BioSystems xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*

Cheng-Long Chuang[a,b,1], Chung-Ming Chen[a,1], Wai-Si Wong[c], Kun-Nan Tsai[a], Err-Cheng Chan[d], Joe-Air Jiang[b,*]

[a] *Institute of Biomedical Engineering, National Taiwan University, Taiwan*
[b] *Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan*
[c] *Kiang Wu Nursing College of Macau, Macau*
[d] *Department of Medical Biotechnology and Laboratory Science, Chang Gung University, Taiwan*

ARTICLE INFO

Article history:
Received 30 November 2008
Received in revised form 8 May 2009
Accepted 28 May 2009

Keywords:
Genetic interactions
Gene regulatory networks
Nonlinear fitting
Microarray gene expression

ABSTRACT

In order to identify genes involved in complex diseases, it is crucial to study the genetic interactions at the systems biology level. By utilizing modern high throughput microarray technology, it has become feasible to obtain gene expressions data and turn it into knowledge that explains the regulatory behavior of genes. In this study, an unsupervised nonlinear model was proposed to infer gene regulatory networks on a genome-wide scale. The proposed model consists of two components, a robust correlation estimator and a nonlinear recurrent model. The robust correlation estimator was used to initialize the parameters of the nonlinear recurrent curve-fitting model. Then the initialized model was used to fit the microarray data. The model was used to simulate the underlying nonlinear regulatory mechanisms in biological organisms. The proposed algorithm was applied to infer the regulatory mechanisms of the general network in *Saccharomyces cerevisiae* and the pulmonary disease pathways in *Homo sapiens*. The proposed algorithm requires no prior biological knowledge to predict linkages between genes. The prediction results were checked against true positive links obtained from the YEASTRACT database, the TRANSFAC database, and the KEGG database. By checking the results with known interactions, we showed that the proposed algorithm could determine some meaningful pathways, many of which are supported by the existing literature.

1. Introduction

Large-scale analyses of gene regulatory networks are built from a number of edges and vertices. The validation of whole-genome functional interactions requires labor, time, and an adequate budget. Thus, bioinformatics research plays an essential role in automating the reconstruction of genetic regulatory networks using available biological information (Zweigenbaum et al., 2007). In recent years, the advance in research of modern molecular biotechnology has allowed scientists to design various types of microarrays, enabling a deeper insight into the processes of gene regulatory mechanisms on a whole-genome scale (Zhong and Sternberg, 2006). With an abundance of biological data, a number of computational approaches have been proposed to analyze these data, and more importantly, to infer functional interactions among gene regulatory networks. Discovering the underlying regulatory process in an organism is a difficult task because the functional gene–gene interactions are more complex than those of physical protein–protein interactions. Therefore, a successful computational model might help with speeding up the genome-wide exploration or with the validation of the genetic interactions.

Genome-wide gene expression analysis has been widely used to investigate the gene regulatory networks of a given organism (Spellman et al., 1998). Many models have been proposed to analyze time-course gene expression data. Clustering has been applied to gene expression data to find co-expressed or co-regulated genes, examples include k-means, hierarchical clustering (Eisen et al., 1998), self-organizing maps (Törönen et al., 1999), bioclustering (Getz et al., 2000), principal component analysis (Raychaudhuri et al., 2000), and independent component analysis (Liebermeister, 2002). Some approaches incorporate sequence data and functional annotations to reconstruct gene clusters (Ihmels et al., 2002; Pilpel et al., 2001). Some earlier studies were built based on model-based approaches, such as random Boolean networks (Liang et al., 1998), Bayesians networks (Dojer et al., 2006; Friedman et al., 2000;

* Corresponding author. Tel.: +886 2 3366 5341; fax: +886 2 2362 7620.
E-mail address: jajiang@ntu.edu.tw (J.-A. Jiang).
[1] These two authors contributed equally to this work.

0303-2647/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2009.05.013

Hughes et al., 2000; Friedman, 2004; Imoto et al., 2002a,b; Perrin et al., 2003; Tamada et al., 2003; Segal et al., 2003; Husmeier, 2003; Kim et al., 2004; Beal et al., 2005; Lähdesmäki et al., 2006; Lähdesmäki and Shmulevich, 2008), and state-space models (Rangel et al., 2004a,b; Yamaguchi and Higuchi, 2006; Yamaguchi et al., 2007) to infer the architecture of gene regulatory networks. Some researchers used a system of differential equations to model the behavior of the regulatory networks, such as deterministic differential systems (Chen and Aihara, 2002; de Jong, 2002; Kimura et al., 2005), linear differential systems (Wu et al., 2004), nonlinear differential systems (Vu and Vohradsky, 2007), and machine learning approaches (Chuang et al., 2007, 2008). These models have been shown to be informative for understanding genetic interactions. However, their disadvantages are that some of them provide no direct way to obtain information about how multiple cooperating transcriptional factors (TFs) interact with a specific target gene. Furthermore, some approaches suffer from excessive computational overhead due to complicated mathematical modeling not suitable for applications of whole-genome analysis. Graphic models (Whittaker, 1990), on the other hand, provide the possibility to infer gene networks using a simplified model formulation. In recent studies (Wong et al., 2003; Dobra et al., 2004; Schafer and Strimmer, 2005), the theory of graphical models has been extended such that it is possible to infer large-scale gene networks using small sample data.

