbf{y}_1 \ \dots \ \mathbf{y}_n] &= \begin{bmatrix} y_1(t_1) & \dots & y_n(t_1) \\ y_1(t_2) & \dots & y_n(t_2) \\ \dots & \dots & \dots \\ y_1(t_M) & \dots & y_n(t_M) \end{bmatrix} \in \mathbb{R}^{M \times n} \\ \Phi &\triangleq \begin{bmatrix} f_{b,1}(\mathbf{x}(t_1)) & \dots & f_{b,N}(\mathbf{x}(t_1)) \\ f_{b,1}(\mathbf{x}(t_2)) & \dots & f_{b,N}(\mathbf{x}(t_2)) \\ \dots & \dots & \dots \\ f_{b,1}(\mathbf{x}(t_M)) & \dots & f_{b,N}(\mathbf{x}(t_M)) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{F}_b^T(\mathbf{x}(t_1)) \\ \mathbf{F}_b^T(\mathbf{x}(t_2)) \\ \dots \\ \mathbf{F}_b^T(\mathbf{x}(t_M)) \end{bmatrix} \in \mathbb{R}^{M \times N} \\ [\mathbf{v}_1 \ \dots \ \mathbf{v}_n] &\triangleq \begin{bmatrix} \mathbf{v}(t_1)^T \\ \mathbf{v}(t_2)^T \\ \dots \\ \mathbf{v}(t_M)^T \end{bmatrix} \in \mathbb{R}^{M \times n} \\ [\mathbf{w}_1 \ \dots \ \mathbf{w}_n] &\triangleq \begin{bmatrix} \mathbf{w}(t_1)^T \\ \mathbf{w}(t_2)^T \\ \dots \\ \mathbf{w}(t_M)^T \end{bmatrix} \in \mathbb{R}^{M \times n} \end{aligned} \tag{10}$$

This leads to  $n$  independent equations:

$$\mathbf{y}_i = \Phi \mathbf{q}_i + \mathbf{v}_i + \mathbf{w}_i \quad (i = 1, \dots, n) \tag{11}$$

where  $\mathbf{y}_i = [y_i(t_1), y_i(t_2), \dots, y_i(t_M)]^T \in \mathbb{R}^M$  represents the  $M$  successive data points,  $\Phi \in \mathbb{R}^{M \times N}$  consists of  $N$  possible candidate bases which are functions of given time series data  $\mathbf{x}$  and  $\mathbf{q}_i \in \mathbb{R}^N$  represents the unknown influence map corresponding to the  $i$ -th species. Since a biochemical reaction network is typically sparse, as a consequence,  $\mathbf{q}_i$  is sparse and we have  $N \gg M$  for  $\Phi$  because we assume the limited time series data and may choose a larger set of basis functions than necessary.

Although we formulate  $n$  independent linear regression problems in Equation (11), we consider  $n$  independent



**Figure 1** A graph representation of nonlinear ODEs. (A) (Example 1,  $n = 3$ ,  $\mathbf{S}_q \in \mathbb{R}^{3 \times 6}$ ): among 18 ( $= 3 \times 6$ ) components, only 6 components are non-zero (B) (Example 1 with hidden node and measurement noise) there exists hidden node  $\mathbf{x}_h$ , affecting  $\mathbf{x}_1$ , and process noise  $\mathbf{w}$ .

equations in Equation (11) together by stacking  $\mathbf{y}_i$  as follows:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \dots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \Phi & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \Phi & \dots & \mathbf{O} \\ \dots & \dots & \dots & \dots \\ \mathbf{O} & \mathbf{O} & \dots & \Phi \end{bmatrix} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \dots \\ \mathbf{v}_n \end{bmatrix} + \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \dots \\ \mathbf{w}_n \end{bmatrix}$$

$$\triangleq \mathbf{\Omega} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \mathbf{v} + \mathbf{w} \tag{12}$$

Now, Equation (12) is in the form of the CS formulation in Equation (1) where  $\mathbf{Y} \in \mathbb{R}^{n \times M}$  is the measurement,  $\mathbf{\Omega} \in \mathbb{R}^{n \times M \times n \times N}$  is the sensing matrix which consists of basis functions for the given time series data,  $\mathbf{v} \triangleq [\mathbf{v}_1; \mathbf{v}_2; \dots; \mathbf{v}_n] \in \mathbb{R}^{n \times M}$  represents possibly large corruption from hidden nodes and  $\mathbf{w} \triangleq [\mathbf{w}_1; \mathbf{w}_2; \dots; \mathbf{w}_n] \in \mathbb{R}^{n \times M}$  represents process or measurement noise.

In this paper, we make the following assumptions:

**Assumption 1.** We consider two cases:

- (ideal case) We assume that we can measure all states  $\mathbf{x}$  (i.e., there is no hidden node ( $\mathbf{x}_h$ ), thus  $\mathbf{v} = \mathbf{0}$ ) and there is no measurement noise ( $\mathbf{w} = \mathbf{0}$ ). Also, there are enough columns in  $\Phi$  in Equation (11) to represent the underlying system in Equation (6).
- (extension) We consider hidden node ( $\mathbf{x}_h$ ) where  $\mathbf{v}(= \mathbf{g}(\mathbf{x}_h))$  is assumed to be sparse, and process noise ( $\mathbf{w} \neq \mathbf{0}$ ). Here, we can also consider the case where the columns (also known as the dictionaries in the CS literature [28]) of  $\Phi$  may not be able to represent the underlying system. In this case, we consider the influence from the missing dictionaries as  $\mathbf{v}$ .

We first consider the ideal case for simplicity in explaining the main results, and then extend the proposed method to the more general case.

**Results and discussion**

**Formulating GRN identification problem into CS**

Many existing algorithms [16,24] consider the  $n$  independent linear regression problems in Equation (11) separately. Since the columns of matrix  $\Phi$  are composed of time series data as in Equation (10), it is difficult to *a priori* guarantee low correlation and sometimes  $\Phi$  even suffers rank deficiency. Also, Pan *et al.* [24] pointed out that correlation between the columns of  $\Phi$  is usually high ( $\mu(\Phi)$  is close to 1).

Intuitively, if two columns of the sensing matrix are highly correlated, it is hard to distinguish the corresponding components in the sparse signal  $\mathbf{q}$  (note that the measurement  $\mathbf{Y}$  is a linear combination of each column of the sensing matrix with the components in  $\mathbf{q}$  as coefficients). In order to deal with high coherence in Equation (11), many method combines CS with different techniques such as Bayesian formulation [24], Kalman filter [25], and Granger causality [26]. Also, it is a well known problem in the lasso formulation of network inference: if there is high coherence in the sensing matrix, one can use an elastic net which combines the  $l_1$  and  $l_2$  norms. Although each reconstruction result [24-26] might be the optimal solution in the sense of its formulation (i.e., maximum likelihood), the identified graph may not represent the underlying GRN. In other words, if the data set is not informative enough to fully explore the underlying system, while the identified graph structure based on the given data set may be an optimal solution of the particular optimization problem, it may not represent the true system.

In this paper, our goal is to get the smallest data informative enough to recover the underlying graph structure exactly. Since the proposed method maintains the CS framework by reducing coherence of the sensing matrix, the method is fundamentally different than any other methods which make use of different techniques in conjunction with the  $l_1$  optimization [24-26] which leads to a sparse representation of the network. Hence, we use all the properties of CS in order to access the ability to exactly reconstruct the underlying graph structure, reveal deficiencies in the data and model, and design new experiments to remedy the deficiencies if necessary.

While maintaining the CS framework, in order to deal with high coherence, we formulate Equation (12) instead of Equation (11) since we have strongly uncorrelated columns in  $\mathbf{\Omega}$ . In other words, since  $\mathbf{\Omega}$  has many independent columns, we have more degrees of freedom to reduce coherence of  $\mathbf{\Omega}$ . We will show that by using a transformation, the components of the sensing matrix can be made more uniformly distributed so that we could reduce coherence.

Moreover, since each  $\mathbf{q}_i$  has different degrees of sparsity in general, if we consider  $n$  independent equations in Equation (11),  $\Phi$  should satisfy the incoherence condition  $M \geq 2 \max_i (\|\mathbf{q}_i\|_0)$  stated in **Proposition 1**. On the other hand,  $\mathbf{\Omega}$  only needs to satisfy the condition  $M \cdot n \geq 2 \cdot \sum_{i=1}^n \|\mathbf{q}_i\|_{l_0}$ . Since the averaged sparsity  $\left( = \frac{\sum_{i=1}^n \|\mathbf{q}_i\|_{l_0}}{n} \leq \max_i (\|\mathbf{q}_i\|_{l_0}) \right)$  is smaller than  $\max_i (\|\mathbf{q}_i\|_{l_0})$ , we can reduce the required number of samples ( $M$ ). Also, in case of rank deficiency, we can simply remove the corresponding rows (note that the rank deficiency is more likely caused by the row since  $N \gg M$ ).

### Optimal design of sensing matrix

#### Reducing coherence by transformation

In order to reduce coherence, first we rearrange  $\Omega$  with respect to a spatial information, and then we consider the transformation in order to reduce coherence (see **Method: Rearranging the sensing matrix** for mathematical details):

$$\mathbf{Z} \triangleq \begin{bmatrix} \bar{\mathbf{z}}(1) \\ \bar{\mathbf{z}}(2) \\ \dots \\ \bar{\mathbf{z}}(M) \end{bmatrix} = \begin{bmatrix} \Psi_1^s(\mathbf{I}_n \otimes \Phi_{1,:}) \\ \Psi_2^s(\mathbf{I}_n \otimes \Phi_{2,:}) \\ \dots \\ \Psi_M^s(\mathbf{I}_n \otimes \Phi_{M,:}) \end{bmatrix} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} \begin{bmatrix} \bar{\mathbf{v}}(1) \\ \bar{\mathbf{v}}(2) \\ \dots \\ \bar{\mathbf{v}}(M) \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{w}}(1) \\ \bar{\mathbf{w}}(2) \\ \dots \\ \bar{\mathbf{w}}(M) \end{bmatrix} \triangleq \bar{\Omega} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \bar{\mathbf{v}} + \bar{\mathbf{w}} \quad (13)$$

where  $\otimes$  represents the Kronecker product<sup>a</sup>,  $\bar{\Omega} \in \mathbb{R}^{M \cdot n \times N \cdot n}$  represents the rearranged sensing matrix multiplied by the transformation  $\Psi_j^s$ , and  $\bar{\mathbf{z}}(j)$ ,  $\bar{\mathbf{v}}$  and  $\bar{\mathbf{w}}$  are defined in Equation (20). We want to find  $\Psi_j^s$  ( $j = 1, \dots, M$ ) to minimize

$$\min \mu(\bar{\Omega}) \text{ where } \bar{\Omega} = \begin{bmatrix} \psi_{11}^s \Phi_{1,:} & \psi_{12}^s \Phi_{1,:} & \dots & \psi_{1n}^s \Phi_{1,:} \\ \psi_{21}^s \Phi_{2,:} & \psi_{22}^s \Phi_{2,:} & \dots & \psi_{2n}^s \Phi_{2,:} \\ \dots & \dots & \dots & \dots \\ \psi_{M1}^s \Phi_{M,:} & \psi_{M2}^s \Phi_{M,:} & \dots & \psi_{Mn}^s \Phi_{M,:} \end{bmatrix} \quad (14)$$

has full row rank

where  $\psi_{ij}^s \in \mathbb{R}^n$  represents the  $j$ -th column of  $\Psi_i^s$ . In this paper, we propose a heuristic approach and a novel way to find  $\Psi_i^s$  by solving the optimization iteratively to reduce coherence (see Method: Section ‘Randomly chosen matrix  $\Psi$ ’ and ‘Finding the optimal transformation’ for details).

