

RESEARCH

Open Access



# Reverse engineering gene regulatory networks from measurement with missing values

Oyetunji E. Ogundijo, Abdulkadir Elmas and Xiaodong Wang\*

## Abstract

**Background:** Gene expression time series data are usually in the form of high-dimensional arrays. Unfortunately, the data may sometimes contain missing values: for either the expression values of some genes at some time points or the entire expression values of a single time point or some sets of consecutive time points. This significantly affects the performance of many algorithms for gene expression analysis that take as an input, the complete matrix of gene expression measurement. For instance, previous works have shown that gene regulatory interactions can be estimated from the complete matrix of gene expression measurement. Yet, till date, few algorithms have been proposed for the inference of gene regulatory network from gene expression data with missing values.

**Results:** We describe a nonlinear dynamic stochastic model for the evolution of gene expression. The model captures the structural, dynamical, and the nonlinear natures of the underlying biomolecular systems. We present point-based Gaussian approximation (PBGA) filters for joint state and parameter estimation of the system with *one-step* or *two-step missing measurements*. The PBGA filters use Gaussian approximation and various quadrature rules, such as the unscented transform (UT), the third-degree cubature rule and the central difference rule for computing the related posteriors. The proposed algorithm is evaluated with satisfying results for synthetic networks, in silico networks released as a part of the DREAM project, and the real biological network, the in vivo reverse engineering and modeling assessment (IRMA) network of yeast *Saccharomyces cerevisiae*.

**Conclusion:** PBGA filters are proposed to elucidate the underlying gene regulatory network (GRN) from time series gene expression data that contain missing values. In our state-space model, we proposed a measurement model that incorporates the effect of the missing data points into the sequential algorithm. This approach produces a better inference of the model parameters and hence, more accurate prediction of the underlying GRN compared to when using the conventional Gaussian approximation (GA) filters ignoring the missing data points.

**Keywords:** Gene expression, Missing data, Bayesian inference, Gaussian filters, Network inference

## 1 Introduction

Gene regulation happens to be one of the most important processes that take place in living cells [1, 2]. For instance, it includes controls over the transcription of messenger RNA (mRNA) and the eventual translation of mRNA into protein via gene regulatory networks (GRNs). A detailed network may depict various inter-dependencies among genes where nodes of the network represent the genes and the edges correspond to interactions among the genes [3].

The strength of these interactions represents the extent to which a gene is affected by other genes in the network. For instance, some of the genes encode specific proteins, known as the transcription factors that can bind deoxyribonucleic acid (DNA) as part of a complex or independently and regulate their rate of transcription [4, 5]. Binding of the DNA by the transcription factors may, in some occasions, include genes that encode for other transcription factors and also genes that encode proteins for other functions. Hence, this results in a complex level of interaction among the genes in the cell. Among others, understanding the complex intracellular network in

\*Correspondence: wangx@ee.columbia.edu  
Department of Electrical Engineering, Columbia University, 500 W 120th Street, New York, NY 10027, USA

a human cell may lead to the identification of diseased genes, drug targets, and biomarkers for complex diseases [6]. Thus, identifying the structure of GRNs has become a major focus in the systems approach to biology [7–10].

The generation of high throughput time series measurement of transcript levels (e.g., via microarray experiments) has become an increasingly powerful tool for investigating complex biological processes and a useful resource for GRN inference [11]. Modeling of the gene networks with gene expression data can be loosely categorized into static and dynamic models. A static approach to modeling gene expressions makes use of the following properties: correlation, statistical independence for clustering, and mutual information [12, 13]. Particularly, the clustering approach has gained significant popularity [14, 15]. On the other hand, the dynamic modeling of GRNs from time series data has also received considerable interest. For instance, Boolean network models quantize the empirical gene expression data into binary values [16] and view the network structures as constraints. Further, via the estimation of the parameters in S-systems, a kind of nonlinear mathematical models based on power law, few authors like [17, 18] have performed the reverse engineering of GRNs. Probabilistic Boolean network models are an extension to the Boolean network models which incorporate the inherent stochasticity of gene expression and the uncertainties introduced by the measurement noise [19]. Also, dynamic Bayesian networks (DBNs) have been proposed to model the time series gene expression data [20, 21] because DBNs can model stochasticity and handle noisy/hidden variables.

The state-space approach, an extension of the DBNs, is a popular technique to model the GRNs [22, 23], where the hidden state of the network can be estimated by Gaussian approximation (GA) filters. The conventional Kalman filter, being optimal for a linear Gaussian system [24], requires some modifications to be able to cope with the nonlinearity of the activation function that regulates the gene activity profile. For instance, the extended Kalman filter (EKF) uses the first-order terms of the Taylor's series expansion [25] to linearize the nonlinear functions in the model. The EKF only calculates the posterior densities accurately to the first order with all higher moments truncated. A different paradigm of the GA filtering approach is the point-based filtering technique, which involves numerically integrating nonlinear functions by using a set of deterministic points. This approach lowers the computational complexity when compared to the Monte Carlo numerical integration which relies on randomly generated points, since it requires much less number of points with the same accuracy.

However, in reality, gene expression time series data may not contain sufficient quantity of data in the appropriate format for the inference of GRNs because of the missing

data points [26]. For example, in microarray measurement of gene expression, errors such as insufficient resolution and image corruption or simply due to dust or scratches on the slide of a microarray chip may occur in the experimental process which lead to corruption or absence of some expression measurements. In the engineering literature, similar problems are inherent in networked control systems (NCS) and sensor networks where packet dropouts and time delays are an unavoidable phenomenon during data transmission [27]. Classical methods fail to solve the filtering and estimation problems for such cases with delays and missing data and cannot accurately infer the underlying network structure.

In this paper, we present a class of GA filters for inferring GRN from data with missing measurement values, which can be modeled in the same unifying framework as in the case of state estimation from one-step or two-step randomly delayed measurements [28]. A general framework is presented through augmenting the state variables and with Gaussian assumptions on the posterior state and missing measurement. To make GRN inference from measurements that contain missing data, we describe the network by a nonlinear model and a measurement model that incorporates the missing data. The inferred parameter set can be used to identify the underlying regulatory network structure.

In the literature, several point-based Gaussian approximation (PBGA) filters have been used for solving the GRN inference problem from DNA microarray gene expression data and genome-wide knockout fitness data [29, 30]; however, there is no solution that outperforms all other counterparts. Thus, one has to pick the filter balancing the estimation performance, implementation complexity, and filter stability. Prominent among the PBGA filters are the cubature Kalman filter (CKF) that makes use of the third-degree cubature rule [31], the unscented Kalman filter (UKF) that makes use of the unscented transformation [30, 32], and the central difference Kalman filter (CDKF) that makes use of the difference rule.

The remainder of this paper is organized as follows. In Section 2, we describe the system model and problem formulation. In Section 3, we describe the corresponding GA filter. In Section 4, we investigate the performance of the proposed algorithm on a synthetic network and a diverse set of in silico networks released as a part of the DREAM project, from which observations can be made for benchmarking purposes [33, 34]. In addition, we present results on a real data obtained from the IRMA network of yeast *Saccharomyces cerevisiae* [35]. Finally, Section 5 concludes the paper.

In this paper, we use the following notations:

1.  $\mathcal{N}(x; \mu, \Sigma)$  denotes the Gaussian probability density function with mean  $\mu$  and covariance  $\Sigma$ .

2.  $\mathbb{E}_g\{\cdot|\mu, \Sigma\}$  denotes the Gaussian integral with respect to  $\mathcal{N}(x; \mu, \Sigma)$ .
3.  $\hat{x}$  represents the estimate of variable  $x$ ,  $\tilde{x} = x - \hat{x}$  is the estimation error, and  $\mathbb{E}[\cdot]$  denotes the expectation operation.
4.  $\mathbf{X}^{-1}$  and  $\mathbf{X}^T$  represent the inverse and transpose of matrix  $\mathbf{X}$ , respectively, and  $I_n$  denotes the  $n$ -dimensional identity matrix.