Traditional techniques for microarray technology have provided valuable insights towards deciphering gene regulatory networks in living organisms. In recent years, the ChIP-on-chip technique has become more popular in identifying physical interactions between TFs and the upstream sequence of genes on a genome-wide basis. Many earlier proposed studies (Simon et al., 2001; Lee et al., 2002; Harbison et al., 2004) investigated the transcriptional regulatory elements and networks using ChIP-on-chip data. However, further analyses are required to obtain information regarding the role of a given TF (as an activator or as a repressor).

In this study, we proposed an unsupervised computational model, which consists of a robust correlation estimator and a nonlinear recurrent model, to infer gene regulatory interactions using time-course gene expression data. The robust correlation estimator assesses the similarity of the expression patterns of all possible paired genes. The nonlinear recurrent curve-fitting model, which is initialized by the robust correlation estimator, infers the gene regulatory interactions by discovering the causal effects of the gene expression levels. The remainder of this paper is organized as follows. In the following sections, we describe the derivations of our approach, followed by the experimental results of the proposed method in *Saccharomyces cerevisiae* and *Homo sapiens*. Then the data sets and normalization procedures used in this work are presented. The gene interactions confirmed by YEASTRACT database (Teixeira et al., 2006) and TRANSFAC database (Matys et al., 2003), and KEGG database (Kanehisa et al., 2006) are then used to evaluate the performance of the proposed approach. Finally, a discussion regarding the advantages of the proposed approach and future work for possible improvement are provided in Section 5.

2. Methods

2.1. Overview

In this study, we develop a nonlinear recurrent model fitting approach to infer genetic interactions on a whole-genome scale. The proposed approach encompasses two primary components: a robust correlation estimator and a nonlinear recurrent model. The primary goal of the former was to initialize the model parameters, and the latter was to detect regulatory linkages using time-course gene expression data. The conceptual system diagram of the proposed approach is shown in Fig. 1. The proposed model is a graphical model. Since some of essential model parameters are estimated *a priori*, the prediction result yielded by the proposed approach is very stable under the same data set. Detailed descriptions of the robust correlation estimator and the nonlinear recurrent model are given in the following subsections.

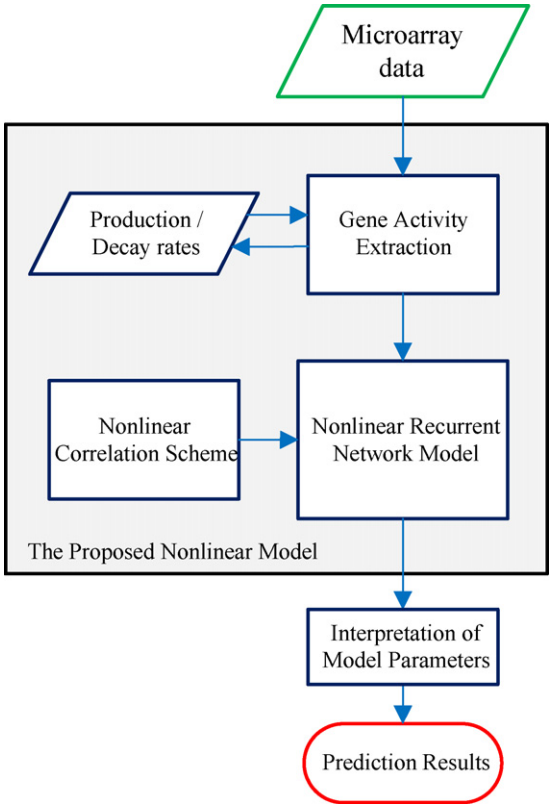

Fig. 1. Conceptual block diagram of the proposed approach.

2.2. Robust correlation estimator

The expression levels of genes are commonly regulated by the level of transcription initiation. Most studies focus on exploring the interactions between significantly expressed genes. However, it is known that some genes with low expression levels (e.g., housekeeping genes) might play an essential role in pathway regulation. To reveal the underlying function of those genes, a standardization step for gene expression data is applied. Let $G_x(t)$ denote the gene expression of gene x at time point t. We can then transform $G_x(t)$ into the standardized form by

$$\hat{G}_x(t) = \frac{G_x(t) - \mu_x}{\sigma_x},$$

where $\hat{G}_x(t)$ is the standardized expression level of gene x at time t, and μ_x and σ_x are the mean and the standard deviation of $G_x(t)$, respectively. Note that this step is optional. Those who tend to focus on the interaction between significantly expressed genes can skip the standardization step.