**Example 2.** (reducing coherence for a linear system) consider a simple linear system ( $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ ) where  $n = 5$ ,  $N = 5$ ,  $M = 4$ ,  $s = 10$  (note that  $n \cdot M \geq 2s$ ), and the elements of  $\mathbf{A}$  are randomly chosen such that there is no isolated node. Figure 2 shows the sensing matrix for both  $\Omega$  (top left) and the transformed sensing matrix (top right, denoted by  $\Omega_\Psi$ ). By reducing coherence, the components of the transformed sensing matrix (top right) are more uniformly distributed and the coherence is reduced by up to 0.6 although  $\mu(\Omega)$  is close to 1 (bottom left) in Figure 2(A). Also, Figure 2(B) shows the result of the inferred graph structure based on given time series data without any a priori information where the  $x$ -axis represents indices of the influence map (i.e., the  $1_{st}, 2_{nd}, \dots, n_{th}$  rows of influence map  $\mathbf{S}_q$ ; note that for a linear system,  $\mathbf{S}_q = \mathbf{A}$ ). Here, there are 5 states ( $n = 5$ ) in a linear system so the influence map  $\mathbf{A}$  has 25 elements. Although L1 and L2<sup>b</sup>

norm minimizations fail to recover the exact signal, CS in Equation (17) (see **Recovery of gene regulatory networks** for details) recovers the exact signal (bottom right) by reducing coherence of the sensing matrix. Note that (L1) solves the  $n$  independent equations in Equation (11) without reducing coherence and (L2) solves Equation (12) with  $l_2$ -regularization.

**Example 3.** (reducing coherence for a nonlinear system,  $n = 3$ ,  $N = 9$ ,  $M = 5$ ,  $s = 6$ ) consider simple nonlinear ODEs as follows:

$$\begin{aligned} \dot{x}_1 &= \gamma_1 x_1 + k_{+12} \frac{x_2^{n_{act}}}{1 + x_2^{n_{act}}} \\ \dot{x}_2 &= \gamma_2 x_2 + k_{-23} \frac{1}{1 + x_3^{n_{ihb}}} \\ \dot{x}_3 &= \gamma_3 x_3 + k_{+13} \frac{x_1^{n_{act}}}{1 + x_1^{n_{act}}} \end{aligned} \quad (15)$$

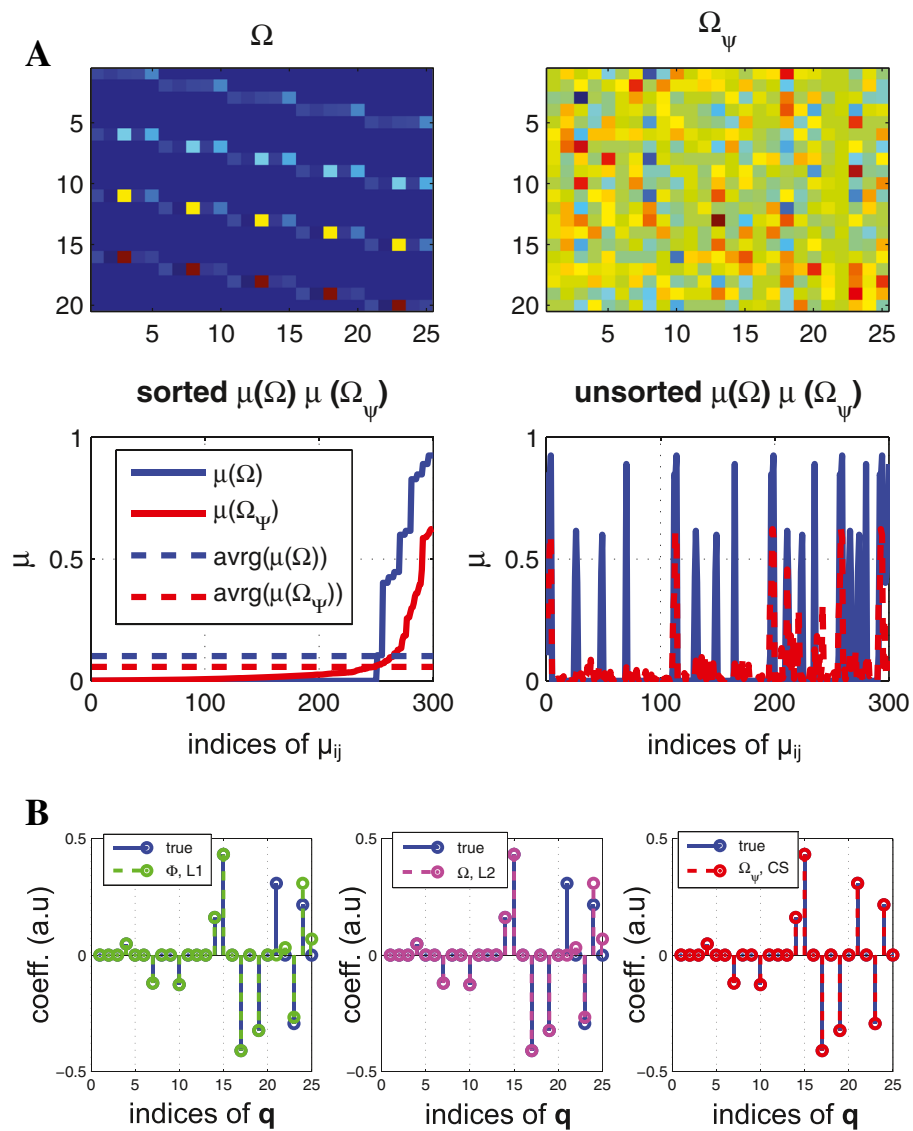
where  $(\gamma_1, \gamma_2, \gamma_3) = (-0.3, -0.25, -0.35)$ ,  $k_{+12} = 1.2$ ,  $k_{+13} = 0.9$ ,  $k_{-23} = 2.2$ ,  $n_{act} = 4$  (activation Hill coefficient) and  $n_{ihb} = 4$  (inhibition Hill coefficient). The set of ODEs corresponds to a topology ( $x_2 \rightarrow x_1 \rightarrow x_3 \dashv x_2$ ) as shown in Figure 3(A). Figure 3(B) shows that we can reduce the coherence by up to 0.8 and (C) shows that only CS recovers the exact graph structure, and L2 regularization does not encourage sparsity but distributes the coefficients to be more similar to each other.

#### Designing effective experiments

Consider the case where the sensing matrix is not incoherent. If the coherence condition  $(M \cdot n > c\mu(\bar{\Omega})^2 \cdot s \log(N \cdot n))$  is not satisfied, exact recovery cannot be guaranteed [34]. We use the transformation ( $\Psi_i^s$  or  $\mathbf{P} = \mathcal{P}^\top \mathcal{P}$  in Equation (24); see Method section ‘Reducing coherence by transformation’) in order to reduce the coherence but obviously, sometimes we might have inherent limits to how much the coherence can be reduced. There are possible reasons:

- Since we solve the relaxed problem in Equation (24) iteratively,  $\mathbf{P}$  might be sub-optimal.
- If the time series data of two different gene expressions,  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are highly correlated, it might be difficult to reduce coherence. In this case, we need to design a new experiment to remedy deficiencies in the data.

As we mentioned, the incoherence of the sensing matrix can be used not only as a good metric to guarantee exact recovery but also as a guideline for designing new experiments. For example, from the coherence distribution, we can identify which columns of the sensing matrix have high coherence, i.e.,  $f_{b,i}(\mathbf{x})$  and  $f_{b,j}(\mathbf{x})$  in Equation (6). Intuitively, in order to reduce ambiguities from the highly



**Figure 2 Reducing coherence for a linear system.** (Example 2) (A) the sensing matrix  $\Omega \in \mathbb{R}^{20 \times 25}$ , without (top left) and with transformation  $\Psi$  (top right) and the corresponding coherence distribution (bottom) (B) reconstruction result (left) L1 (middle) L2 with  $\Omega$  (right)  $l_1$  optimization with  $\Omega_\Psi$  (CS) where x-axis represents indices of  $q \in \mathbb{R}^{25}$  and y-axis represents coefficient of  $q$  (a.u). By reducing coherence of the sensing matrix with transformation, we can recover the exact structure in (B) (right). However, both L1 and L2 fail to recover the exact structure.

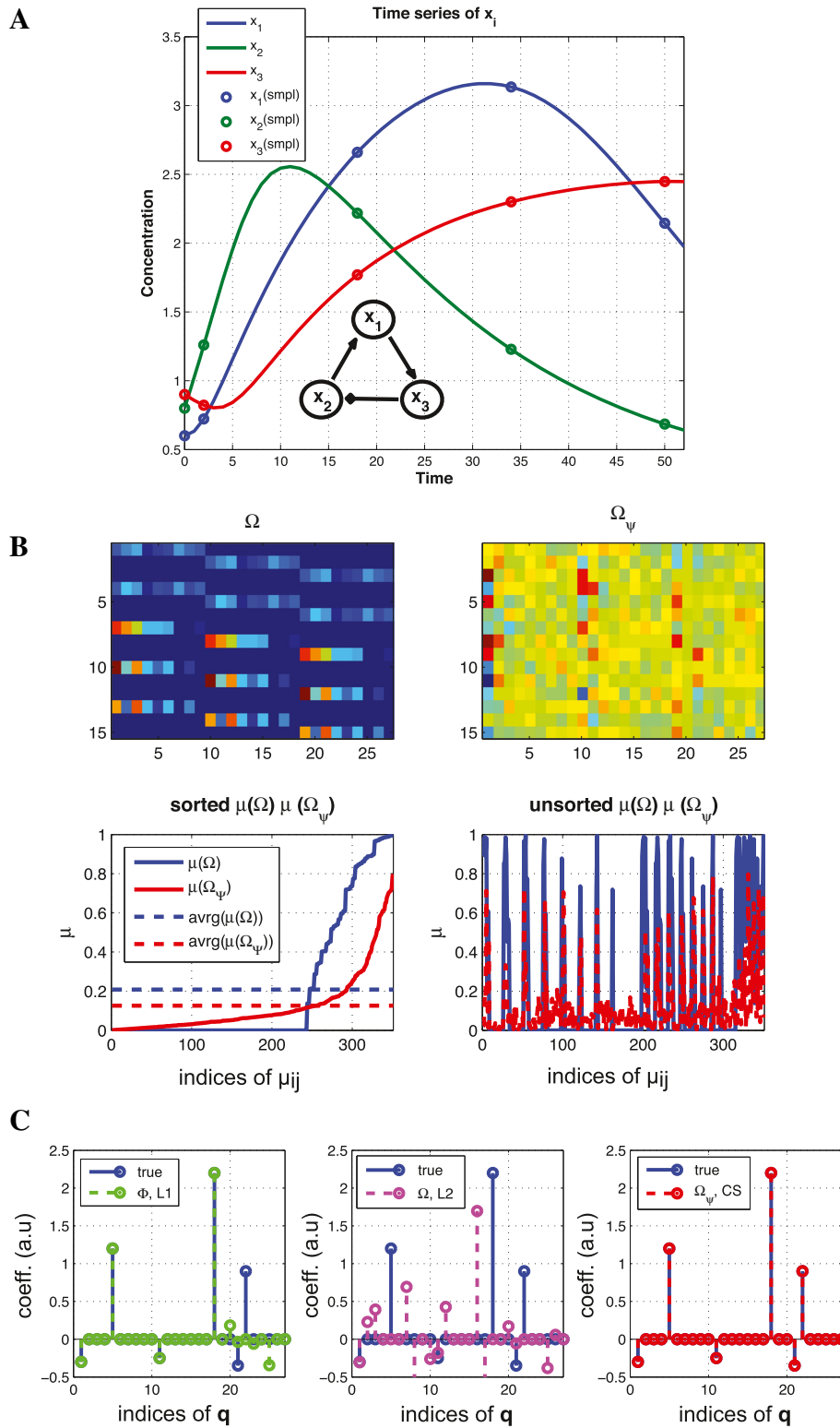
correlated columns of the sensing matrix, we should perturb either  $f_{b,i}(x)$  or  $f_{b,j}(x)$ . Thus, it is possible to use this property of coherence to guide biological experiment design, to collect a more informative dataset. By doing this, we can minimize the number of experiments and reduce the cost of experiments.