## 2 Methods

### 2.1 Problem formulation and system model

Gene regulatory networks can be modeled as either static or dynamic systems. In this paper, the state-space model is used which is an instance of the dynamic modeling and can effectively cope with time variations in the gene expression data. Consider a GRN consisting of  $N$  genes. Let  $g_{i,k}$ ,  $i = 1, \dots, N, k = 1, \dots, K$  denote the gene expression level for the  $i$ th gene at time step  $k$  where  $K$  is the total number of data points available. Here, “time” is a discrete index enumerating data points sampled at regular intervals. A well-adopted nonlinear model [25, 30] that captures the gene interactions and the evolution of gene expression values effectively is the discrete-time nonlinear stochastic dynamical system which is proposed in [36] as follows:

$$g_{k,i} = \sum_{j=1}^N a_{ij}g_{k-1,j} + \sum_{j=1}^N b_{ij}f(g_{k-1,j}, \mu_j) + I_{0i} + e_{k-1,i} \quad (1)$$

$i, j = 1, \dots, N, \quad k = 1, \dots, K,$

where  $a_{ij}$  is the linear regulatory coefficient from gene  $j$  to gene  $i$ ,  $b_{ij}$  is the nonlinear regulatory coefficient from gene  $j$  to gene  $i$ ,  $N$  is the total number of genes in the gene network, and  $f(g, \mu)$  is a nonlinear sigmoid function defined as

$$f(g, \mu) = \frac{1}{1 + e^{-\mu g}}, \quad (2)$$

with  $\mu$  being a parameter to be identified and  $I_{0i}$  being the external bias on the  $i$ th gene. The noise vector  $\mathbf{e}_k = [e_{k,1}, e_{k,2}, \dots, e_{k,N}]^T$  is Gaussian distributed with zero mean and covariance matrix  $\mathbf{Q}'_k$ , for  $k = 1, \dots, K$ .

The goal of inference is to estimate the parameters (coefficients) of the model in (1), which form the basis of the GRN. To that end, the state vector is concatenated with the model parameters to form augmented state vector as follows. Denote  $\mathbf{A} = [a_{11}, \dots, a_{1N}, a_{21}, \dots, a_{2N}, \dots, a_{N1}, \dots, a_{NN}]^T$ ,  $\mathbf{B} = [b_{11}, \dots, b_{1N}, b_{21}, \dots, b_{2N}, \dots, b_{N1}, \dots, b_{NN}]^T$ ,  $\boldsymbol{\mu} = [\mu_1, \dots, \mu_N]^T$  and  $\mathbf{I}_0 = [I_{01}, \dots, I_{0N}]^T$  and we denote the expression level for all genes at time step  $k$  by  $\mathbf{g}_k = [g_{1,k}, \dots, g_{N,k}]^T$ . Then, the augmented state vector can be described by

$$\mathbf{x}_k \triangleq \left[ \mathbf{g}_k^T, \mathbf{A}^T, \mathbf{B}^T, \boldsymbol{\mu}^T, \mathbf{I}_0^T \right]^T \in \mathbb{R}^{(2N^2+3N)}. \quad (3)$$

The augmented version of the state transition equations include (1) and the following

$$\begin{aligned} a_{ij,k} &= a_{ij,k-1}, \quad b_{ij,k} = b_{ij,k-1}, \\ \mu_{i,k} &= \mu_{i,k-1}, \quad I_{0i,k} = I_{0i,k-1}, \quad i, j = 1, \dots, N. \end{aligned} \quad (4)$$

Succinctly, the state transition of the dynamic model is written as

$$\mathbf{x}_k = f(\mathbf{x}_{k-1}) + \mathbf{w}_{k-1}, \quad (5)$$

where  $f(\cdot)$  is the nonlinear function associated with (2) and (4);  $\mathbf{w}_k = [e_{k,1}, \dots, e_{k,N}, 0, \dots, 0]$  is the augmented noise vector with covariance matrix  $\mathbf{Q}_k = \text{diag}([\mathbf{Q}'_k \quad \mathbf{0}_{2N+2N^2}])$ , where  $\mathbf{0}_m$  denotes an  $m \times m$  all-zero matrix.

The measured gene expression levels can be modeled as

$$z_k = h(\mathbf{x}_k) + \mathbf{v}_k, \quad (6)$$

where  $z_k$  is the output data from the experiments at time  $k$ ,  $h(\mathbf{x}_k) = \mathbf{g}_k$  and  $\mathbf{v}_k \in \mathbb{R}^N$  is Gaussian distributed noise with zero mean and covariance matrix  $\mathbf{R}_k \in \mathbb{R}^{N \times N}$ .

Now, we consider the case that some measurement outputs  $z_k$ , are missing and the estimation is made from the available measurements,  $y_k$ . We assume that  $z_1$  is available. At time  $k = 2$ , if the measurement output is missing, estimation is done with  $z_1$  and at any time instant  $k \geq 3$ , maximum of two consecutive time points may be missing. In summary, if  $z_k$  is missing estimation is done with  $z_{k-1}$  and if  $z_{k-1}$  is unavailable, estimation is done with  $z_{k-2}$ . Thus, the measurement output at each time can be modeled as [27, 37]

$$y_k = \sum_{d=0}^{\min(k-1,2)} \gamma_k^d z_{k-d} \quad (k \geq 1) \quad (7)$$

with

$$\gamma_k^0 = 1 - \varsigma_k, \quad \gamma_k^1 = \varsigma_k(1 - \varsigma_{k-1}), \quad \text{and} \quad \gamma_k^2 = \varsigma_k \varsigma_{k-1}, \quad (8)$$

where  $\varsigma_1 = 0$ ,  $\varsigma_k$  is a Bernoulli random variable with probability  $p(\varsigma_k = 1)(k \geq 2) = q$ . Moreover, it is assumed that  $\mathbf{x}_0, \{\mathbf{w}_k, k \geq 0\}, \{\mathbf{v}_k, k \geq 1\}, \{\varsigma_k, k \geq 2\}$  are mutually independent. Denote  $p_k^d(d = 0, 1, 2)$  as the probabilities that measurements  $z_k, z_{k-1}$ , and  $z_{k-2}$  are used at time  $k$ . Then, we have

$$\begin{aligned} p_k^0 &\triangleq p(\gamma_k^0 = 1) = \mathbb{E}[\gamma_k^0] = 1 - q, \\ p_k^1 &\triangleq p(\gamma_k^1 = 1) = \mathbb{E}[\gamma_k^1] = q(1 - q), \\ p_k^2 &\triangleq p(\gamma_k^2 = 1) = \mathbb{E}[\gamma_k^2] = q^2, \end{aligned} \quad (9)$$

Finally, (5)–(8) describe the dynamic model we propose for inferring GRNs with one-step or two-step missing measurements.