Then, we take the 1st-order derivatives of the standardized gene expression levels with respect to t. The primary goal of this step is to extract the gene expression pattern without influencing the baseline shift effect. In addition, it is widely known that gene expression data are nonlinear. Many famous metrics, such as Euclidian distance (Krause, 1986) and Pearson correlation (Biedl et al., 2001), have been proven to be ineffective in the analysis of gene expression data. To remove the nonlinearity in the data, we slice $\hat{G}_x(t)$ into small pieces, and extract the expression features of a pair of genes. The 1st-order partial derivative with respect to t can be obtained by

$$\hat{G}'_x(t) = \frac{\partial \hat{G}_x(t)}{\partial t}.$$

Since time-course gene expression data $\hat{G}_x(t)$ are discretized in time t, the partial derivative terms can be reformulated by

$$\hat{G}'_x(t) = \frac{\partial \hat{G}_x(t)}{\partial t} = \frac{\hat{G}_x(t + \Delta t) - \hat{G}_x(t)}{\Delta t},$$

where $t + \Delta t$ denotes the $(t + \Delta t)$th time point and Δt is the time interval between time points t and $t + \Delta t$, and $\Delta t = 1$. For a pair of genes $\{A, B\}$, where A is the regulating gene and B is the target gene, we use the procedure of the linear regression model

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

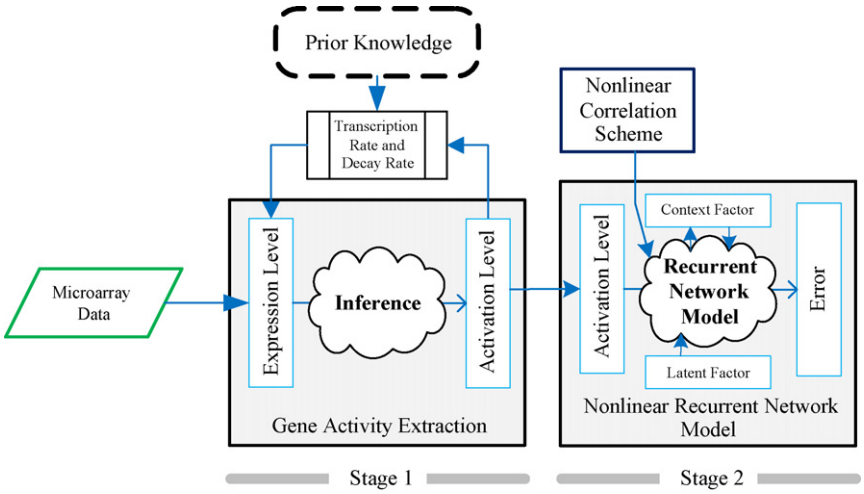

Fig. 2. The schematic view of the nonlinear recurrent network model. The model consists of two stages, gene activity extraction and recurrent network model.

to fit $\hat{G}'_A(t)$ and $\hat{G}'_B(t)$ by

$$\hat{G}'_B(t) = \beta_0 + \beta_{\{A,B\}}\hat{G}'_A(t) + \varepsilon(\beta_0, \beta_{\{A,B\}}, t),$$

where β_0 is the intercept, $\beta_{\{A,B\}}$ is the slope of the regression line, and $\varepsilon(\beta_0, \beta_{\{A,B\}}, t)$ denotes the error term. Estimation of β_0 and $\beta_{\{A,B\}}$ can be obtained by the method of the general least squares estimator (LSE). The LSE estimates β_0 and $\beta_{\{A,B\}}$ by minimizing the sum of the squares of the residuals $\sum_t \varepsilon^2(\beta_0, \beta_{\{A,B\}}, t) = \sum_t (\hat{G}'_B(t) - \beta_0 - \beta_{\{A,B\}}\hat{G}'_A(t))^2$. However, the LSE approach is known to have a major drawback in that it is very sensitive to outliers. To overcome this problem, here, we utilize a robust linear regression model (Huber, 1981) to reduce the influence of outliers. Thereby we can reformulate the error term by the Cauchy–Lorentz distribution (Spiegel, 1992) as follows:

$$\varepsilon^*(\beta_0, \beta_{\{A,B\}}, t) = \frac{1}{1 + \varepsilon^2(\beta_0, \beta_{\{A,B\}}, t)} = \frac{1}{1 + (\hat{G}'_B(t) - \beta_0 - \beta_{\{A,B\}}\hat{G}'_A(t))^2}. \quad (1)$$

Since the 1st-order partial derivative step neutralizes the baseline shift effect, we can omit β_0 from Eq. (1). The optimal $\beta_{\{A,B\}}$ can be obtained by following the procedures of the robust linear regression model. The reason for using the Cauchy–Lorentz distribution is that the probability density function of the Cauchy–Lorentz distribution suppresses the error caused by the outliers and leads the regression line to closer to the data points. This is one of the major robust regression techniques that can easily be implemented by scientific computing software, e.g., MATLAB, S+, or R. Indeed, there are other robust regression models can be used in this step to obtain similar values of optimal $\beta_{\{A,B\}}$, e.g., least median squares estimator (Rousseeuw, 1984) and least trimmed squares estimators (Rousseeuw, 1985).