**Example 4.** (limitation of reducing coherence by  $\Psi$ , linear dynamics,  $n = 5, N = 5, M = 4, s = 10$ ) In Figure 4, Exp#1 represents the original experimental data set which has the limitation of reducing coherence by  $\Psi$ . Since we consider linear dynamics, i.e.,  $f_{b,i}(x) = x_i$  and  $f_{b,j}(x) = x_j$ ,

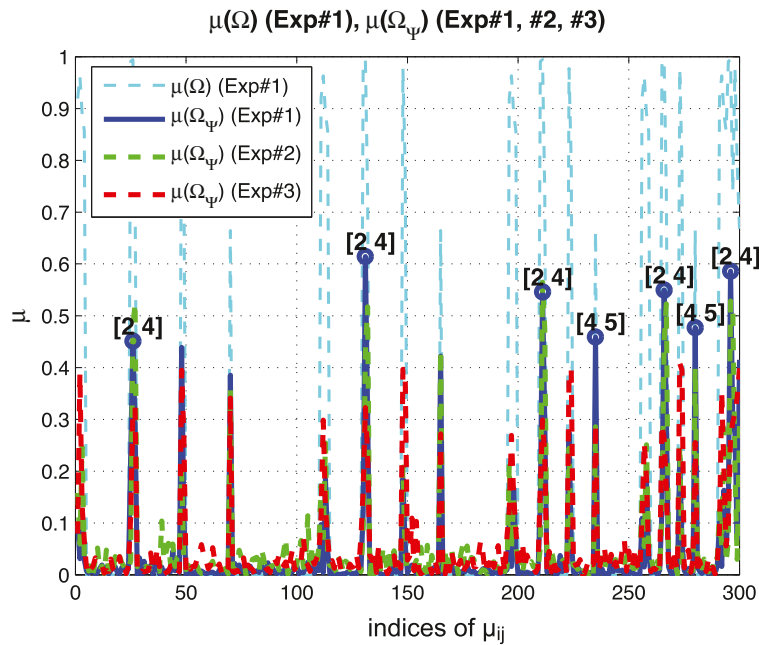
we found that  $x_2$  and  $x_4$  cause high coherence as shown in Figure 4 (circle marker). Thus, in order to reduce this high coherence, we should perturb either  $x_2$  or  $x_4$ .

To show the effectiveness of the new experiment, we design two different experiment sets and compare the reconstruction results with each other; for Exp#2, we perturb  $x_3$  and for Exp#3, we perturb  $x_2$ . As we mentioned earlier, intuitively, we expect that Exp#3 to be a more informative experiment to identify the graph structure since we would like to reduce the coherence between  $x_2$  and  $x_4$ . As we expected, in Exp#3, we can reduce the coherence more than that of Exp#2 as shown in Figure 4, and recover the exact





**Figure 3 Reducing coherence for a nonlinear system.** (Example 3) **(A)** Time series of  $x_1, x_2, x_3$  for model in Equation (15) **(B)** the sensing matrix without (top left) and with transformation  $\Psi$  matrix (top right) and the corresponding coherence distribution (bottom) **(C)** reconstruction results. By constructing the sensing matrix  $\Omega \in \mathbb{R}^{M \times n \times N}$  where  $n = 3, N = 9, M = 5$  and  $s = 6$ , we recover the  $s$ -sparse signal in **(B)** (right). However, both  $L1$  and  $L2$  fail to recover the exact structure  $\mathbf{q}$ .



**Figure 4** Guideline for designing new experiment based on coherence distribution. (Example 4) coherence distribution of the original  $\Omega$  based on Exp#1 show that  $x_2$  and  $x_4$  cause high coherence. Thus, in order to reduce this high coherence, we should perturb either  $x_2$  or  $x_4$ . We design two different experiment sets; for Exp#2, we perturb  $x_3$  and for Exp#3, we perturb  $x_2$ . As we expected, Exp#3 is a more effective experiment to reduce the coherence between  $x_2$  and  $x_4$  and we recover the exact graph structure as shown in Figure 5. On the other hand, the coherence of Exp#2 remains almost the same as that of Exp#1 and we fail to recover the exact graph structure in Figure 5.

graph structure as shown in Figure 5. On the other hand, the coherence of Exp#2 remains almost the same as that of Exp#1 and we fail to recover the exact graph structure. This numerical example illustrates that by using the property of coherence, we can guide biological experiment design more effectively.

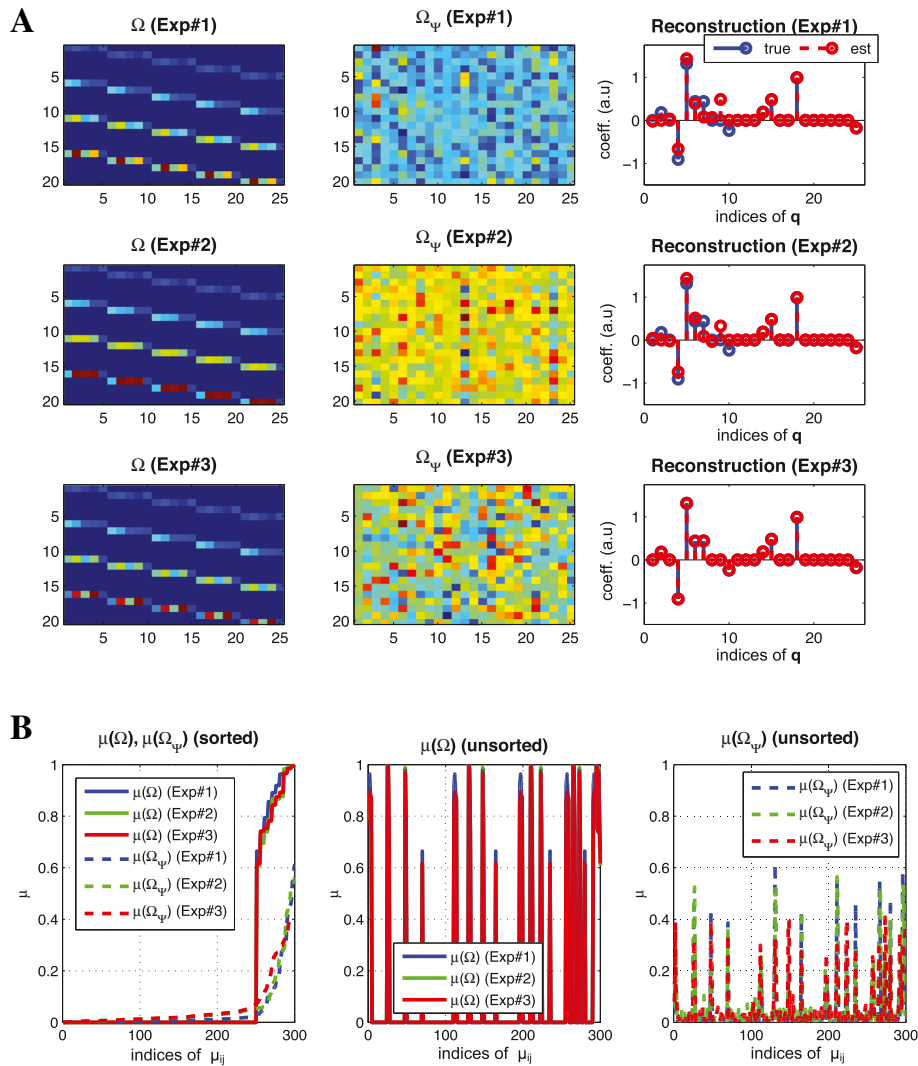
**Example 5.** (limitation of reducing coherence by  $\Psi$ , nonlinear dynamics,  $n = 3, N = 9, M = 5, s = 9$ ) consider the following set of ODEs:

$$\begin{aligned} \dot{x}_1 &= \gamma_1 x_1 + k_{+12} \frac{x_2^{n_{act}}}{1 + x_2^{n_{act}}} + k_{+13} \frac{x_3^{n_{act}}}{1 + x_3^{n_{act}}} \\ \dot{x}_2 &= \gamma_2 x_2 + k_{-21} \frac{1}{1 + x_1^{n_{ihb}}} + k_{-23} \frac{1}{1 + x_3^{n_{ihb}}} \\ \dot{x}_3 &= \gamma_3 x_3 + k_{-31} \frac{1}{1 + x_1^{n_{ihb}}} + k_{-32} \frac{1}{1 + x_2^{n_{ihb}}} \end{aligned}$$

where  $(\gamma_1, \gamma_2, \gamma_3) = (-0.25, -0.23, -0.26)$ ,  $k_{+12} = 1.2$ ,  $k_{+13} = 1.25$ ,  $k_{-21} = 2.8$ ,  $k_{-23} = 2.1$ ,  $k_{-31} = 2.7$ ,  $k_{-32} = 1.8$ ,  $n_{act} = 4$  (activation Hill coefficient) and  $n_{ihb} = 4$  (inhibition Hill coefficient). The corresponding topology is shown in Figure 6(B) (note that we intentionally

choose the symmetric structure and similar parameters). The reconstruction error using Exp#1 data is shown in Figure 6(A) (left, bottom) and the reconstruction error illustrates difficulties of resolving ambiguities from  $x_2$  and  $x_3$ . This can be captured by the coherence distribution of the sensing matrix based on the Exp#1 dataset; the correlation between the columns corresponding to  $x_1$  and  $x_2$  is close to the correlation between the columns corresponding to  $x_1$  and  $x_3$ . Based on the coherence distribution, we design two trials; for Exp#2, we perturb  $x_1$  and for Exp#3, we perturb  $x_2$ . As we expected, Exp#2 is not an effective experiment in terms of information. On the other hand, by using Exp#3, we can reduce both the maximum coherence and the averaged coherence, and reconstruct the exact graph structure as shown in Figure 6(A).

Both Examples 4 and 5 illustrate that if the transformed sensing matrix is not incoherent enough to guarantee exact recovery, we can design a new experiment based on the distribution of coherence. Also, we show that the proposed experiment can help to reduce coherence more and thus reconstruct the exact graph structure. For a fair comparison, we use the same number of time points as  $M$  here. However, in practice, we can also stack all the experimental data sets together if we assume that the linear map  $S_q$



**Figure 5 Design effective experiment for a linear system.** (Example 4) Reconstruction results based on different experiments: Exp#1 represents the original experimental dataset which has limitation to reduce coherence; Exp#2 represents non-effective experimental dataset; and Exp#3 represents the effective experimental dataset **(A)** the sensing matrix and reconstruction result **(B)** coherence distribution comparison. This example demonstrates that if there is a limitation to reduce coherence in the original dataset, we can also design a more effective experiments based on the coherence distribution. By revealing deficiency in the data, the proposed method can help reduce the cost of experiments.

and the set of basis functions  $F_b(x)$  does not change for different experiment:

$$\begin{bmatrix} Z_1 \\ \dots \\ Z_k \end{bmatrix} = \begin{bmatrix} \bar{\Omega}_1 \\ \dots \\ \bar{\Omega}_k \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ \dots \\ q_n \end{bmatrix} + \begin{bmatrix} \bar{v}_1 \\ \dots \\ \bar{v}_k \end{bmatrix} + \begin{bmatrix} \bar{w}_1 \\ \dots \\ \bar{w}_k \end{bmatrix} \quad (16)$$

where the subscript  $Z_i, \bar{\Omega}_i, \bar{v}_i, \bar{w}_i$  represents the  $i$ -th experiment. As the number of measurements increase ( $M$ ), one may be able to reduce the coherence. However, one can reduce the coherence only if the additional measurements provide us more useful information. As a trivial example, one could stack exactly the same data on top of the first,