To estimate the GRN based on (5)–(8), we solve the optimal filtering problem by finding the estimator  $\mathbb{E}[x_k|Y_k]$ , where  $Y_k \triangleq (y_1, \dots, y_k)$ . With the Bayes rule, the conditional probability density function (PDF)  $p(x_k|Y_k)$ , and subsequently its first two moments, i.e.,  $\hat{x}_{k|k} = \mathbb{E}[x_k|Y_k]$  and  $P_{k|k}^{xx} = \mathbb{E}[\tilde{x}_{k|k}\tilde{x}_{k|k}^T|Y_k]$ , are recursively obtained through estimating the posterior predictive PDF of the state  $p(x_k|Y_{k-1})$  and the measurement  $p(y_k|Y_{k-1})$ , where  $\tilde{x} = x - \hat{x}$  is the estimation error. For the purpose of filtering, we will make use of the following Gaussian assumptions:

1. The one-step posterior predictive PDF of the state  $x_k$  conditioned on  $Y_{k-1}$  is Gaussian, i.e.,

$$p(x_k|Y_{k-1}) = \mathcal{N}(x_k; \hat{x}_{k|k-1}, P_{k|k-1}^{xx}), \quad (10)$$

where

$$\hat{x}_{k|k-1} = \mathbb{E}[x_k|Y_{k-1}], \quad P_{k|k-1}^{xx} = \mathbb{E}[\tilde{x}_{k|k-1}\tilde{x}_{k|k-1}^T|Y_{k-1}]. \quad (11)$$

2. The one-step posterior predictive PDF of  $y_k$  conditioned on  $Y_{k-1}$  is Gaussian, i.e.,

$$p(y_k|Y_{k-1}) = \mathcal{N}(y_k; \hat{y}_{k|k-1}, P_{k|k-1}^{yy}), \quad (12)$$

where

$$\hat{y}_{k|k-1} = \mathbb{E}[y_k|Y_{k-1}], \quad P_{k|k-1}^{yy} = \mathbb{E}[\tilde{y}_{k|k-1}\tilde{y}_{k|k-1}^T|Y_{k-1}]. \quad (13)$$

## 2.2 Gaussian approximation filters with missing measurements

In this section, we briefly present the general GA filtering framework for the PBGA filters with one-step or two-step missing measurements for the state-space dynamic model. In Additional file 1, we detail its derivation, we review different numerical techniques for approximating multidimensional Gaussian weighted integrals that involve nonlinear transformation of random vectors, and we present the algorithm that implements the UKF version of the filter. Given all the measurements up to the present time in the system described in (5) and (6), the standard Gaussian filter operates by updating only the posterior PDF of the state, i.e.,  $p(x_k|Y_k)$  [38]. However, in the case that the measurements are randomly delayed (or missing) by one or two sampling times as described in (7), apart from  $p(x_k|Y_k)$ , the posterior PDFs  $p(v_k|Y_k)$ ,  $p(x_{k-1}|Y_k)$ , and  $p(v_{k-1}|Y_k)$  also must be updated. Specifically, substituting (6) and (8) into (7), we obtain

$$y_k = \sum_{d=0}^2 \gamma_k^d [h(x_{k-d}) + v_{k-d}] \quad (k \geq 3). \quad (14)$$

Substituting (14) into (13) to incorporate the delayed measurement in the GA filter, whereby  $\hat{y}_{k|k-1}$  and  $P_{k|k-1}^{yy}$  depend on the estimates  $\hat{x}_{k-d}$  and  $\hat{v}_{k-d}$ ,  $d = 0, 1, 2$ . By the Gaussian assumptions, it boils down to computing the first two moments of  $p(v_{k-1}|Y_{k-1})$ ,  $p(x_{k-2}|Y_{k-1})$ , and  $p(v_{k-2}|Y_{k-1})$ . This is achieved through augmenting the state  $x_k$  as follows:

$$x_k^a = \begin{bmatrix} x_k \\ v_k \end{bmatrix}, \quad \mathfrak{X}_k = \begin{bmatrix} x_{k-1}^a \\ x_k^a \end{bmatrix}. \quad (15)$$

Given the Gaussian approximations to  $p(x_k|Y_k)$ ,  $p(v_k|Y_k)$ ,  $p(x_{k-1}|Y_k)$ , and  $p(v_{k-1}|Y_k)$ , the posterior PDFs  $p(x_{k-1}^a|Y_k)$ ,  $p(x_k^a|Y_k)$ , and  $p(\mathfrak{X}_k|Y_k)$  of the augmented states  $x_{k-1}^a$ ,  $x_k^a$ , and  $\mathfrak{X}_k$  are approximated as Gaussian respectively as

$$\begin{aligned} p(x_{k-1}^a|Y_k) &= \mathcal{N}(x_{k-1}^a; \hat{x}_{k-1|k}^a, P_{k-1|k}^{aaa}), \\ p(x_k^a|Y_k) &= \mathcal{N}(x_k^a; \hat{x}_{k|k}^a, P_{k|k}^{aaa}), \\ p(\mathfrak{X}_k|Y_k) &= \mathcal{N}(\mathfrak{X}_k; \hat{\mathfrak{X}}_{k|k}, P_{k|k}^{\mathfrak{X}\mathfrak{X}}), \end{aligned} \quad (16)$$

where

$$\hat{x}_{k-1|k}^a = \begin{bmatrix} \hat{x}_{k-1|k} \\ \hat{v}_{k-1|k} \end{bmatrix}, \quad P_{k-1|k}^{aaa} = \begin{bmatrix} P_{k-1|k}^{xx} & P_{k-1|k}^{xv} \\ (P_{k-1|k}^{xv})^T & P_{k-1|k}^{vv} \end{bmatrix},$$

$$\text{with } P_{k-1|k}^{xv} = \mathbb{E}[\tilde{x}_{k-1|k}\tilde{v}_{k-1|k}^T|Y_k], \quad (17)$$

$$\hat{x}_{k|k}^a = \begin{bmatrix} \hat{x}_{k|k} \\ \hat{v}_{k|k} \end{bmatrix}, \quad P_{k|k}^{aaa} = \begin{bmatrix} P_{k|k}^{xx} & P_{k|k}^{xv} \\ (P_{k|k}^{xv})^T & P_{k|k}^{vv} \end{bmatrix}, \quad (18)$$

$$\text{with } P_{k|k}^{xv} = \mathbb{E}[\tilde{x}_{k|k}\tilde{v}_{k|k}^T|Y_k],$$

and

$$\begin{aligned} \hat{\mathfrak{X}}_{k|k} &= \begin{bmatrix} \hat{x}_{k-1|k}^a \\ \hat{x}_{k|k}^a \end{bmatrix}, \quad P_{k|k}^{\mathfrak{X}\mathfrak{X}} = \begin{bmatrix} P_{k-1|k}^{aaa} & P_{k-1,k|k}^{aaa} \\ (P_{k-1,k|k}^{aaa})^T & P_{k|k}^{aaa} \end{bmatrix}, \\ \text{with } P_{k-1,k|k}^{aaa} &= \mathbb{E}[\tilde{x}_{k-1|k}^a\tilde{x}_{k|k}^{aT}|Y_k]. \end{aligned} \quad (19)$$

As with the general GA filtering, the filtering procedure consists of the *state update* and *measurement update*.

### 2.2.1 State update

Given the augmented state PDF  $p(\mathfrak{X}_{k-1}|Y_{k-1})$  at time  $k-1$ , with its mean and covariance defined as

$$\begin{aligned} \hat{\mathfrak{X}}_{k-1|k-1} &= \begin{bmatrix} \hat{x}_{k-2|k-1}^a \\ \hat{x}_{k-1|k-1}^a \end{bmatrix}, \\ P_{k-1|k-1}^{\mathfrak{X}\mathfrak{X}} &= \begin{bmatrix} P_{k-2|k-1}^{aaa} & P_{k-2,k-1|k-1}^{aaa} \\ (P_{k-2,k-1|k-1}^{aaa})^T & P_{k-1|k-1}^{aaa} \end{bmatrix}, \end{aligned}$$

$$\text{with } P_{k-2,k-1|k-1}^{aaa} = \mathbb{E}[\tilde{x}_{k-2|k-1}^a\tilde{x}_{k-1|k-1}^{aT}|Y_{k-1}], \quad (20)$$

the predicted conditional PDF is  $p(\mathfrak{X}_k|Y_{k-1}) = \mathcal{N}(\mathfrak{X}_k; \hat{\mathfrak{X}}_{k|k-1}, \mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}})$ , with

$$\hat{\mathfrak{X}}_{k|k-1} = \begin{bmatrix} \hat{x}_{k-1|k-1}^a \\ \hat{x}_{k|k-1} \\ 0_{N \times 1} \end{bmatrix}$$

$$\mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}} = \begin{bmatrix} \mathbf{P}_{k-1|k-1}^{aa} & \mathbf{P}_{k-1,k|k-1}^{ax} & \mathbf{0}_{(2N^2+4N) \times N} \\ (\mathbf{P}_{k-1,k|k-1}^{ax})^T & \mathbf{P}_{k|k-1}^{xx} & \mathbf{0}_{(2N^2+3N) \times N} \\ 0_{N \times (2N^2+4N)} & 0_{N \times (2N^2+3N)} & \mathbf{R}_k \end{bmatrix} \quad (21)$$