After we obtain the optimal value of the parameter $\beta_{\{A,B\}}$, which is denoted by $\hat{\beta}_{\{A,B\}}$, the robust correlation coefficient in the time-course gene expression levels of paired genes $\{A, B\}$ can be calculated by transforming $\hat{\beta}_{\{A,B\}}$ into an interval ranging from 1 to −1 as

$$r_{\mathrm{rLSE}}(G_A, G_B) = \begin{cases} \dfrac{\hat{\beta}_{\{A,B\}}}{\max(\sigma_A/\sigma_B, \sigma_B/\sigma_A)} & \text{if } |\hat{\beta}_{\{A,B\}}| \le 1 \\ \dfrac{1/\hat{\beta}_{\{A,B\}}}{\max(\sigma_A/\sigma_B, \sigma_B/\sigma_A)} & \text{otherwise} \end{cases},$$

where $r_{\mathrm{rLSE}}(G_A, G_B)$ is the robust correlation coefficient in the expression levels of paired genes $\{A, B\}$. The positive (negative) sign of $r_{\mathrm{rLSE}}(G_A, G_B)$ suggests that paired genes $\{A, B\}$ have a similar (reverse) pattern of variation in their expression level. The amplitude of $r_{\mathrm{rLSE}}(G_A, G_B)$ measures the strength of the pattern. For example, $r_{\mathrm{rLSE}}(G_A, G_B) = 1$ (−1) represents paired genes $\{A, B\}$ have a perfect positive (negative) correlation, and 0 means that there is no significant relation between paired genes $\{A, B\}$. The coefficient r_{rLSE} plays an essential role in the initialization of the nonlinear recurrent model, which is described as follows.

3. Nonlinear recurrent model

It has been widely known that the genetic regulatory networks are nonlinear. The regulatory mechanism of genes encompasses complex molecular interactions, including transcription, translation, post-translational modification, transcription cofactor interactions, and unknown factor interference. Thus, the time-course gene expression data from the microarray experiments are nonlinear. In the present paper, we used a nonlinear recurrent model to formulate the effects in a gene regulatory network.

The proposed nonlinear recurrent model is a graphical network model that uses gene expression data to refine the regulatory activity of genes in the organism. Most previous network-based studies initialized model parameters using a random number generator. This may cause the training performance to be unstable. Therefore, in this study we initialize the model parameters by the robust correlation coefficient. The entire model consists of two stages, as shown in Fig. 2.

3.1. Gene activity extraction

In the first stage, a scheme is used to reverse-engineer gene activities on a genome-wide scale using gene expression data from the microarray experiments. The concentrations of the mRNAs reflect the expression levels of the genes. In addition to the famous Michaelis–Menten kinetics (Michaelis and Menten, 1913), there are many mRNA kinetic models that have been proposed to determine the activity level of genes. A nonlinear least square method has been proposed to fit the experimental data for uncovering the mRNA intensities for the repressor-only, activator-only and repressor–activator systems (Nicholas et al., 2006). A stochastic model has been proposed to estimate the mean intensity of gene products based on the effect of oscillatory stimulus (Lipan and Wong, 2005). A correlation-based model was proposed to explain the complex behavior of genes in a systematical gene network (Pedraza and van Oudenaarden, 2005). A quasiequilibrium approximation has been proposed to model the reaction kinetics of fast and slow reactions in stochastic biochemical systems (Haseltine and Rawlings, 2002; Rao and Arkin, 2003; Goutsias, 2005). A Markov chain-based has been proposed for simulating the self-regulated gene with arbitraries feedback mechanism (Fournier et al., 2009). These methods worked well in modeling the mRNA reaction kinetics in biological systems. In this study, we assume that the mRNA degradation and production rates are known *a priori*. To simplify the problem in this study, we developed a simple mRNA reaction kinetics that is able to find the explicitly approximation of the dynamics of the gene expressions.

The mRNA degradation model is formulated by ordinary differential equations that contain the terms of mRNA degradation and production rates.