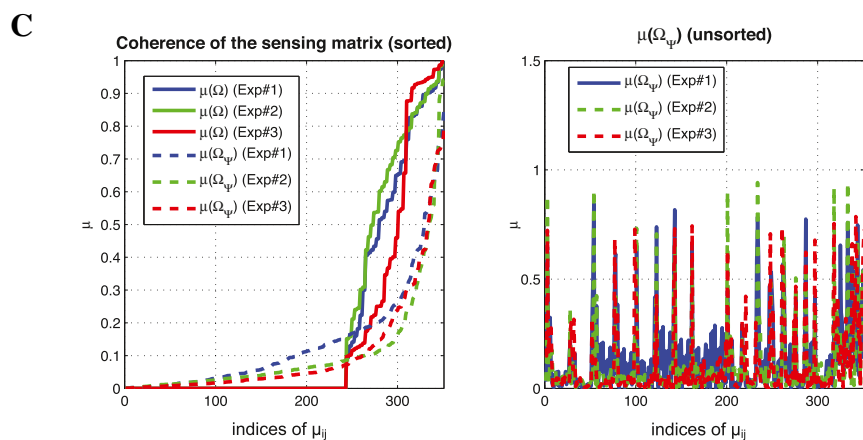
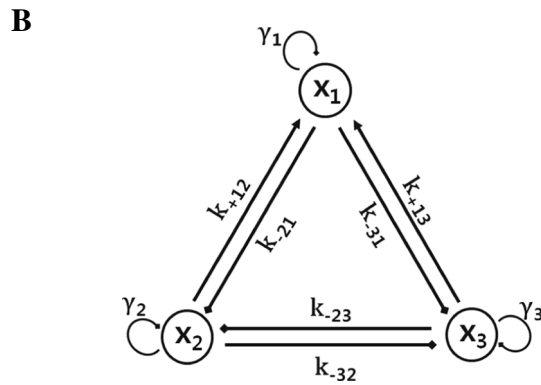
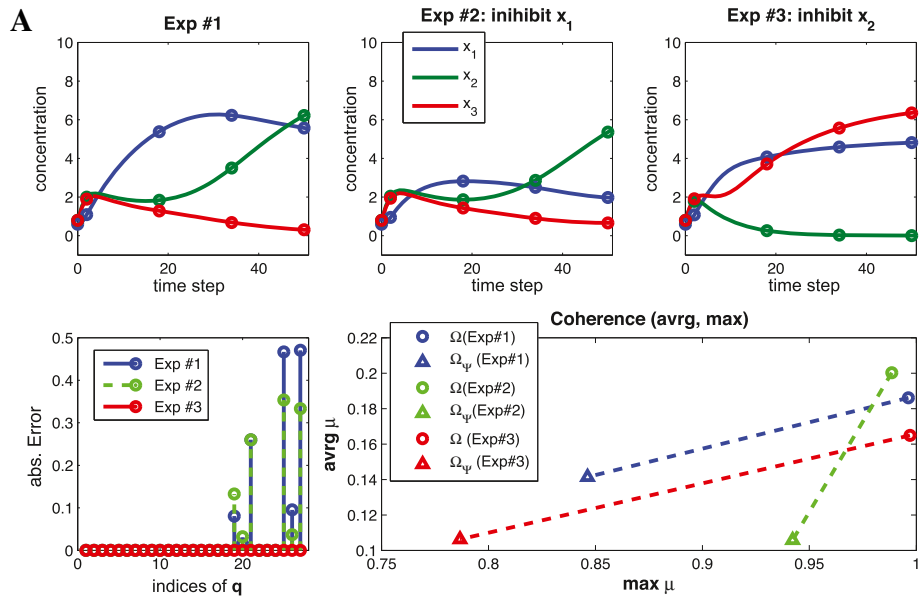
and increase  $M$  to  $2M$ , however the coherence is exactly the same as that of the original dataset.

### Recovery of gene regulatory networks

In this section, we present reconstruction of the exact graph structure and show how the condition for exact recovery will be used. First, we consider the ideal case where there are no hidden nodes and no measurement noise. Second, we extend the ideal case to the more general case.

#### Reconstructing gene regulatory networks (ideal case)

In Equation (13),  $q$  represents the  $s$ -sparse network structure which we want to reconstruct from the



**Figure 6 Design effective experiment for a nonlinear system.** (Example 5) Reconstruction result based on different experiment (nonlinear case): Exp#1 represents the original experimental dataset which has limitation to reduce coherence; Exp#2 represents non-effective experimental dataset (inhibit  $x_1$ ); and Exp#3 represents the effective experimental dataset (inhibit  $x_2$ ) **(A)** the time series of  $x_1, x_2, x_3$  for each experiment (top), the reconstruction error (bottom left) and coherence comparison for each experiment (bottom right) **(B)** the corresponding topology **(C)** coherence distribution of the sensing matrix. In Exp#3, we reduce the coherence below 0.8.

time series gene expression by solving the  $l_1$ -norm optimization:

$$\min \|\mathbf{q}\|_{l_1} \text{ s.t. } \tilde{\mathbf{Z}} = \tilde{\mathbf{\Omega}}\mathbf{q} \quad (CS) \quad (17)$$

where  $\tilde{\mathbf{Z}} \triangleq \mathcal{P}\mathbf{Z}$ ,  $\tilde{\mathbf{\Omega}} \triangleq \mathcal{P}\mathbf{\bar{\Omega}}$  and  $\mathcal{P}$  is the optimal transformation in Equation (24).

**Proposition 2.** *If the sensing matrix  $\tilde{\mathbf{\Omega}}$  constructed from time series data, multiplied by the optimal transformation,  $\mathcal{P}$ , has  $2s$  linearly independent columns, then any  $s$ -sparse network structure  $\mathbf{q}$  can be reconstructed uniquely from  $\tilde{\mathbf{Z}} = \tilde{\mathbf{\Omega}}\mathbf{q}$ .*

*Proof.* (Suppose not), then there are two  $s$ -sparse graph structures  $\mathbf{q}_1, \mathbf{q}_2$  with  $\tilde{\mathbf{\Omega}}\mathbf{q}_1 = \tilde{\mathbf{\Omega}}\mathbf{q}_2$  (or  $\tilde{\mathbf{\Omega}}(\mathbf{q}_1 - \mathbf{q}_2) = 0$ ). However,  $\mathbf{q}_1 - \mathbf{q}_2$  is  $2s$ -sparse, so there is a linear dependence between  $2s$  columns of  $\tilde{\mathbf{\Omega}}$  (contradiction).  $\square$

The requirement of  $2s$  linearly independent columns in **Proposition 2** may be translated to an incoherence condition on the sensing matrix. That is, if the unknown  $s$ -sparse signal  $\mathbf{q}$  is reasonably sparse, it is possible to recover  $\mathbf{q}$  under the incoherence condition on the sensing matrix. Although the sensing matrix consists of redundant dictionaries, the coherence of the sensing matrix can be reduced. In a heuristic way, we multiply the redundant dictionaries  $\Phi_{k,:}$  by a randomly chosen matrix  $\Psi_k^s$  at each time step  $k$  and iterate this step until the coherence is decreased. Or, we can find the optimal transformation  $\mathcal{P}$  (or  $\mathbf{P}$ ) in Equation (24) to reduce the coherence. In the previous numerical examples, we illustrated that the coherence of the sensing matrix is decreased by transformation and showed the exact reconstruction of graph structure.

**Example 6.** (Statistics) *Here, we compare the success rate of the proposed method with other methods such as L1 and L2. Figure 7 shows statistics of 50 trials for a simple linear case (for each trial, we randomly generate the influence map). Here, we count the number of successes of each method when any of the methods recover the exact structure. By reducing coherence, we can improve the success rate as shown in Figure 7(A). Also, L2-regularization does not encourage sparsity but distributes the coefficients to be more similar to each other as shown in Figure 7(B).*

**Graph reconstruction with hidden nodes**

The main contributions of the proposed method in the previous section is the conversion of the problem of inferring graph structure into the CS framework. Then, we demonstrate that one could recover sparse graph structures from only a few measurements. However, for practical use, the proposed method needs to be able to deal

with both sparsely corrupted signals ( $\mathbf{v} = \mathbf{g}(\mathbf{x}_i)$ ) and measurement noise ( $\mathbf{w}$ ) in Equation (5).

In general, the assumption of accessibility or observability of all nodes [17] is not satisfied. Thus, we focus on the case in which the hidden node affects observable nodes directly as shown in Figure 1(B). Also, without loss of generality, the hidden node dynamics could be any arbitrary dynamic model. Or, even if there is no hidden node, a small portion of the biological experiment dataset could be in practice contaminated by large error resulting from, for example, mislabeling, or improper use of markers or antibodies. Moreover, all biological datasets are contaminated by at least a small amount of noise from measurement devices. Therefore, the proposed method should be robust. We note this goes beyond the results in [11] due to the consideration of hidden node dynamics with measurement noise.

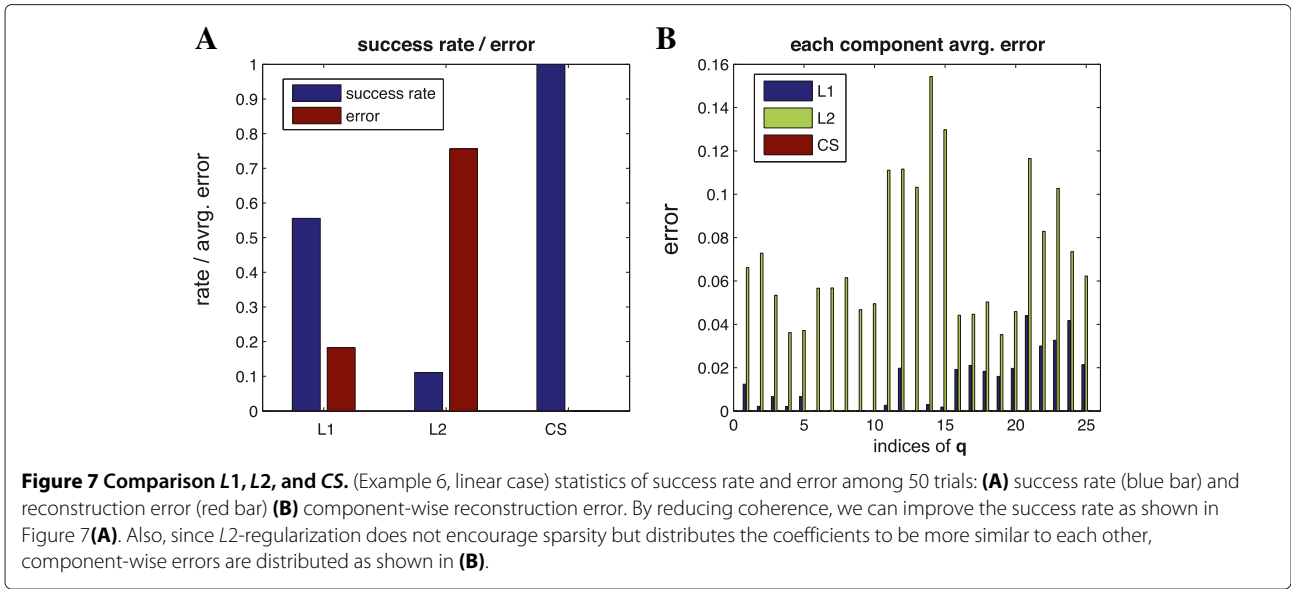
Here, the question is whether it is still possible to reconstruct the graph structure reliably when measurements are corrupted. Since hidden nodes and measurement noise are considered, the number of time points is assumed to be greater than that of the previous case [17] (i.e., no corruption and no measurement noise). Thus, the number of rows of the sensing matrix is assumed to be greater than the number of columns. If the number of the time points  $M (< N)$  is limited, then we can stack  $\bar{\mathbf{z}}(j)$  with different  $\Psi_k^s$  or including different dataset as described in Equation (16).

In CS literature [31], two decoding strategies for recovering the signal from a corrupted measurement are introduced, where the corruption includes both a possible sparse vector of large errors and a vector of small error affecting all the entries. It is shown that two decoding schemes allow the recovery of the signal with nearly the same accuracy as if no sparse large errors occurred. Our contribution is converting the problem of inferring the graph structure with hidden nodes into the highly robust error correction method framework [31] (see Method: Section ‘Two-step refinements’) and showing how this can improve the reliability of reconstruction.

**Example 7.** (arbitrary corruption with no measurement noise) *Consider nonlinear ODEs as follows:*

$$\begin{aligned} \dot{x}_1 &= -\gamma_1 x_1 + \alpha_{13} h_{act}(x_3) + v_1 \\ \dot{x}_2 &= -\gamma_2 x_2 + \alpha_{21} h_{act}(x_1) + v_2 \\ \dot{x}_3 &= -\gamma_3 x_3 + \alpha_{34} h_{act}(x_4) + \beta_{32} h_{ihb}(x_2) + v_3 \\ \dot{x}_4 &= -\gamma_4 x_4 + \beta_{41} h_{ihb}(x_1) + v_4 \end{aligned} \quad (18)$$

where  $h_{act}, h_{ihb}$  represents Hill functions for activation and inhibition respectively, and  $v_i$  represents arbitrary corruption shown in Figure 8(A) and assumed to be sparse



(at each time step, we choose  $\text{card}(v(j)) = 1$ ). The magnitude of  $v_i$  are about 50% of the magnitude of  $\dot{x}$ . Since we consider arbitrary corruption, we need more time points ( $M > N$ ). By using two-step refinements, first we estimate sparse large corruption  $\mathbf{v}$  as shown in Figure 8(B) (top) and then, we reconstruct  $\mathbf{q}$  (bottom).