where  $\hat{x}_{k-1|k-1}^a$  and  $\mathbf{P}_{k-1|k-1}^{aa}$  in (21) are available from  $\hat{\mathfrak{X}}_{k-1|k-1}$  and  $\mathbf{P}_{k-1|k-1}^{\mathfrak{X}\mathfrak{X}}$  in (20), and

$$\hat{x}_{k|k-1} = \mathbb{E}_g\{f(x_{k-1})|\hat{\mathfrak{X}}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}^{\mathfrak{X}\mathfrak{X}}\},$$

$$\mathbf{P}_{k|k-1}^{xx} = \mathbb{E}_g\{f(x_{k-1})f_{k-1}^T(x_{k-1})|\hat{\mathfrak{X}}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}^{\mathfrak{X}\mathfrak{X}}\} - \hat{x}_{k|k-1}\hat{x}_{k|k-1}^T + \mathbf{Q}_{k-1},$$

$$\mathbf{P}_{k-1,k|k-1}^{ax} = \mathbb{E}_g\{x_{k-1}^a f_{k-1}^T(x_{k-1})|\hat{\mathfrak{X}}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}^{\mathfrak{X}\mathfrak{X}}\} - \hat{x}_{k-1|k-1}^a \hat{x}_{k|k-1}^T. \quad (22)$$

For the detailed derivations, see Additional file 1.

### 2.2.2 Measurement update

After obtaining the approximation to the predictive PDF  $p(\mathfrak{X}_k|Y_{k-1})$ , the Gaussian approximation of the augmented state posterior PDF  $p(\mathfrak{X}_k|Y_k)$  is obtained by the Kalman filter equations:

$$\hat{\mathfrak{X}}_{k|k} = \hat{\mathfrak{X}}_{k|k-1} + \mathbf{K}_k^{\mathfrak{X}}(y_k - \hat{y}_{k|k-1}),$$

$$\mathbf{P}_{k|k}^{\mathfrak{X}\mathfrak{X}} = \mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}} - \mathbf{K}_k^{\mathfrak{X}} \mathbf{P}_{k|k-1}^{yy} (\mathbf{K}_k^{\mathfrak{X}})^T, \quad (23)$$

$$\mathbf{K}_k^{\mathfrak{X}} = \mathbf{P}_{k|k-1}^{\mathfrak{X}y} (\mathbf{P}_{k|k-1}^{yy})^{-1},$$

where  $\mathbf{K}_k^{\mathfrak{X}}$  is the Kalman gain and

$$\hat{y}_{k|k-1} = \sum_{d=0}^{\min(k-1,2)} p_k^d \hat{z}_{k-d|k-1},$$

$$\mathbf{P}_{k|k-1}^{yy} = \sum_{d=0}^{\min(k-1,2)} p_k^d \mathbf{P}_{k-d|k-1}^{zz} + \sum_{d=0}^{\min(k-1,2)} (p_k^d \hat{z}_{k-d|k-1} \hat{z}_{k-d|k-1}^T - \hat{y}_{k|k-1} \hat{y}_{k|k-1}^T),$$

$$\mathbf{P}_{k|k-1}^{\mathfrak{X}y} = \sum_{d=0}^{\min(k-1,2)} p_k^d \mathbf{P}_{k,k-d|k-1}^{\mathfrak{X}z}. \quad (24)$$

The delayed/missing measurement statistics  $\hat{z}_{k-d|k-1}$ ,  $\mathbf{P}_{k-d|k-1}^{zz}$ , and  $\mathbf{P}_{k,k-d|k-1}^{\mathfrak{X}z}$  are defined as follows.

For  $d = 0$ :

$$\hat{z}_{k|k-1} = \mathbb{E}_g\{h(x_k)|\hat{\mathfrak{X}}_{k|k-1}, \mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}}\},$$

$$\mathbf{P}_{k|k-1}^{zz} = \mathbb{E}_g\{h(x_k)h_k^T(x_k)|\hat{\mathfrak{X}}_{k|k-1}, \mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}}\} - \hat{z}_{k|k-1}\hat{z}_{k|k-1}^T + \mathbf{R}_k,$$

$$\mathbf{P}_{k|k-1}^{\mathfrak{X}z} = \mathbb{E}_g\{\mathfrak{X}_k(h_k(x_k) + v_k)^T|\hat{\mathfrak{X}}_{k|k-1}, \mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}}\} - \hat{\mathfrak{X}}_{k|k-1}\hat{z}_{k|k-1}^T, \quad (25)$$

for  $d = 1$ :

$$\hat{z}_{k-1|k-1} = \mathbb{E}_g\{h(x_{k-1} + v_{k-1})|\hat{\mathfrak{X}}_{k|k-1}, \mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}}\},$$

$$\mathbf{P}_{k-1|k-1}^{zz} = \mathbb{E}_g\{(h(x_{k-1}) + v_{k-1})(h(x_{k-1}) + v_{k-1})^T|\hat{\mathfrak{X}}_{k|k-1}, \mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}}\} - \hat{z}_{k-1|k-1}\hat{z}_{k-1|k-1}^T,$$

$$\mathbf{P}_{k,k-1|k-1}^{\mathfrak{X}z} = \mathbb{E}_g\{\mathfrak{X}_k(h(x_{k-1}) + v_{k-1})^T|\hat{\mathfrak{X}}_{k|k-1}, \mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}}\} - \hat{\mathfrak{X}}_{k|k-1}\hat{z}_{k|k-1}^T, \quad (26)$$

and for  $d = 2$ :

$$\hat{z}_{k-2|k-1} = \mathbb{E}_g\{h(x_{k-2} + v_{k-2})|\hat{\mathfrak{X}}_{k|k-1}, \mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}}\},$$

$$\mathbf{P}_{k-2|k-1}^{zz} = \mathbb{E}_g\{(h(x_{k-2}) + v_{k-2})(h(x_{k-2}) + v_{k-2})^T|\hat{\mathfrak{X}}_{k|k-1}, \mathbf{P}_{k|k-1}^{\mathfrak{X}\mathfrak{X}}\} - \hat{z}_{k-2|k-1}\hat{z}_{k-2|k-1}^T,$$

$$\mathbf{P}_{k,k-2|k-1}^{\mathfrak{X}z} = \begin{bmatrix} \mathbb{E}_g\{x_{k-1}^a z_{k-2}^T|Y_{k-1}, \hat{\mathfrak{X}}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}^{\mathfrak{X}\mathfrak{X}}\} \\ \mathbb{E}_g\{x_k z_{k-2}^T|Y_{k-1}, \hat{\mathfrak{X}}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}^{\mathfrak{X}\mathfrak{X}}\} \\ 0_{N \times N} \end{bmatrix} - \hat{\mathfrak{X}}_{k|k-1}\hat{z}_{k-2|k-1}^T. \quad (27)$$

The filtering estimate  $\hat{\mathfrak{X}}_{k|k}$  and covariance  $\mathbf{P}_{k|k}^{\mathfrak{X}\mathfrak{X}}$  of the system state are obtained from  $\hat{\mathfrak{X}}_{k|k}$  and  $\mathbf{P}_{k|k}^{\mathfrak{X}\mathfrak{X}}$  respectively. (See Additional file 1 for derivations).

However, the Gaussian weighted integrals in (22) and (25)–(27) contain nonlinear functions which render the analytical calculation infeasible and the algorithm becomes intractable. To deal with this, we employ the point-based numerical integration techniques, which is presented in Additional file 1.