$$G^*(t') = \exp(-\lambda t'), \quad (2)$$

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

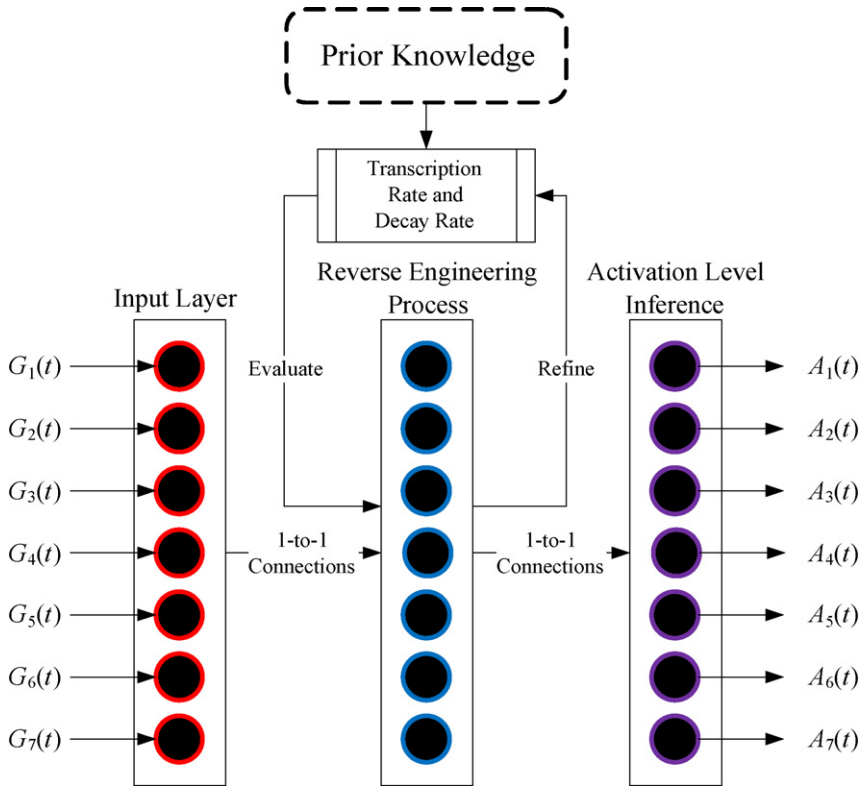

Fig. 3. Process of the gene activity extraction stage.

where $G^*(t')$ represents the quantity of newly produced mRNAs of a given gene at time t', t' denotes the duration of the degradation, $\lambda = \ln(1/2)/t^*$ is the mRNA self-degradation rate, and t^* is the half-life of the mRNA. This forms an mRNA decay model $G^*(t')$ to estimate the remaining mature mRNA quantity at time t' after a certain amount of the mRNAs have been transcribed.The levels of mRNAs are mature mRNAs accumulated over time in the nucleus and in the cytoplasm. Thus, we formulate the expression level of mRNAs by a convolution model

$$G(t) = \int_{-\infty}^{\infty} P(\tau)G^*(t-\tau)d\tau + \varepsilon(t),$$

where $G(t)$ is the expression level of mRNAs, $P(t)$ represents the quantity of mRNAs newly transcribed into the nucleus and the cytoplasm, $G^*(t)$ is the mRNA decay model described in Eq. (2), and $\varepsilon(t)$ is the unknown noise in the data. In order to extract the quantity of mRNAs that have been transcribed at any given time t, we deconvolve the gene expression level $G(t)$ from the mRNA decay model $G^*(t)$ as

$$P(t) = \int_{-\infty}^{\infty} G(\tau)[G^*(t-\tau)]^{inv}d\tau. \quad (3)$$

Note that the unknown noise $\varepsilon(t)$ is eliminated to simplify the model. In this study, we assume that the quantity of mRNAs increases or decreases due to the transcription or degradation processes, respectively. Thus, the value of $P(t)$ should be positive for all time t. If the minimal value of $P(t)$ over time t is negative, it is necessary to adjust the value of the mRNA self-degradation rate λ to satisfy our assumptions. If necessary, we adjust the value of λ by $\lambda^* = \psi\lambda$, where λ^* is the adjusted λ, and ψ is the refinement rate (default value is 0.99). We use the updated λ to compute the mRNA degradation model and $P(t)$. This procedure is repeated until all values of $P(t)$ over t are positive.

According to the Michaelis–Menten kinetics, the maximum rate of reaction mediated by enzymes has the saturation effects at the steady state (Michaelis and Menten, 1913). However, there is no clearly defined method to determine the maximum reaction rate (V_{max}) and the Michaelis constant (K_M) in the kinetics model using only microarray data. In this study, we formulate the mRNA production rate changes by the level of the transcription rate and the magnitude of the regulation from activators and repressors. In addition, we add a saturation characteristic for the mRNA production rate, such that

$$\frac{dP(t)}{dA(t)} = c_1 SA(t)P^{-1}(t)t,$$

where c_1 is a constant, S represents the transcription rate of the gene (mRNA/h), and $A(t)$ is the magnitude of regulation of the gene at time t. Thus, we can derive the magnitude of regulation $A(t)$ is

$$A(t) = \sqrt{\frac{P^2(t) - c_2}{c_1 St}}, \quad (4)$$

where c_2 is the constant of integration. Constants c_1 and c_2 can be used to express Eq. (4) under different initial conditions, and can be solved by two initial conditions: (1) when $t = 1$ min, $A(1) = 1$, $P(1) = 1$, and (2) when $t = 60$ min, $A(60) = 1$, $P(60) = S$. Thus, according to Eq. (4), we can determine the magnitude of regulation $A(t)$ simply by putting the value of the quantity of the newly produced mRNAs $P(t)$ into Eq. (3). The processes described above are shown in Fig. 3.