In practice, a specific node is corrupted by a hidden node and a small portion of the dataset can be largely corrupted by human error. Also, since we choose the set of possible candidate basis functions of the sensing matrix in Equation (6), the columns of the sensing matrix may not be able to represent the underlying system (i.e., missing dictionaries). Then, we can consider the influence from these missing dictionaries as  $\mathbf{v}$ .

**Example 8.** (Arbitrary corruption with measurement noise) Recall a model in Equation (18) with different parameters and consider sparse large corruption  $\mathbf{v}$  and small magnitude noise  $\mathbf{w}$  (1% of the magnitude of  $\dot{x}$ ). Figure 9 shows the time series data and reconstruction result.

**Geometric view**

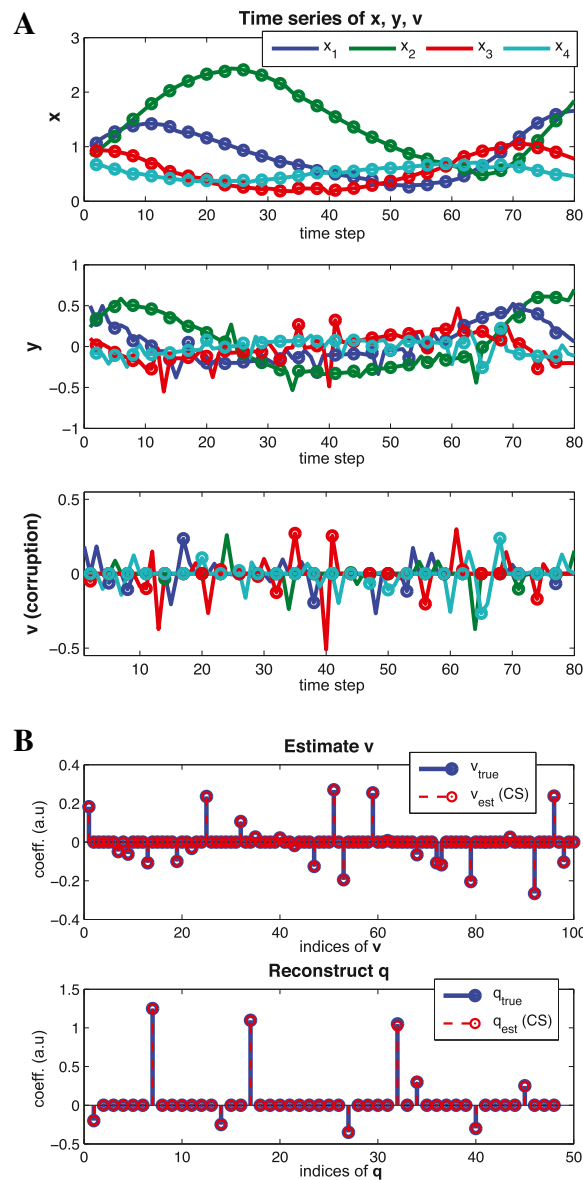
In Equation (17), since we assume all nodes are accessible and perfect measurement (meaning that there is no hidden node,  $\mathbf{v} = 0$  and no measurement noise,  $\mathbf{w} = 0$ ), we can solve Equation (27) directly without filtering out the unmodelled dynamics in Equation (26) (i.e.,  $\mathbf{Z} = \bar{\Omega}\mathbf{q}$  in Equation (27)). If there exist hidden nodes or measurement noise, we can still provide an unambiguous indication of the existence of these corruption ( $\mathbf{e} \neq 0$ ).

The intuition is that  $\mathbf{Z}(= \bar{\Omega}\mathbf{q} + \bar{\Psi}\mathbf{e})$  can be decomposed as the superposition of an arbitrary element in  $V(= \mathbf{Z} - \mathbf{Q}\mathbf{Q}^*\mathbf{e})$  and of an element in  $V^\perp(= \mathbf{Q}\mathbf{Q}^*\mathbf{e})$  as shown in Figure 10. In other words,  $\mathbf{Z}$  can be decomposed as the superposition of modelled dynamics and anomalies caused by hidden node or unmodelled dynamics. This geometric view enables us to understand how we could reveal deficiencies in our model:

- $\bar{\mathbf{Z}} = \mathbf{Q}^*\mathbf{Z} = \mathbf{0}$ : there is no hidden node  $\implies \mathbf{Z} \in \mathcal{R}(\bar{\Omega})$
- $\bar{\mathbf{Z}} \neq \mathbf{0}$ :  $\mathbf{Z}$  cannot be represented by  $\bar{\Omega}\mathbf{q}$  so there might be hidden nodes or our dictionaries in the sensing matrix  $\bar{\Omega}$  are not sufficient to represent  $\mathbf{Z}$  (revealing deficiencies in our model or dictionaries).

**HER2 Overexpressed breast cancer**

We apply the proposed algorithm to study a breast cancer signaling pathway by reconstructing the graph structure using an RPPA dataset [32] as shown in Figure 11 (see also Additional file 1: Figure S1, S2, and S3 for details: each figure presents the RPPA dataset and the result of graph reconstruction compared with L1, L2-optimization). Here, we choose small networks which are composed of 3 nodes and known to be sparsely connected, i.e.,  $\text{PI3K} \rightarrow \text{PDK} \rightarrow \text{Akt}$  and  $\text{PDK} \rightarrow \text{Akt} \rightarrow \text{mTOR}$  in order to satisfy our assumption such that the influence on observable nodes from a hidden node should be sparse (i.e.,  $\mathbf{v}$  is sparse). The graph structures identified by the proposed method are consistent with the current understanding of the networks, whereas those found using L1- and L2-optimizations fail to reconstruct the known structure as shown in Figure 12.



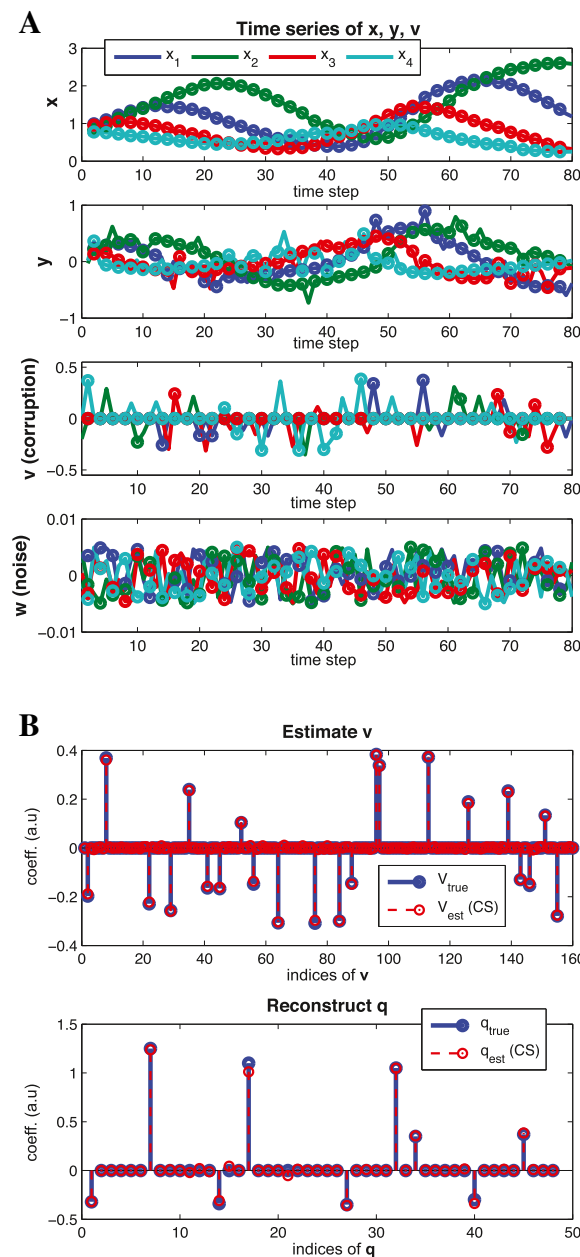
**Figure 8 Arbitrary corruption with no measurement noise.** (Example 7) Reconstruction with corrupted signal. **(A)** time series of  $x, y, v$  **(B)** reconstruction results of  $v$  and  $q$  where each circle represents sampled time points ( $n = 4, M = 25, N = 12, s = 9$ ). By using two-step refinements, first we estimate sparse large corruption  $v$  and then, we reconstruct  $q$ . Note that since we consider arbitrary corruption, we have more time points ( $M > N$ ).

Also, an abstract model of the breast cancer signal pathway proposed by M. Moasser [38] is considered, as shown in Additional file 1: Figure S3(B) where PHLPP isoforms are a pair of protein phosphatases, PHLPP1 and PHLPP2, which are important regulators of Akt serine-threonine kinases (Akt1, Akt2, Akt3) and conventional protein kinase C (PKC) isoforms. PHLPP may act as a tumor suppressor in several types of cancer due to its ability to block growth factor-induced signaling in cancer cells [39]. PHLPP dephosphorylates Ser473 (the hydrophobic motif) in Akt, thus partially inactivating the kinase [40]. Unfortunately, in our RPPA dataset, we do not have PHLPP

so we simply consider three nodes (Akt<sub>pT308</sub>, Akt<sub>pS473</sub> and mTOR). Figure 11(right) shows the result of the proposed method using the RPPA dataset. The reconstructed graph structure matches up to the known structure (Additional file 1: Figure S3(B)). Specifically, our result can capture the partial inactivating characteristics of PHLPP (i.e., mTOR( $\rightarrow$  PHLPP)  $\dashv$  Akt<sub>pS473</sub>).

### Discussion

Many network inference algorithms use  $l_1$ -norm optimization to find the actual network structure, but they cannot guarantee exact recovery. In this paper, we propose



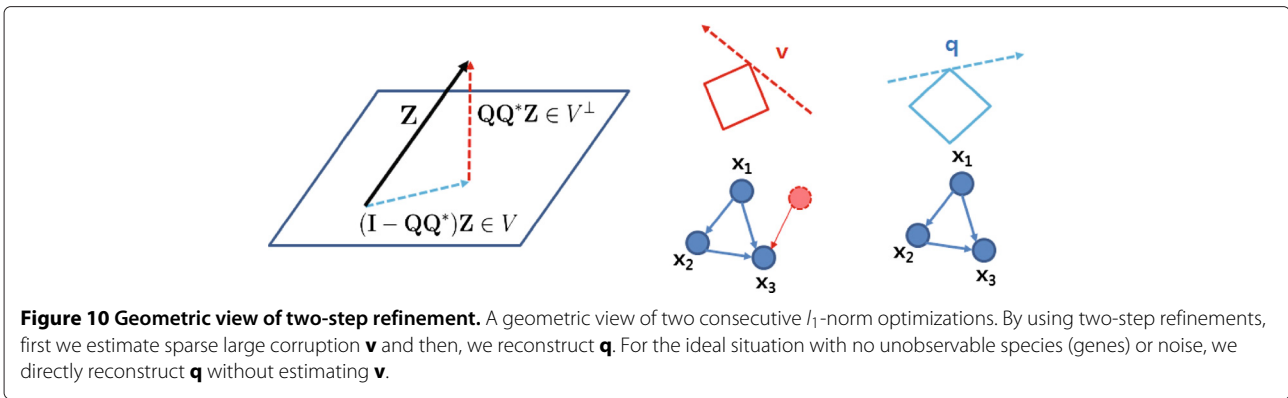
**Figure 9 Arbitrary corruption with measurement noise.** (Example 8) Reconstruction with corrupted signal. **(A)** time series of  $x, y, v$  **(B)** reconstruction results of  $v$  and  $q$  where each circle represents sample time points ( $n = 4, M = 40, N = 12, s = 9$ ). Here we consider arbitrary large corruption with measurement noise. By choosing the parameters  $\epsilon_1, \epsilon_2$  properly in Equation (30), we can recover the graph structure  $q$  within the small error bound.

a novel approach to reconstruct GRNs based on compressive sensing with a dynamic system model. We demonstrate that the incoherence condition is essential for exact recovery. This is not properly considered in inference methods based on  $l_1$ -norm optimization. Also, we illustrate how the incoherence condition can be used to design new experiments effectively. Finally, we consider a more general setting, in which hidden genes exist and all measurements are contaminated by noise, and

we show that the proposed method leads to reliable reconstruction.