## 3 Results

We assess the proposed algorithm using both synthetic data and real data. Gold standards or the ground-truths are provided for both categories of data and the inferred networks are “benchmarked” against the gold standards. Benchmarking is done by counting the number of links correctly predicted by the algorithm (true positives, TP), the number of incorrectly predicted links (false positives, FP), the number of true links missed in the inferred network (false negatives, FN), and the number of correctly identified non-existing links (true negatives, TN). Thus, the following performance metrics will be defined

accordingly: true positive rate or recall also known as the sensitivity ( $\text{TPR} = \text{TP}/(\text{TP}+\text{FN})$ ), positive predictive value or precision ( $\text{PPV} = \text{TP}/(\text{TP}+\text{FP})$ ), and false positive rate ( $\text{FPR} = \text{FP}/(\text{FP}+\text{TN})$ , where specificity =  $1-\text{FPR}$ ). All the metrics are computed for different thresholds and the area under the receiver operating characteristic (AUROC) curve and the area under the precision-recall (AUPR) curve are estimated to illustrate the overall inference performance of the algorithms. As the inference result comprises of the estimates of both the linear and nonlinear regulatory coefficients among the genes, if at least one of the regulatory coefficients between any two genes is recovered, the link is designated as TP.

In addition,  $y_1 = z_1$ ; at time  $k = 2$  the measurement output can be missing by one-step; and at any time instant  $k \geq 3$  it can be missing by one-step or two-step. With the prior knowledge of the number of missing data points to be replaced in the experimental output, an estimate of the value of  $q$ , the success probability of the Bernoulli variable  $\varsigma_k$  can be made. Specifically, if the number of missing data points is less than 20% of the total number of data points, a  $q$  value chosen in the interval  $[0.05, 0.2]$  is a good choice. In our experiments,  $q = 0.1$ , so that the probability that  $z_k$  is used in the estimation is  $p_k^0 = 0.9$ , the probability that  $z_{k-1}$  is used in the estimation is  $p_k^1 = 0.09$ , and the probability that  $z_{k-2}$  is used in the estimation is  $p_k^2 = 0.01$ . In the remainder of this paper, we denote the datasets that have no missing values as the complete measurements (CM) and we denote the datasets with missing but replaced data points as the missing measurements (MM). The MM is created in the following manner: at time  $k$ , if  $z_k$  is missing and  $z_{k-1}$  is available, we replace  $z_k$  with  $z_{k-1}$ ; otherwise, we replace  $z_k$  with  $z_{k-2}$ , as there can be maximum of two consecutive missing data points in the measurement.

### 3.1 Synthetic network

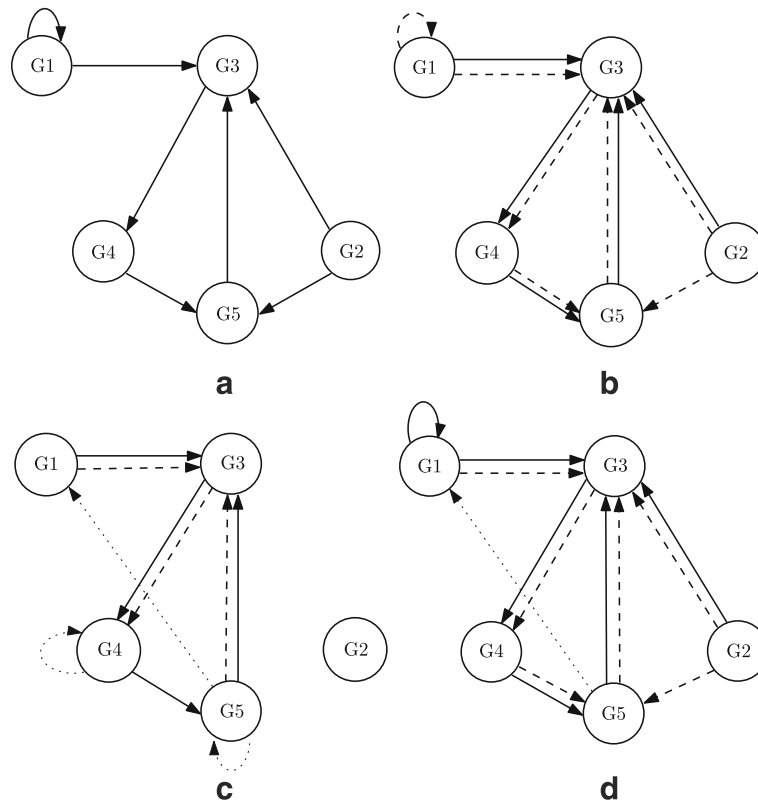
The synthetic network in Fig. 1a is assumed to have both linear and nonlinear connections. The dynamics of the network are based on the model given by (5)–(8), with arrows denoting the direction of regulatory interactions. The parameters of the network, i.e., the linear connection coefficients (LCC) and the nonlinear connection coefficients (NCC), are given in the second column in Table 1 with the NCC in parentheses. The underlying zero-mean Gaussian process noise has a covariance matrix  $\mathbf{Q}_k = 0.004\mathbf{I}$ , and the zero-mean Gaussian measurement noise has a covariance matrix  $\mathbf{R}_k = 0.001\mathbf{I}$ ,  $k = 1, \dots, M$ . Time series data are generated for a total of  $M = 50$  time points. To quantify the results more rigorously, we set the noise threshold at 40% of the maximal variation for linear and nonlinear coefficients such that if an inferred link is less than this threshold, it is considered noise and subsequently filtered off. In the end, we come up with sparse

networks and the TPR and PPV metrics are calculated for the networks.

First, we supplied the CM data to the UKF algorithm. The inferred model parameters are shown in the third column in Table 1, with the NCC in parentheses. The corresponding network is displayed in Fig. 1b where the solid edges indicate the inferred linear connections and the dashed edges indicate the inferred nonlinear connections. Next, we create the MM data by removing data points 10, 11, 25, 35, 36, and 40 from the time series data; the removed data points are then replaced accordingly. To investigate the impact of missing data points on the performance of inference algorithms, we supplied the MM data to the UKF algorithm. The inferred model parameters are shown in the fourth column in Table 1 and the network structure is shown in Fig. 1c. The black dotted arrows indicate the false positives, i.e., incorrectly predicted links. Finally, using the same MM data we tested the proposed UKF with one-step or two-step missing measurements (UKFMM). The inferred model parameters are shown in the fifth column in Table 1 and the inferred network is displayed in Fig. 1d. It is observed that the missing data points have great impact on the performance of the UKF algorithm; whereas the proposed UKFMM algorithm can deal with the missing data effectively by displaying a robust performance which is in fact at par with the performance of the UKF with CM. To average out the influence of random data deletion, we run the experiment 1000 times, where at each run, we randomly deleted up to five data points, with maximum of two consecutive data points, and replaced the deleted data points in similar manner as described above. For all the runs, we record the TPR and the PPV, and the average TPR and PPV with their standard deviations (shown in parentheses) are shown in Table 2.

### 3.2 DREAM4 in silico gene regulatory networks

In order to assess the performance of GRN inference algorithms, several in silico gene networks have been produced as the benchmarking data sets, specifically, the DREAM in silico gene networks [39–41]. We made use of the 10-gene networks by the DREAM4 challenge to test the efficacy of the proposed algorithm. All networks and data were generated with version 2.0 of GeneNetWeaver (GNW) [42]. In total, there are five separate networks, each with 10 genes, whose topologies were extracted from the known GRNs in *Escherichia coli* and *Saccharomyces cerevisiae*. The time series measurements were generated using parametrized stochastic differential equations (SDEs), with observations uniformly sampled (21 time points, single replicate) under five different perturbations, for a total of 105 observations per gene. The inference is performed by using all the perturbations. Self-interaction/autoregulatory edges were not expected in the



**Fig. 1** Synthetic network. *Solid black edges* denote the linear connections, *dashed blue edges* denote the nonlinear connections, and the *dotted red arrows* indicate false positives. **a** Gold standard for the synthetic network. **b** Inferred linear and nonlinear connections by the UKF with CM. **c** Inferred linear and nonlinear connections by the UKF with MM. **d** Inferred linear and nonlinear connections with the proposed UKFMM with MM

predictions and were subsequently removed. Since the number of possible edges in an  $N$ -gene network without autoregulatory interactions is  $N(N - 1)$ , the length of a complete list of predictions is 90 edges for a network of size 10 [33, 34].