4. Recurrent network model

4.1. Feedforward prediction process

In complex genetic regulatory networks, observing feedback loops between mRNAs and TFs are common. In this study, we use a nonlinear recurrent network to construct a modulated model with

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

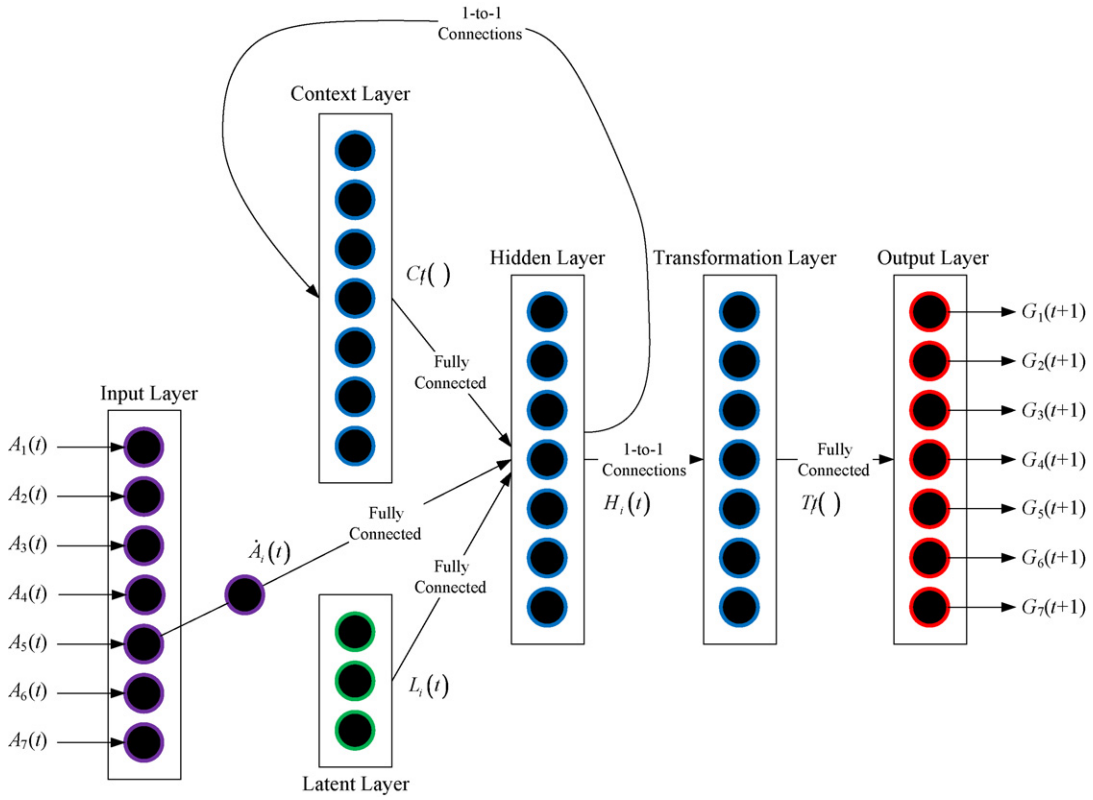

Fig. 4. Architecture of the nonlinear recurrent network model that consists of five layers, which are input, context and latent, hidden, transformation, and output layers.

feedback connections and latent factors that provides the foundation for the accurate representation of gene regulatory networks.

The proposed recurrent model consists of five layers as depicted in Fig. 4. The first layer is the input layer. The magnitudes of regulations of all genes (e.g., $A_i(t)$ for magnitude of regulation of gene i at time t) are input to the model via this layer. We transform $A_i(t)$ into a nonlinear space using a sigmoid function which results in the saturation characteristic of the biochemical reactions in a living organism as

$$\dot{A}_i(t) = sig(A_i(t)s_{i,j}) + b_i, \tag{5}$$

where $\dot{A}_i(t)$ denotes the normalized reaction rate of gene i specified for gene j at time t, sig is the sigmoid function, $s_{i,j}$ is the scale factor associated with the interaction between regulating gene i and target gene j, and b_i is the offset factor for gene i. The scale factor is used to adjust the difference of the change ratios between genes, and the offset factor is utilized to remove the constant offset effect in $A_i(t)$. We can determine the change ratio of the magnitude of the regulation between gene i and gene j over time from

$$s_{i,j} = \left| \frac{\max_t(A_j(t)) - \min_t(A_j(t))}{\max_t(A_i(t)) - \min_t(A_i(t))} \right|.$$

In addition, the value of the constant offset b_i is fixed at 0.