We compare all the results with  $l_1$ - and  $l_2$ -norm optimization, since  $l_1$ -norm optimization is widely used in network inference and many inference methods combine the  $l_1$ - and  $l_2$ -norms (for example, elastic net). We highlight the improvement by the proposed method compared with these commonly used methods. Our use of the method to propose new experiments, and its extension





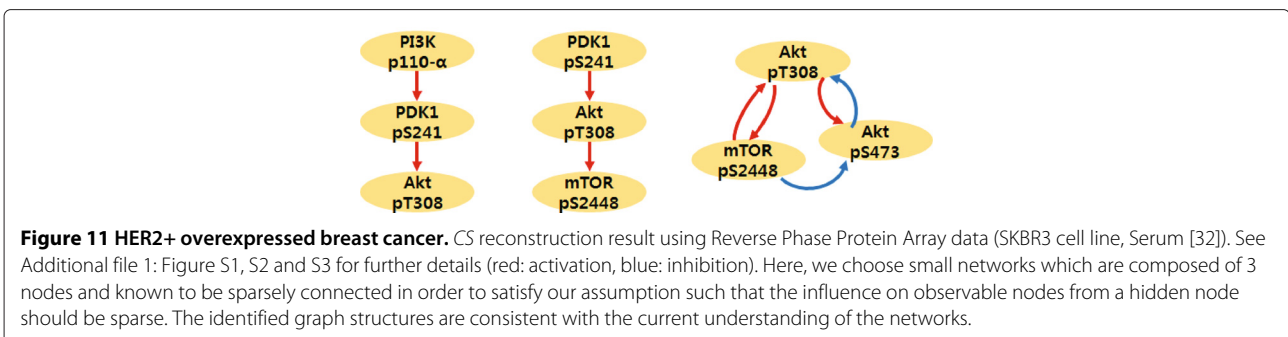
to a more general setting goes beyond the results in any network inference methods.

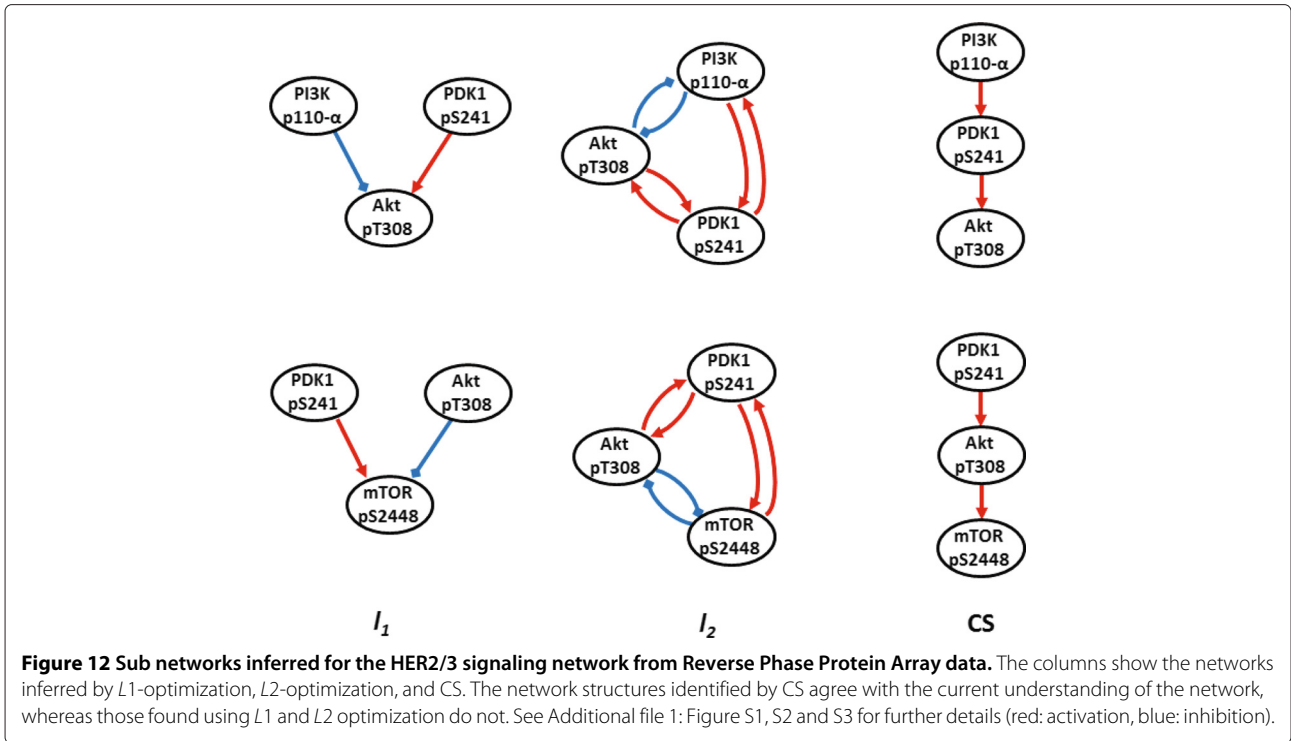
From the lessons learned from the numerical study using both synthetic datasets and real RPPA datasets of HER2 overexpressed breast cancer signaling pathways, the proposed method can be applied to sub-networks in order to satisfy our assumption about the number of hidden nodes relative to the network size, and has great potential for reconstructing GRN by designing effective experiments. Note that the proposed method is not limited to the network size itself but rather by this assumption and the incoherence condition. In the ideal case with no unobservable species (genes) and no noise, the method can be applied to entire networks if the incoherence condition is satisfied (i.e.,  $n \cdot M \geq 2s$ , practically  $n \cdot M \geq 4s$ ) where  $n$  is the number of nodes,  $M$  is the number of time points and  $s$  represents the sparsity of the network structure. For example, suppose we only have one time point measurement ( $M = 1$ ) and then, the proposed method cannot be applicable for any network ( $n > 2$  where  $n$  is the number of nodes) since it is impossible to satisfy the incoherence condition. As an example, if we consider  $n$  nodes which are simply connected (i.e.,  $s = n - 1$ ), then  $n \cdot M < 2s$ . With a reasonable number of time points ( $M$ ), the incoherence condition could be satisfied. Similarly, in the practical case with unobservable species (genes) and measurement noise, in order to guarantee the exact recovery of the graph structure, we should

satisfy Assumption 1 for a given network (hidden node only affects relatively few nodes in the given network) and the incoherence condition. Suppose we choose a small sub-network where a hidden node or unobservable node affects all nodes in the chosen sub-network: then, it is impossible to infer the exact structure of the sub-network. However, if a hidden node only affects relatively few nodes in the given network, we can still reconstruct the exact structure of the sub-network by inferring the hidden node influence first. Therefore, the proposed method is not limited to the network size itself but rather by Assumption 1 and the incoherence condition. In practice, we can work on a sub-network at a time and integrate the identified sub-networks.

Finally, we evaluate performance with respect to the size of the problem under typical parameters (the number of genes, the number of kinetic features and connectivity) in (Additional file 1: Table S1) which provides us a brief guideline of application, e.g., for an  $n$ -genes network with certain kinetic features, how many time points/how much resolution are appropriate for the reconstruction.

In this paper, we do not consider any *a priori* knowledge of connectivity. However, since we may have partial information of connectivity, we can also use prior knowledge of connectivity in the form of known nonzero elements in  $\mathbf{q}$  allowing fewer experiments to be performed. Also, here we only consider protein expression or gene expression levels but we can also consider many other





types of useful information that can be incorporated into the network reconstruction process, such as sequence motifs and direct binding measurements (e.g., ChIP-chip and ChIP-seq). Since it has been shown that ChIP-seq signals of Histone modification are more correlated with transcription factor motifs at promoter sites in comparison to RNA level, time series Histone modification ChIP-seq could provide a more reliable inference of GRNs in comparison to method based on expression level. We are interested in continuing this research direction for large-scale networks.

### Conclusion

We proposed a method for reconstructing sparse graph structures based on time series gene expression data without any *a priori* information. We demonstrated that the proposed method can reconstruct graph structure reliably. Also, we illustrated that coherence in the sensing matrix can be used as a guideline for designing effective experiments.

Second, the proposed method is extended to the cases in which dynamics is corrupted by hidden nodes and the measurement is corrupted by human error in addition to the measurement noise. Using a two-step refinement procedure, we demonstrate good performance for the reconstruction of graph structure. A set of numerical examples is implemented to illustrate the method and its performance. Also, a biological example of HER2 overexpressed breast cancer using an RPPA dataset is studied. We are

currently applying our method to recover the HER2 signaling pathway, where a significant part of the network is currently unknown.

### Method

#### Reducing coherence by transformation

##### Rearranging the sensing matrix

Define  $\mathbf{z}(j)$  as a vector of each component of  $\mathbf{y}_i$  at the  $j$ -th time point:

$$\begin{aligned} \mathbf{z}(j) &\triangleq \begin{bmatrix} \mathbf{y}_1^\top \mathbf{e}_j \\ \mathbf{y}_2^\top \mathbf{e}_j \\ \dots \\ \mathbf{y}_n^\top \mathbf{e}_j \end{bmatrix} = \begin{bmatrix} y_1(t_j) \\ y_2(t_j) \\ \dots \\ y_n(t_j) \end{bmatrix} \\ &= \begin{bmatrix} \Phi_{j,:} \mathbf{q}_1 \\ \Phi_{j,:} \mathbf{q}_2 \\ \dots \\ \Phi_{j,:} \mathbf{q}_n \end{bmatrix} + \begin{bmatrix} v_1(t_j) \\ v_2(t_j) \\ \dots \\ v_n(t_j) \end{bmatrix} + \begin{bmatrix} w_1(t_j) \\ w_2(t_j) \\ \dots \\ w_n(t_j) \end{bmatrix} \quad (19) \\ &= (\mathbf{I}_n \otimes \Phi_{j,:}) \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \mathbf{v}(j) + \mathbf{w}(j) \end{aligned}$$

where  $\Phi_{j,:}$  represents the  $j$ -th row of  $\Phi$ . Consider  $\Psi_j^s \in \mathbb{R}^{n \times n}$  and multiply Equation (19) by  $\Psi_j^s$ :

$$\bar{\mathbf{z}}(j) \triangleq \Psi_j^s \mathbf{z}(j) = \Psi_j^s (\mathbf{I}_n \otimes \Phi_{j,:}) \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \bar{\mathbf{v}}(j) + \bar{\mathbf{w}}(j) \quad (20)$$

where  $\bar{\mathbf{v}}(j) \triangleq \Psi_j^s \mathbf{v}(j)$ ,  $\bar{\mathbf{w}}(j) \triangleq \Psi_j^s \mathbf{w}(j)$  and  $\text{rank}(\Psi_i^s (\mathbf{I}_n \otimes \Phi_{i,:})) = \text{rank}(\Psi_i^s)$  since  $\text{rank}(\mathbf{I}_n \otimes \Phi_{i,:}) = \text{rank}(\mathbf{I}_n) \cdot \text{rank}(\Phi_{i,:}) = n$ . By stacking  $\bar{\mathbf{z}}(j)$ ,

$$\mathbf{Z} \triangleq \begin{bmatrix} \bar{\mathbf{z}}(1) \\ \bar{\mathbf{z}}(2) \\ \dots \\ \bar{\mathbf{z}}(M) \end{bmatrix} = \begin{bmatrix} \Psi_1^s (\mathbf{I}_n \otimes \Phi_{1,:}) \\ \Psi_2^s (\mathbf{I}_n \otimes \Phi_{2,:}) \\ \dots \\ \Psi_M^s (\mathbf{I}_n \otimes \Phi_{M,:}) \end{bmatrix} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{v}}(1) \\ \bar{\mathbf{v}}(2) \\ \dots \\ \bar{\mathbf{v}}(M) \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{w}}(1) \\ \bar{\mathbf{w}}(2) \\ \dots \\ \bar{\mathbf{w}}(M) \end{bmatrix} \triangleq \bar{\mathbf{\Omega}} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \bar{\mathbf{v}} + \bar{\mathbf{w}} \quad (21)$$

**Randomly chosen matrix  $\Psi$**

The optimization problem in Equation (14) is not trivial because of the constraint. One simple and heuristic approach is that we select  $\Psi_i^s$  by (normalized) randomly chosen matrix with independent identically distributed (i.i.d.) random variable where  $\bar{\mathbf{\Omega}}$  has full row rank, calculate  $\mu(\bar{\mathbf{\Omega}})$  and run this with several times to reduce coherence. Since the randomly chosen matrix,  $\Psi_i^s$ , spreads out the component of  $\Phi_{i,:}$  uniformly, we can reduce coherence.