We first test the UKF algorithm on the five 10-gene network data sets (CM) and the result is shown in column 2 in Table 3. To average out the influence of random data deletion, we ran 1000 experiments where at each run, we created the MM by randomly deleting up to five data

**Table 1** Network parameters for the synthetic network

Edge	LCC and NCC	UKF with CM	UKF with MM	UKFMM with MM
(1,1)	0.5 (0.4)	— (0.5880)	— —	0.7313 —
(3,1)	0.5 (0.4)	0.3837 (0.4391)	0.7357 (0.5043)	0.4079 (0.4223)
(3,2)	0.5 (0.4)	0.7380 (0.4390)	— —	0.7380 (0.4192)
(3,5)	0.5 (0.4)	0.6098 (0.4391)	0.6623 (0.5043)	0.8285 (0.4354)
(4,3)	0.5 (0.4)	0.7257 (0.3059)	0.3953 (0.2123)	0.7256 (0.3235)
(5,2)	0.5 (0.4)	— (0.3837)	— —	— (0.3235)
(5,4)	0.5 (0.4)	0.6677 (0.3839)	0.5813 (0.3706)	0.7850 (0.3464)
<b>(4,4)</b>	— —	— —	— <b>(0.8417)</b>	— —
<b>(5,1)</b>	— —	— —	— <b>(0.5916)</b>	— —
<b>(5,5)</b>	— —	— —	— <b>(0.3705)</b>	— —
<b>(1,5)</b>	— —	— —	— —	<b>0.5722</b> —

Parameters of the synthetic network and the networks inferred by the UKF algorithm with CM, UKF algorithm with MM, and the proposed UKFMM with MM. The bold edges do not exist in the original network. The false negatives are represented by (non-bold) dashes, and false positives are given in bold numbers

**Table 2** Average TPR and PPV for the synthetic network (standard deviations are shown in parentheses)

	UKF with CM	UKF with MM	UKFMM with MM
TPR	1.00	0.48 (0.035)	0.91 (0.017)
PPV	1.00	0.53 (0.028)	0.86 (0.013)

points, with maximum of two consecutive data points, and replaced the deleted data points accordingly. For each run, we fed both the UKF and the proposed UKFMM algorithms with the MM and we record the average AUROC and AUPR scores for each of the five networks, where the empirical averages and standard deviations over 1000 experiments are shown in columns three and six, respectively in Table 3. Again, it is seen from Table 3 that the proposed UKFMM algorithm is robust against the missing data conditions where it can infer the network as accurately as the UKF algorithm that uses the CM.

We also compared our algorithm against a relevant computational method designed for the GRN network inference, i.e., [43], which is based on the use of Bayesian analysis with ordinary differential equations (ODEs) and non-parametric Gaussian process, an algorithm referred to as GP4GRN. The inference result of GP4GRN with CM is shown in the fourth column in Table 3. Similarly, we tested GP4GRN with the MM where we ran 1000 experiments. At each run, we created the MM by randomly deleting up to five data points with maximum of two consecutive data points and replaced the deleted data points accordingly. The averages and standard deviations of AUROC and AUPR are obtained and the corresponding results are summarized in the fifth column in Table 3. We conclude that the GP4GRN method has comparable performance to the UKF in all data sets, and similarly it is outperformed by the proposed UKFMM algorithm under missing data conditions.

### 3.3 *Saccharomyces cerevisiae* IRMA network

*Saccharomyces cerevisiae* GAL network in yeast is one of the most prominent model systems due to its importance for the studies of eukaryotic regulation and relatively self-contained nature [44–47]. A synthetic GRN that contains

5 genes has previously been constructed in the budding yeast [35]. In the well studied network, popularly referred to as in vivo reverse engineering and modeling assessment (IRMA) network, each of the genes regulate at least one other gene in the network. Expression within the network is activated in the presence of galactose and then switched to glucose to obtain the switch-off data which represents the expressive samples at 21 time points. The switch-on data consists of 16 sample points and is obtained by growing the cells in a glucose medium and then changing to galactose.

The true interactions is shown in Fig. 2a. The real biological data is first supplied to the UKF algorithm and the inferred network is shown in Fig. 2b. As standard, some data points are randomly discarded from the input and they are replaced accordingly to generate the MM. The UKF and the proposed algorithm UKFMM are tested on the generated data set (MM) and the inferred networks are shown in Fig. 2c, d, respectively, and the corresponding results are summarized in Table 4. Again, on the missing data condition, the proposed algorithm shows a better performance compared to the UKF. In addition, we also test the GP4GRN algorithm with both CM and MM and the results are presented in the fourth and fifth columns in Table 4, which further affirms the impact of missing measurements in the GRN inference methods and the relative robustness of the proposed UKFMM algorithm.

## 4 Discussion

This work presents a novel algorithm for GRN inference from time-series gene expression data with *one-step* or *two-step missing measurements*. Gene regulation is assumed to follow a nonlinear state evolution model described in (1). The parameters of the model, which are assumed to be the regulatory coefficients between the genes, are estimated with a modified unscented Kalman filtering algorithm. We considered the experimental scenarios that lead to total loss of expression values for all genes at a particular time point or few successive time points which may significantly diminish the performance of GRN inference algorithms.

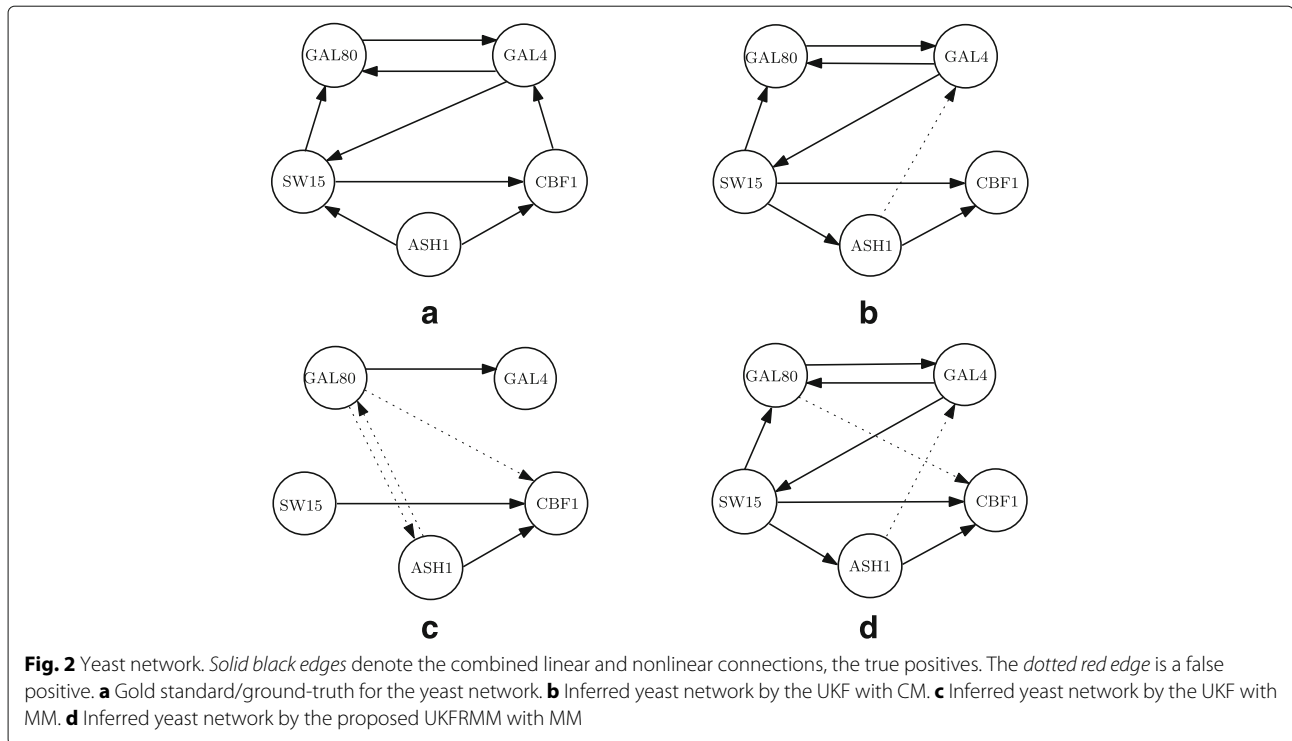
In the proposed algorithm, the state vector which is the gene expression at each time point in (1) is concatenated