The outputs of the first layer feed directly into the second layer, referred to as the hidden layer. The hidden layer is not only connected to the following layer in a feedforward manner, but also to a further layer, called the context layer, in a simple 1-to-1 connection. To form recurrent connections, the outputs of the context layer $C_i(t)$ are also inputted into the hidden layer. These recurrent connections of the context layer provide the model with a short-term memory (e.g., $C_i(t) = H_i(t-1)$). It allows the hidden layer to not only observe the actual inputs, but also to, via the context layer, obtain information on their own state with time lag-1. Besides the context layer, a latent layer is connected to the hidden layer. The outputs of the latent layer $L_i(t)$ are used to imitate the influences of unknown factors on the regulation of the gene network. Each input of the node in the hidden layer is multiplied by its corresponding weights. Furthermore, the output is the linear sum of the weighted activations from other layers, which is given by

$$H_i(t) = \sum_j w^a_{i,j}\dot{A}_j(t) + \sum_k w^c_{i,k}C_k(t) + \sum_l w^l_{i,l}L_l(t) + h_i,$$

where $H_i(t)$ represents the activation level of gene i, h_i is the offset factor used to formulate the bias influence in the hidden layer, and $w^a_{i,j}$, $w^c_{i,k}$ and $w^l_{i,l}$ are weights associated with connections from the input layer, context layer, and latent layer, respectively. Weights between the input layer and hidden layer $w^a_{i,j}$ are initialized by robust correlation as

$$w^a_{i,j} = \frac{r_{\mathrm{rLSE}}(G_i, G_j)}{\max(\sigma_{G_j}/\sigma_{G_i}, \sigma_{G_i}/\sigma_{G_j})},$$

where $r_{\mathrm{rLSE}}(G_i, G_j)$ is the robust correlation coefficient in the expression levels of paired genes i and j, and σ_{G_i} and σ_{G_j} are variances of the expression levels of paired genes i and j. Other weights are initialized in a random manner. The output of the hidden layer is then fed to the third layer, called the transformation layer. The transformation layer converts the activation level of a gene $H_i(t)$ by a sigmoid function as

$$T_i(t) = sig(H_i(t)),$$

where $T_i(t)$ is the transformed activation level of gene i.

The outputs of the transformation layer are used to predict the activation level of other genes at time $t+1$. Each output is multiplied by its corresponding weights. In addition, each node is connected to the nodes in the output layer in a full-connected manner. Thus, the output of the network is the linear sum of the weighed activations

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

from the transformation layer as given by

$$\hat{A}_i(t+1) = \sum_j w^t_{i,j} T_j(t) + o_i,$$

where $\hat{A}_i(t+1)$ is the predicted activation level of gene i at time $t+1$, $w^t_{i,j}$ denotes the weights associated with the connection between the output layer and transformation layer, and o_i represents the offset factor for the prediction process.

4.2. Tuning process

For a given set of inputs, the squared error produced by given weight and bias distributions can be defined as

$$E_i(t) = (A_i(t+1) - \hat{A}_i(t+1))^2.$$

The error is thus computed backwards across the entire model for each time point. To simplify the notation in this section, time t will no longer be explicitly stated in the rest of this paper. The change of the weight associated with the connection between output node i and transformation node j can be determined by

$$\Delta w^t_{i,j} = \eta \frac{\partial E_i}{\partial w^t_{i,j}} = \eta \frac{\partial E_i}{\partial \hat{A}_i} \frac{\partial \hat{A}_i}{\partial w^t_{i,j}} = 2\eta T_j (A_i - \hat{A}_i),$$

where $\Delta w^t_{i,j}$ is the change of weight, and η is the tuning rate. The change of the bias associated with the same link is

$$\Delta o_i = \eta \frac{\partial E_i}{\partial o_i} = \eta \frac{\partial E_i}{\partial \hat{A}_i} \frac{\partial \hat{A}_i}{\partial o_i} = 2\eta (A_i - \hat{A}_i),$$

where Δo_i is the change of bias. The changes of the weights for the incoming connections to the hidden layer can be described as a linear combination of errors produced by each node produced at the output layer.

$$\Delta w^a_{i,j} = \eta \sum_m \frac{\partial E_m}{\partial w^a_{i,j}} = \eta \sum_m \frac{\partial E_m}{\partial \hat{A}_m} \frac{\partial \hat{A}_m}{\partial T_i} \frac{\partial T_i}{\partial H_i} \frac{\partial H_i}{\partial w^a_{i,j}} = 2\eta \dot{A}_j (1 - \dot{A}_j) \sum_m (A_m - \hat{A}_m) \quad (6)$$

Similar to Eq. (6), the weights associated with the connections between the hidden layer, context layer, and latent layer can be determined by

$$\Delta w^c_{i,k} = 2\eta C_k (1 - C_k) \sum_m (A_m - \hat{A}_m),$$

$$\Delta w^l_{i,l} = 2\eta L_l (1 - L_l) \sum_m (A_m - \hat{A}_m),$$

where $C_k(t) = H_k(t-1)$, and L_l is the activation level of the latent factors. The change in scale factor in Eq. (5) is given by

$$\Delta s_{i,j} = \eta \sum_m \frac{\partial E_m}{\partial s_{i,j}} = \eta \sum_m \frac{\partial E_m}{\partial \hat{A}_m} \frac{\partial \hat{A}_m}{\partial T_i} \frac{\partial T_i}{\partial H_i} \frac{\partial H_i}{\partial \dot{A}_i} \frac{\partial \dot{A}_i}{\partial s_{i,j}} = 2\eta A_i (1 - 4(A_i s_{i,j}) + 6(A_i s_{i,j})^2 - 4(A_i s_{i,j})^3) \sum_m (A_m - \hat{A}_m).$$