**Finding the optimal transformation**

This heuristic approach may not be enough to reduce the coherence. Consider a nonsingular matrix  $\mathcal{P} \in \mathbb{R}^{M \cdot n \times M \cdot n}$  and  $\tilde{\mathbf{\Omega}} = \mathcal{P} \bar{\mathbf{\Omega}}$  where  $\bar{\mathbf{\Omega}}$  is constructed by heuristic way, i.e., randomly chosen matrix.

$$\begin{aligned} \tilde{\mu}_{i,j} &\triangleq \frac{|\langle \tilde{\mathbf{\Omega}}_i, \tilde{\mathbf{\Omega}}_j \rangle|}{\|\tilde{\mathbf{\Omega}}_i\|_2 \|\tilde{\mathbf{\Omega}}_j\|_2} = \frac{|\langle \mathcal{P} \bar{\mathbf{\Omega}}_i, \mathcal{P} \bar{\mathbf{\Omega}}_j \rangle|}{\|\mathcal{P} \bar{\mathbf{\Omega}}_i\|_2 \|\mathcal{P} \bar{\mathbf{\Omega}}_j\|_2} \\ &\triangleq \frac{|\bar{\mathbf{\Omega}}_i^\top \mathbf{P} \bar{\mathbf{\Omega}}_j|}{\sqrt{(\bar{\mathbf{\Omega}}_i^\top \mathbf{P} \bar{\mathbf{\Omega}}_i) (\bar{\mathbf{\Omega}}_j^\top \mathbf{P} \bar{\mathbf{\Omega}}_j)}} \end{aligned} \quad (22)$$

where  $\mathbf{P} \triangleq \mathcal{P}^\top \mathcal{P} = \mathbf{P}^\top \in \mathbb{S}^{M \cdot n}$  is positive definite and  $\tilde{\mathbf{\Omega}}_i$  denote the  $i$ -th column of  $\tilde{\mathbf{\Omega}}$ . Note that if  $\mathcal{P}^\top \mathcal{P} = \mathbf{I}$ ,  $\tilde{\mu}_{i,j} = \frac{|\langle \bar{\mathbf{\Omega}}_i, \bar{\mathbf{\Omega}}_j \rangle|}{\|\bar{\mathbf{\Omega}}_i\|_2 \|\bar{\mathbf{\Omega}}_j\|_2} = \frac{|\langle \bar{\mathbf{\Omega}}_i, \bar{\mathbf{\Omega}}_j \rangle|}{\|\bar{\mathbf{\Omega}}_i\|_2 \|\bar{\mathbf{\Omega}}_j\|_2} \triangleq \bar{\mu}_{i,j}$ . Therefore, our goal is finding  $\mathbf{P} (= \mathcal{P}^\top \mathcal{P}) \in \mathbb{S}^+$  such that

$$\min \mu(\tilde{\mathbf{\Omega}}) = \min_{\mathbf{P}} \max_{j < k} \tilde{\mu}_{j,k} = \min_{\mathbf{P}} \|\Gamma\|_\infty \quad (23)$$

where  $\Gamma \in \mathbb{R}^{(n \cdot N) \times C_2}$  can be defined as follows:

$$\Gamma \triangleq \begin{bmatrix} \frac{\text{tr}(\mathcal{P} \bar{\mathbf{\Omega}}_1 \bar{\mathbf{\Omega}}_2^\top \mathcal{P}^\top)}{\|\mathcal{P} \bar{\mathbf{\Omega}}_1 \bar{\mathbf{\Omega}}_2^\top \mathcal{P}^\top\|_F} & \dots & \frac{\text{tr}(\mathcal{P} \bar{\mathbf{\Omega}}_1 \bar{\mathbf{\Omega}}_{N \cdot n}^\top \mathcal{P}^\top)}{\|\mathcal{P} \bar{\mathbf{\Omega}}_1 \bar{\mathbf{\Omega}}_{N \cdot n}^\top \mathcal{P}^\top\|_F} \\ \dots & \dots & \dots \\ \frac{\text{tr}(\mathcal{P} \bar{\mathbf{\Omega}}_2 \bar{\mathbf{\Omega}}_3^\top \mathcal{P}^\top)}{\|\mathcal{P} \bar{\mathbf{\Omega}}_2 \bar{\mathbf{\Omega}}_3^\top \mathcal{P}^\top\|_F} & \dots & \frac{\text{tr}(\mathcal{P} \bar{\mathbf{\Omega}}_{N \cdot n - 1} \bar{\mathbf{\Omega}}_{N \cdot n}^\top \mathcal{P}^\top)}{\|\mathcal{P} \bar{\mathbf{\Omega}}_{N \cdot n - 1} \bar{\mathbf{\Omega}}_{N \cdot n}^\top \mathcal{P}^\top\|_F} \end{bmatrix}^\top$$

In practice, we ignore the denominator and solve the following problem:

$$\min_{\mathbf{P} = \mathbf{P}^\top > 0} \left\| \left[ \text{vec}(\bar{\mathbf{\Omega}}_1 \bar{\mathbf{\Omega}}_2^\top) \dots \text{vec}(\bar{\mathbf{\Omega}}_1 \bar{\mathbf{\Omega}}_{N \cdot n}^\top) \dots \text{vec}(\bar{\mathbf{\Omega}}_{N \cdot n - 1} \bar{\mathbf{\Omega}}_{N \cdot n}^\top) \right]^\top \text{vec}(\mathbf{P}) \right\|_\infty \quad (24)$$

We can also combine  $\|\cdot\|_\infty$  and  $\|\cdot\|_1$  to reduce the coherence. Note that for  $\|\cdot\|_\infty$ , we minimize the maxim coherence and for  $\|\cdot\|_1$ , we minimize the sum of the all possible combinations of the columns of  $\bar{\mathbf{\Omega}}$ , i.e.,  $\tilde{\mu}_{ij}$ . Thus, if certain bases are highly correlated,  $\mathbf{P}$  or  $\mathcal{P}$  makes the components of the sensing matrix spread out enough to differentiate the influences from those bases. Since we ignore the denominator, the optimal solution of Equation (24) may be suboptimal. Thus, we can also combine heuristic way and Equation (24) iteratively to reduce coherence of the sensing matrix in practice.

**Two-step refinements**

**Sparse large corruption with no measurement noise**

Recall Equation (13) where  $\bar{\mathbf{w}} = \mathbf{0}$ :

$$\begin{aligned} \mathbf{Z} &= \bar{\mathbf{\Omega}} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \text{diag}\{\Psi_1^s, \Psi_2^s, \dots, \Psi_M^s\} \begin{bmatrix} \mathbf{v}(1) \\ \mathbf{v}(2) \\ \dots \\ \mathbf{v}(M) \end{bmatrix} \\ &= \bar{\mathbf{\Omega}} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \bar{\Psi} \mathbf{v} \end{aligned} \quad (25)$$

where  $\bar{\Psi} \triangleq \text{diag}\{\Psi_1^s, \Psi_2^s, \dots, \Psi_M^s\}$  and  $\mathbf{v}(j)$  represents sparse large corruption at the  $j$ -th time point, that could result from the existence of hidden nodes. We assume that the influence from hidden nodes is sparse and unknown (i.e.,  $\mathbf{v}(j)$  is assumed to be sparse). In other words, hidden nodes can affect only a few nodes' dynamics (intuitively, if hidden nodes affect all nodes, there is no way to reconstruct the graph structure). Then, we consider the reconstruction of graph structure  $\mathbf{q}$  from the corrupted signal  $\mathbf{Z}$ .

Since we assume the number of rows of the sensing matrix  $\bar{\Omega}$  is greater than the number of columns ( $M \cdot n > N \cdot n$ ), we consider  $\mathbf{Q}^*$  which annihilates the sensing matrix  $\bar{\Omega}$  on the left ( $\mathbf{Q}^* \bar{\Omega} = 0$ ) where  $\mathbf{Q}^*$  is any  $(M \cdot n - N \cdot n) \times M \cdot n$  matrix whose kernel is the range of  $\bar{\Omega}$  in  $\mathbb{R}^{M \cdot n}$  ( $\text{rank}(\mathbf{Q}^*) + \text{nullity}(\mathbf{Q}^*) = M \cdot n$ ):

$$\bar{\mathbf{Z}} \triangleq \mathbf{Q}^* \mathbf{Z} = \mathbf{Q}^* \bar{\Omega} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \mathbf{Q}^* \bar{\Psi} \mathbf{v}$$

Then, the following two-step optimization problem enables us to compute  $\mathbf{q}$ :

- Filter unmodelled dynamics out from the measurement:

$$\hat{\mathbf{v}} = \arg \min \|\mathbf{v}\|_{l_1} \text{ s.t. } \bar{\mathbf{Z}} = \mathbf{Q}^* \bar{\Psi} \mathbf{v} \quad (26)$$

- Reconstruct  $\mathbf{q}$  :

$$\begin{aligned} \min \|\mathbf{q}\|_{l_1} \text{ s.t. } \quad & \mathbf{Z} - \bar{\Psi} \hat{\mathbf{v}} = \bar{\Omega} \mathbf{q} \\ \text{or } \bar{\mathbf{Z}} = \mathcal{P}(\mathbf{Z} - \bar{\Psi} \hat{\mathbf{v}}) = \mathcal{P} \bar{\Omega} \mathbf{q} = \tilde{\Omega} \mathbf{q} \quad & (\mathcal{P} \text{ in Equation (23)}) \end{aligned} \quad (27)$$

If we could somehow get an accurate estimate  $\hat{\mathbf{v}}$  from Equation (26), Equation (27) represents the problem of reconstructing the graph structure  $\mathbf{q}$ . The intuition is that  $\mathbf{Z} (= \bar{\Omega} \mathbf{q} + \bar{\Psi} \mathbf{v})$  can be decomposed as the superposition of modelled dynamics and anomalies caused by hidden node or unmodelled dynamics.

The two step convex optimization problems in Equation (26) and Equation (27) are  $l_1$ -norm optimization problems in CS. Thus, if the sensing matrix  $\mathbf{Q}^* \bar{\Psi}$  and  $\bar{\Omega}$  (or  $\mathcal{P} \bar{\Omega}$ ) satisfy the incoherence condition, signals  $\mathbf{v}$  and  $\mathbf{q}$  can be recovered exactly [17,34]. Here, there are many possible choices of  $\mathbf{Q}^*$  but we have to choose  $\mathbf{Q}^*$  to satisfy the incoherence condition for the exact recovery of  $\mathbf{v}$  [17]. To choose such a  $\mathbf{Q}^*$ , we observe that  $\bar{\Omega}$  can be denoted as follows using Singular Value Decomposition (SVD):

$$\bar{\Omega} = [\mathbf{U}_1 \ \mathbf{U}_2] \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix} = \mathbf{U}_1 \Sigma \mathbf{V}_1^T \quad (28)$$

where  $\mathbf{U}_1 \in \mathbb{R}^{M \cdot n \times r}$ ,  $\mathbf{U}_2 \in \mathbb{R}^{M \cdot n \times (M \cdot n - r)}$ ,  $\Sigma \in \mathbb{R}^{r \times r}$ ,  $\mathbf{V}_1 \in \mathbb{R}^{N \cdot n \times r}$ ,  $\mathbf{V}_2 \in \mathbb{R}^{N \cdot n \times (N \cdot n - r)}$  and  $r$  is the rank of  $\bar{\Omega}$ . Suppose we choose  $\mathbf{Q}^*$  such that  $\mathbf{Q}^* = \Xi \mathbf{U}_2^T$ . Then:

$$\begin{aligned} \mathbf{Q}^* \mathbf{Z} &= \mathbf{Q}^* \bar{\Omega} \mathbf{q} + \mathbf{Q}^* \bar{\Psi} \mathbf{v} \\ \bar{\mathbf{Z}} &= (\Xi \mathbf{U}_2^T) \mathbf{U}_1 \Sigma \mathbf{V}_1^T \mathbf{q} + (\Xi \mathbf{U}_2^T) \bar{\Psi} \mathbf{v} \\ \bar{\mathbf{Z}} &= (\Xi \mathbf{U}_2^T) \bar{\Psi} \mathbf{v} \end{aligned}$$

where  $\Xi$  can be used as a tuning matrix for satisfying the incoherence condition. A geometric view in Results and discussion: Section ‘Geometric view’. enables us to understand how we could reveal deficiencies in our model.