**Table 3** AUROC and AUPR curves for the DREAM4 networks

	UKF with CM	UKF with MM	GP4GRN with CM	GP4GRN with MM	UKFMM with MM
N1	[0.63] [0.42]	[0.44(0.024)][0.24(0.020)]	[0.66] [0.42]	[0.42(0.027)][0.29(0.021)]	[0.61(0.015)][0.42(0.008)]
N2	[0.67] [0.49]	[0.48(0.018)][0.26(0.017)]	[0.69] [0.44]	[0.44(0.015)][0.28(0.018)]	[0.64(0.013)][0.44(0.011)]
N3	[0.72] [0.50]	[0.45(0.020)][0.30(0.012)]	[0.70] [0.47]	[0.50(0.022)][0.33(0.016)]	[0.72(0.021)][0.53(0.012)]
N4	[0.75] [0.52]	[0.56(0.019)][0.28(0.011)]	[0.62] [0.35]	[0.36(0.031)][0.25(0.027)]	[0.72(0.009)][0.50(0.010)]
N5	[0.81] [0.44]	[0.53(0.021)][0.26(0.019)]	[0.86] [0.65]	[0.55(0.022)][0.40(0.019)]	[0.80(0.012)][0.42(0.014)]

Column 1 shows the network number. In columns 3, 5, and 6, average AUROC and average AUPR are presented in the *square brackets* and the standard deviations are in *parentheses*





with the model parameters and an augmented state vector in (3) is defined for the joint estimation of gene expression values and system parameters. We consider the possibility that each real measurement is randomly missing and the estimation is made from the available measurements. The use of the UKF, an instance of the PBGA filters, for the state and parameter estimation renders the algorithm computationally efficient and capable of working offline or online (when all the measurements are readily available, or they become available successively, respectively). The proposed algorithm is tested on both synthetic and real biological data to evaluate the efficacy of the predictions. From the series of results obtained for both synthetic data and the real biological data, we conclude that the gene network structure can be inferred from time series data with missing values.

In this paper, we have applied the proposed algorithm to the time series data generated from the DNA microarray because to our best of knowledge, DNA microarray is still of interest in transcriptome profiling due to its reduced cost and widespread use as compared to the RNA-seq. In addition, it has been shown that there is there is high

correlation between the gene expression profiles generated between the DNA microarray and RNA-seq [48, 49]. Hence, the proposed method can easily be extended to time series gene expression data from RNA-seq.

In general, this work addresses the possibility of having *one-step or two-step missing expression values* by considering them as the delayed observations of the full set of genes. Future work will focus on the inference of the structure of a (potentially larger) network by incorporating a general *s-step missing values* for *s-consecutive* time points, which may address more complex missing data scenarios.

### 5 Conclusions

Time series gene expression data be modeled with state-space model and the model parameters can be estimated using different GA filters. Unfortunately, there are situations which result in loss of expression values for all genes at a particular time point or few successive time points. In this case, conventional filtering approach fails to correctly estimate the model parameters, which are used to elucidate the underlying GRN. We have proposed PBGA

**Table 4** AUROC and AUPR curve for the yeast networks

	UKF with CM	UKF with MM	GP4GRN with CM	GP4GRN with MM	Proposed UKFRMM
AUROC curve	0.70	0.42	0.76	0.49	0.68
AUPR curve	0.46	0.34	0.57	0.38	0.46

filters that treat the missing measurement values as a set of delayed measurements and demonstrated that the modified filter can estimate the model parameters, with missing measurements, as accurate as the conventional filter with no missing measurements.

## Additional file

**Additional file 1:** Supplemental Material for "Reverse engineering gene regulatory networks from measurement with missing values". (PDF 207 kb)

## Abbreviations

AUPR: Area under precision-recall; AUROC: Area under the receiver operating characteristic; CKF: Cubature Kalman filter; CDKF: Central difference Kalman filter; CM: Complete measurements; DBN: Dynamic Bayesian network; DNA: Deoxyribonucleic acid; DREAM: Dialogue on reverse engineering assessment and methods; EKF: Extended Kalman filter; FN: False negative; FP: False positive; FPR: False positive rate; GA: Gaussian approximation; GRN: Gene regulatory network; GWN: GeneNetWeaver; IRMA: In-vivo reverse-engineering and modeling assessment; LCC: Linear connection coefficient; MM: Missing measurements; NCC: Nonlinear connection coefficient; NCS: Networked control systems; ODE: Ordinary differential equation; PBGA: Point-based Gaussian approximation; PDF: Probability density function; PPV: Positive predictive value; RNA: Ribonucleic acid; SDE: Stochastic differential equation; TN: True negative; TP: True positive; TPR: True positive rate; UKF: Unscented Kalman filter; UKFMM: Unscented Kalman filter with one-step or two-step missing measurements; UT: Unscented transform

## Acknowledgements

Not applicable.

## Availability of data and materials

DREAM4 in silico challenge datasets analyzed during the current study are available for download at <http://gnw.sourceforge.net/dreamchallenge.html>. IRMA in vivo dataset analyzed during the current study is available for download at [http://www.cell.com/supplemental/S0092-8674\(09\)00156-1](http://www.cell.com/supplemental/S0092-8674(09)00156-1).

## Authors' contributions

OEO developed the idea, performed the implementation of the algorithm, ran most of the experiments, performed the data analysis and interpretation, and heavily involved in the writing of the manuscript. AE performed some of the simulation experiments, was involved with data analysis and interpretation, and was a major contributor in writing the manuscript. XW conceived the idea, contributed significantly to its development, supervised its implementation, and performed a thorough revision of 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.