### Sparse large corruption with measurement noise

While considering influence from hidden nodes is interesting, it still may not be realistic to assume that except for hidden nodes, one is able to measure the node dynamics with infinite precision. A better model would assume that there is measurement noise. Consider the problem of recovering the graph structure  $\mathbf{q}$  from the vector  $\mathbf{Z}$  which is corrupted by measurement noise  $\bar{\mathbf{w}}$  in Equation (13):

$$\begin{aligned} \mathbf{Z} &= \bar{\Omega} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \bar{\Psi} \begin{bmatrix} \mathbf{v}(1) + \mathbf{w}(1) \\ \mathbf{v}(2) + \mathbf{w}(2) \\ \dots \\ \mathbf{v}(M) + \mathbf{w}(M) \end{bmatrix} \\ &= \bar{\Omega} \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \dots \\ \mathbf{q}_n \end{bmatrix} + \bar{\Psi} \begin{bmatrix} \mathbf{e}(1) \\ \mathbf{e}(2) \\ \dots \\ \mathbf{e}(M) \end{bmatrix} \end{aligned} \quad (29)$$

where  $\mathbf{e}(j) = \mathbf{v}(j) + \mathbf{w}(j)$ ,  $\mathbf{w}$  is Gaussian noise  $\mathcal{N}(0, \sigma)$  assumed to be bounded  $\|\mathbf{w}\|_{l_2} \leq \epsilon$ . In general, we can consider any corruption decomposed into sparse large error  $\mathbf{v}$  and small magnitude error  $\mathbf{w}$  [31]. Then, modified two-step refinements can be applied as follows:

$$\begin{aligned} \bar{\mathbf{Z}} &= \mathbf{Q}^* \bar{\Omega} \mathbf{q} + \mathbf{Q}^* \bar{\Psi} \mathbf{e} = \mathbf{Q}^* \bar{\Psi} (\mathbf{v} + \mathbf{w}) \\ \hat{\mathbf{v}} &= \arg \min \|\mathbf{v}\|_{l_1} \text{ s.t. } \|\bar{\mathbf{Z}} - \mathbf{Q}^* \bar{\Psi} \mathbf{v}\|_{l_2} \leq \epsilon_1 \\ \mathbf{q} &= \arg \min \|\mathbf{q}\|_{l_1} \text{ s.t. } \|\mathbf{Z} - \bar{\Omega} \mathbf{q} - \hat{\mathbf{e}}\|_{l_2} \leq \epsilon_2, \hat{\mathbf{e}} = \bar{\Psi} \hat{\mathbf{v}} \end{aligned} \quad (30)$$

where the parameters  $\epsilon_1, \epsilon_2$  above depend on the magnitude of the small errors  $\epsilon$ , which can be determined as in [31].

### Endnotes

<sup>a</sup>If  $\mathbf{A}$  is an  $m \times n$  matrix and  $\mathbf{B}$  is a  $p \times q$  matrix, then the Kronecker product  $\mathbf{A} \otimes \mathbf{B}$  is the  $mp \times nq$  block matrix:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11} \mathbf{B} & \dots & a_{1n} \mathbf{B} \\ \vdots & \ddots & \vdots \\ A_{m1} \mathbf{B} & \dots & a_{mn} \mathbf{B} \end{bmatrix}$$

<sup>b</sup>We compare the performance of CS with the performance of the  $L_1$  and  $L_2$  optimization as follows:

$$\begin{aligned} \min_{\mathbf{q}_i} \|\mathbf{q}_i\|_{l_1} \text{ subject to } \mathbf{y}_i &= \Phi \mathbf{q}_i \quad (i = 1, \dots, n) \quad (L1) \\ \min_{\mathbf{q}} \|\mathbf{q}\|_{l_2} \text{ subject to } \mathbf{Y} &= \Omega \mathbf{q} \quad (L2) \end{aligned}$$

## Additional file

Additional file 1: Supporting Information.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

Conceived and developed computational method: YHC CJT, Analyzed the data : YHC CJT, Supervised the project: JWG CJT, Wrote the paper: YHC CJT. All authors read and approved the final manuscript.

### Acknowledgements

This research was supported by the NIH NCI under the ICBP and PS-OC programs (5U54CA112970-08), and by the Stand Up To Cancer-American Association for Cancer Research Dream Team Translational Cancer Research Grant SU2C-AACR-DT0409 to JWG.

### Author details

<sup>1</sup>Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 94720 Berkeley, CA, USA. <sup>2</sup>Department of Biomedical Engineering and the Center for Spatial Systems Biomedicine, Oregon Health and Science University, Portland, OR, USA. <sup>3</sup>Faculty Scientist, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

Received: 18 April 2014 Accepted: 27 November 2014

Published online: 14 December 2014

### References

- Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signaling circuits in molecular interaction networks.** *Bioinformatics* 2002, **18**:233–240.
- Wiley HS, Shvartsman SY, Lauffenburger DA: **Computational modeling of the egf-receptor system: a paradigm for systems biology.** *Trends Cell Biol* 2003, **13**(1):43–50.
- Ma L, Iglesias PA: **Quantifying robustness of biochemical network models.** *BMC Bioinformatics* 2002, **3**(38):1–13.
- Sontag ED: **Network reconstruction based on steady-state data.** *Essays Biochem* 2008, **45**:161–176.
- Zechner C, Ruessa J, Krenn P, Pelet S, Peter M, Lygeros J, Koepl H: **Moment-based inference predicts bimodality in transient gene expression.** *Proc Natl Acad Sci U S A* 2012, **109**(21):8340–8345.
- Zavlanos MM, Julius AA, Boyd SP, Pappas GJ: **Identification of stable genetic networks using convex programming.** In *Proceedings of the American Control Conference (ACC)*. Seattle, WA: IEEE; 2008:2755–2760.
- Cooper NG, Belta CA, Julius AA: **Genetic regulatory network identification using multivariate monotone functions.** In *Proceedings of the IEEE conference on Decision and Control and European Control Conference (CDC-ECC)*. Orlando, FL: IEEE; 2011:2208–2213.
- Porreca R, Drulhe S, de Jong H, Ferrari-Trecate G: **Structural identification of piecewise-linear models of genetic regulatory networks.** *J Comput Biol* 2008, **15**(10):1365–1380.
- Bernardo DD, Gardner TS, Collins JJ: **Robust identification of large genetic networks.** *Pac Symp Biocomput* 2004, **9**:486–497.
- Richard G, Julius A, Belta C: **Optimizing regulation functions in gene network identification.** In *IEEE Conference on Decision and Control (CDC)*. Firenze, Italy: IEEE; 2013:745–750.
- Gonçalves JM, Warnick SC: **Necessary and sufficient conditions for dynamical structure reconstruction of lti networks.** *IEEE Trans Automatic Control* 2008, **53**(7):1670–1674.
- Sontag E, Kiyatkin A, Kholodenko BN: **Inferring dynamics architecture of cellular network using time series of gene expression, protein and metabolite data.** *Bioinformatics* 2004, **20**(12):1877–1886.
- Han S, Yoon Y, Cho K-H: **Inferring biomolecular interaction networks based on convex optimization.** *Comput Biol Chem* 2007, **31**(5-6):347–354.
- Napoletani D, Sauer T, Struppa DC, Petricoin E, Liotta L: **Augmented sparse reconstruction of protein signaling networks.** *J Theor Biol* 2008, **255**(1):40–52.
- Yuan Y, Stan G-B, Warnick S, Gonçalves J: **Robust dynamical network structure reconstruction.** *Automatica* 2011, **47**:1230–1235.
- Chang YH, Tomlin CJ: **Data-driven graph reconstruction using compressive sensing.** In *IEEE Conference on Decision and Control (CDC)*. Maui, HI: IEEE; 2012:1035–1040.
- Steinke F, Seeger M, Tsuda K: **Experimental design for efficient identification of gene regulatory networks using sparse bayesian models.** *BMC Syst Biol* 2007, **1**:51. doi: 10.1186/1752-0509-1-51.
- Gardner TS, di Bernardo D, Lorenz D, Collins JJ: **Inferring genetic networks and identifying compound mode of action via expression profiling.** *Science* 2003, **301**(5629):102–105.
- Tegner J, Yeung MKS, Hasty J, Collins JJ: **Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling.** *Proc Natl Acad Sci* 2003, **100**:5944–5949.
- Jeong H, Mason SP, Barabasi A-L, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**(6833):41–42.
- Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: Simple building blocks of complex networks.** *Science* 2002, **298**(5594):824–827.
- Burda Z, Krzywicki A, Martin OC, Zagorski M: **Motifs emerge from function in model gene regulatory networks.** *Proc Natl Acad Sci* 2011, **108**(42):17263–17268.
- Pan W, Yuan Y, Gonçalves J, Stan G-B: **Reconstruction of arbitrary biochemical reaction networks: A compressive sensing approach.** In *IEEE Conference on Decision and Control (CDC)*. Maui, HI: IEEE; 2012:2334–2339.
- Noor A, Serpedin E, Nounou M, Nounou H: **Reverse engineering sparse gene regulatory networks using cubature kalman filter and compressed sensing.** *Adv Bioinformatics* 2013, **2013**:11. doi:10.1155/2013/205763.
- Emad A, Milenkovic O: **CaSPIAN: A causal compressive sensing algorithm for discovering directed interactions in gene networks.** *PLoS ONE* 2014, **9**(3):e90781. doi:10.1371/journal.pone.0090781.
- Wang W-X, Lai Y-C, Grebogi C, Ye J: **Network reconstruction based on evolutionary-game data via compressive sensing.** *Phys Rev X* 2011, **1**:021021.
- Candès E, Romberg J: **Sparsity and incoherence in compressive sampling.** *Inverse Probl* 2006, **23**:969–985.
- Chang YH, Gray J, Tomlin CJ: **Optimization-based inference for temporally evolving boolean networks with applications in biology.** In *Proceedings of the American Control Conference (ACC)*. San Francisco, CA: IEEE; 2011:4129–4134.
- Chang YH, Gray J, Tomlin CJ: **Optimization-based inference for temporally evolving networks with applications in biology.** *J Comput Biol* 2012, **19**(12):1307–1323.
- Candès EJ, Randall PA: **Highly robust error correction by convex programming.** *IEEE Trans Inf Theory* 2008, **54**(7):2829–2840.
- Hennessy BT, Lu Y, Gonzalez-Angulo AM, Carey MS, Myhre S, Ju Z, Davies MA, Liu W, Coombes K, Meric-Bernstam F, Bedrosian I, McGahren M, Agarwal R, Zhang F, Overgaard J, Alsner J, Neve RM, Kuo W-L, Gray JW, Borresen-Dale A-L, Mills GB: **A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers.** *Clinical Proteomics*, **6**(4):129–151.
- Tao T: *Compressed sensing (Or: the equation Ax = b, revisited)*, Mahler Lecture Series. Los Angeles: University of California; 2009.
- Candès E, Wakin M: **An introduction to compressive sampling.** *IEEE Signal Process Mag* 2008, **25**(2):21–30.
- Candès EJ, Romberg JK, Tao T: **Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information.** *IEEE Trans Inf Theory* 2004, **52**:489–509.
- Yeung MKS, Tegner J, Collins JJ: **Reverse engineering gene networks using singular value decomposition and robust regression.** *Proc Natl Acad Sci* 2002, **99**(9):6163–6168.
- Ideker T, Thorsson V, Siegel AF, Hood LE: **Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data.** *J Comput Biol* 2000, **7**(6):805–817.

38. Moasser M: *Understanding the network topology underlying addiction to HER. Integrative Cancer Biology Program (ICBP) Retreat (Days of Science, the Sequel)*. SoutheastWaverly Drive, Portland: Waverly Club; 2012.
39. Brognard J, Newton AC: **Phlipping the switch on akt and protein kinase c signaling**. *Trends Endocrinol Metab* 2008, **19**(6):223–30.
40. Gao T, Furnari F, Newton AC: **Phlpp: a phosphatase that directly dephosphorylates akt, promotes apoptosis, and suppresses tumor growth**. *Mol Cell* 2005, **18**(1):13–24.

Submit your next manuscript to BioMed Central  
and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