Received: 11 August 2016 Accepted: 15 December 2016

Published online: 10 January 2017

## References

- El Palmero, SGP de Campos, M Campos, NC Souza, IDC Guerreiro, AL Carvalho, MMC Marques, Mechanisms and role of microRNA deregulation in cancer onset and progression. *Genet. Mol. Biol.* **34**(3), 363–370 (2011)
- D Baek, J Villén, C Shin, FD Camargo, SP Gygi, DP Bartel, The impact of microRNAs on protein output. *Nature*. **455**(7209), 64–71 (2008)
- J Jin, K He, X Tang, Z Li, L Lv, Y Zhao, J Luo, G Gao, An arabidopsis transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors. *Mol. Biol. Evol.* **32**(7), 1767–1773 (2015)
- VB Teif, K Rippe, Statistical–mechanical lattice models for protein–DNA binding in chromatin. *J. Phys. Condensed Matter*. **22**(41), 414105 (2010)
- RE Moellering, M Cornejo, TN Davis, C Del Bianco, JC Aster, SC Blacklow, AL Kung, DG Gilliland, GL Verdine, JE Bradner, Direct inhibition of the notch transcription factor complex. *Nature*. **462**(7270), 182–188 (2009)
- A-L Barabási, N Gulbahce, J Loscalzo, Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**(1), 56–68 (2011)
- LE Chai, SK Loh, ST Low, MS Mohamad, S Deris, Z Zakaria, A review on the computational approaches for gene regulatory network construction. *Comput. Biol. Med.* **48**, 55–65 (2014)
- F Emmert-Streib, M Dehmer, B Haibe-Kains, Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front. Cell Dev. Biol.* **2**, 38 (2014)
- N Vijesh, SK Chakrabarti, J Sreekumar, et al., Modeling of gene regulatory networks: a review. *J. Biomed. Sci. Eng.* **6**(02), 223 (2013)
- M Hecker, S Lambeck, S Toepfer, E Van Someren, R Guthke, Gene regulatory network inference: data integration in dynamic models? A review. *Biosystems*. **96**(1), 86–103 (2009)
- CE Gagna, WC Lambert, Novel multistranded, alternative, plasmid and helical transitional dna and RNA microarrays: implications for therapeutics. *Pharmacogenomics*. **10**(5), 895–914 (2009)
- W Zhao, E Serpedin, ER Dougherty, Inferring connectivity of genetic regulatory networks using information-theoretic criteria. *Comput. Biol. Bioinform.* *IEEE/ACM Trans.* **5**(2), 262–274 (2008)
- J Dougherty, I Tabus, J Astola, Inference of gene regulatory networks based on a universal minimum description length. *EURASIP J. Bioinforma. Syst. Biol.* **2008**(1), 1 (2008)
- B Godsey, Improved inference of gene regulatory networks through integrated bayesian clustering and dynamic modeling of time-course expression data. *PLoS ONE*. **8**(7), 68358 (2013)
- CD Giurcarneau, I Tabus, J Astola, J Ollila, M Vihinen, Fast iterative gene clustering based on information theoretic criteria for selecting the cluster structure. *J. Comput. Biol.* **11**(4), 660–682 (2004)
- S Kauffman, C Peterson, B Samuelsson, C Troein, Random boolean network models and the yeast transcriptional network. *Proc. Natl. Acad. Sci.* **100**(25), 14796–14799 (2003)
- X Yang, JE Dent, C Nardini, An s-system parameter estimation method (SPeM) for biological networks. *J. Comput. Biol.* **19**(2), 175–187 (2012)
- OR Gonzalez, C Küper, K Jung, PC Naval, E Mendoza, Parameter estimation using simulated annealing for s-system models of biochemical networks. *Bioinformatics*. **23**(4), 480–486 (2007)
- I Shmulevich, ER Dougherty, W Zhang, From boolean to probabilistic boolean networks as models of genetic regulatory networks. *Proc. IEEE*. **90**(11), 1778–1792 (2002)
- Y Huang, J Wang, J Zhang, M Sanchez, Y Wang, Bayesian inference of genetic regulatory networks from time series microarray data using dynamic Bayesian networks. *J. Multimedia*. **2**(3), 46–56 (2007)
- T-F Liu, W-K Sung, A Mittal, Model gene network by semi-fixed Bayesian network. *Expert Syst. Appl.* **30**(1), 42–49 (2006)
- J Angus, M Beal, J Li, C Rangel, D Wild, in *Learning and Inference in Computational Systems Biology*, ed. by ND Lawrence, M Girolami, M Rattray, and G Sanguinetti. Inferring transcriptional networks using prior biological knowledge and constrained state-space models (MIT Press, Cambridge, 2010), pp. 117–152
- O Hirose, R Yoshida, S Imoto, R Yamaguchi, T Higuchi, DS Charnock-Jones, S Miyano, et al., Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*. **24**(7), 932–942 (2008)
- RE Kalman, A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82**(1), 35–45 (1960)
- Z Wang, X Liu, Y Liu, J Liang, V Vinciotti, An extended Kalman filtering approach to modeling nonlinear dynamic gene regulatory networks via short gene expression time series. *IEEE/ACM Trans. Comput. Biol. Bioinformatics (TCBB)*. **6**(3), 410–419 (2009)
- MSB Sehgal, I Gondal, LS Dooley, R Coppel, How to improve postgenomic knowledge discovery using imputation. *EURASIP J. Bioinforma. Syst. Biol.* **2009**(1), 1 (2009)
- S Sun, L Xie, W Xiao, N Xiao, Optimal filtering for systems with multiple packet dropouts. *Circ. Syst. II Express Briefs IEEE Trans.* **55**(7), 695–699 (2008)

28. M Sahebsara, T Chen, SL Shah, Optimal filtering with random sensor delay, multiple packet dropout and uncertain observations. *Int. J. Control.* **80**(2), 292–301 (2007)
29. A Noor, E Serpedin, M Nounou, H Nounou, Reverse engineering sparse gene regulatory networks using cubature Kalman filter and compressed sensing. *Adv. Bioinformatics.* **2013**, 205763 (2013)
30. L Wang, X Wang, AP Arkin, MS Samoilov, Inference of gene regulatory networks from genome-wide knockout fitness data. *Bioinformatics.* **29**(3), 338–346 (2013)
31. I Arasaratnam, S Haykin, Cubature Kalman filters. *IEEE Trans. Autom. Control.* **54**(6), 1254–1269 (2009)
32. EA Wan, R Van Der Merwe, in *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000. The unscented Kalman filter for nonlinear estimation* (IEEE, Alberta, Canada, 2000), pp. 153–158
33. RJ Prill, D Marbach, J Saez-Rodriguez, PK Sorger, LG Alexopoulos, X Xue, ND Clarke, G Altan-Bonnet, G Stolovitzky, Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE.* **5**(2), 9202 (2010)
34. D Marbach, RJ Prill, T Schaffter, C Mattiussi, D Floreano, G Stolovitzky, Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci.* **107**(14), 6286–6291 (2010)
35. I Cantone, L Marucci, F Iorio, MA Ricci, V Belcastro, M Bansal, S Santini, M Di Bernardo, D Di Bernardo, MP Cosma, A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell.* **137**(1), 172–181 (2009)
36. L Chen, K Aihara, Chaos and asymptotical stability in discrete-time neural networks. *Physica D Nonlinear Phenomena.* **104**(3), 286–325 (1997)
37. A Hermoso-Carazo, J Linares-Pérez, Unscented filtering algorithm using two-step randomly delayed observations in nonlinear systems. *Appl. Math. Model.* **33**(9), 3705–3717 (2009)
38. K Ito, K Xiong, Gaussian filters for nonlinear filtering problems. *IEEE Trans. Autom. Control.* **45**(5), 910–927 (2000)
39. D Marbach, T Schaffter, C Mattiussi, D Floreano, Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J. Comput. Biol.* **16**(2), 229–239 (2009)
40. G Stolovitzky, RJ Prill, A Califano, Lessons from the DREAM2 challenges. *Ann. N. Y. Acad. Sci.* **1158**(1), 159–195 (2009)
41. G Stolovitzky, D Monroe, A Califano, Dialogue on reverse-engineering assessment and methods. *Ann. N. Y. Acad. Sci.* **1115**(1), 1–22 (2007)
42. T Schaffter, D Marbach, D Floreano, Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics.* **27**(16), 2263–2270 (2011)
43. T Äijö, H Lähdesmäki, Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics.* **25**(22), 2937–2944 (2009)
44. O Egriboz, F Jiang, JE Hopper, Rapid gal gene switch of *Saccharomyces cerevisiae* depends on nuclear gal3, not nucleocytoplasmic trafficking of GAL3 and GAL80. *Genetics.* **189**(3), 825–836 (2011)
45. M Johnston, JS Flick, T Pexton, Multiple mechanisms provide rapid and stringent glucose repression of GAL gene expression in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **14**(6), 3834–3841 (1994)
46. D Lohr, P Venkov, J Zlatanova, Transcriptional regulation in the yeast gal gene family: a complex genetic network. *FASEB J.* **9**(9), 777–787 (1995)
47. S Ostergaard, L Olsson, J Nielsen, Metabolic engineering of *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **64**(1), 34–50 (2000)
48. S Zhao, W-P Fung-Leung, A Bittner, K Ngo, X Liu, Comparison of RNA-seq and microarray in transcriptome profiling of activated t cells. *PLoS ONE.* **9**(1), 78644 (2014)
49. S Kogenaru, Q Yan, Y Guo, N Wang, RNA-seq and microarray complement each other in transcriptome profiling. *BMC Genomics.* **13**(1), 1 (2012)

Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)

---