After the changes in all parameters are obtained, we can update the parameters as follows:

$$s_{i,j} = s_{i,j} + \Delta s_{i,j}$$
$$w^a_{i,j} = w^a_{i,j} + \Delta w^a_{i,j}$$
$$w^c_{i,j} = w^c_{i,j} + \Delta w^c_{i,j}$$
$$w^l_{i,j} = w^l_{i,j} + \Delta w^l_{i,j}$$
$$o_i = o_i - \Delta o_i$$
$$h_i = h_i - \Delta h_i$$

After we update the model parameters, we can estimate the values of $L_i(t)$ by minimizing the mean square errors of all genes across time, which is given by

$$\Phi = \frac{1}{m} \sum_m \sum_{t=2}^{t_{\max}} E_m(t),$$

where Φ represents the mean square error generated by the model, and $E_m(t)$ is the square error generated by node m in the output layer at time t. Subsequently, in each simulation round (epoch), the updated parameters are used to fit the activation levels of the genes at time $t+1$. The fitting process will loop back to the first time point when t reaches the maximum available time point $t_{\max}$ in the time-course microarray data. An early stopping criterion based on the mean square errors of all genes across time is suggested. We continue the fitting process until the mean square errors is (are) lesser than e.g., 0.001 or until the number of loops reaches 100 epochs.

4.3. Interpretation of the tuned parameters

For the tuned parameters, a new objective is to extract the prediction information embedded inside the model. We use a product of the weights to infer the interaction type and the strength of the link between a given pair of genes. For example, if we consider the interaction between genes i and j, we can determine the properties of the interaction by

$$\Gamma_{i,j} = \sum_m ((s_{i,m} w^a_{i,m}) w^t_{m,j}),$$

where $\Gamma_{i,j}$ denotes the strength of the interaction, and the sign of $\Gamma_{i,j}$ is an indication of the interaction type. There are two types of genetic interactions, the activation–target (AT) interaction and the repressor–target (RT) interaction. AT interactions will result in positive values of $\Gamma_{i,j}$, while for RT interactions $\Gamma_{i,j}$ will result in negative values. In order to distinguish significant links from whole-genome predictions, we may screen out insignificant links using a t-test at the significance level of 1%.

In addition, the latent factors L_i at the latent layer also play an essential role in modeling the hidden components in the gene regulation networks (Sabatti and James, 2005). Similar to a Bayesian version of the state-space model (Beal et al., 2005), these latent factors are analogous to capturing the dynamic changes in gene activations as functions of unobserved biological information, including the activity levels of other unobserved genes or TFs, as well as transcriptional and post-transcription effects. In our study, the analysis resulting from the latent factors L_i are immediately interpreted as the influence of unknown factors that cannot be measured in microarray experiments across time.

5. Results and discussion

5.1. Simulated data

A time-course data from 10-gene regulatory network with 2 latent factors was simulated. The simulation consists of various sample sizes and noise levels. Let $f_i(t)$, $y_i(t)$ and $\varepsilon_i(t)$ denote the

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

expression level of latent factor i, gene i, and noise variable i, respectively. The dynamic factor model for the 6-gene network is defined as:

$$\begin{aligned}
\Delta y_1(t) &= -0.2f_1(t) + 0.5y_1(t-1) + 0.5y_4(t-1) + \varepsilon_1(t)\\
\Delta y_2(t) &= 0.4f_2(t) - 0.5y_2(t-1) + 0.4y_3(t-1) + \varepsilon_2(t)\\
\Delta y_3(t) &= 0.6y_3(t-1) - 0.5y_4(t-1) + \varepsilon_3(t)\\
\Delta y_4(t) &= 0.5y_4(t-1) + 0.5y_5(t-1) + \varepsilon_4(t)\\
\Delta y_5(t) &= 0.4f_1(t) + 0.7y_5(t-1) + \varepsilon_5(t)\\
\Delta y_6(t) &= -0.5y_6(t-1) + 0.5y_7(t-1) + \varepsilon_6(t)\\
\Delta y_7(t) &= -0.5f_2(t) + 0.5y_7(t-1) + 0.4y_8(t-1) + \varepsilon_7(t)\\
\Delta y_8(t) &= -0.5y_8(t-1) + 0.5y_9(t-1) + \varepsilon_8(t)\\
\Delta y_9(t) &= 0.3f_2(t) + 0.5y_9(t-1) - 0.5y_{10}(t-1) + \varepsilon_9(t)\\
\Delta y_{10}(t) &= 0.5f_2(t) - 0.7y_{10}(t-1) + \varepsilon_{10}(t)
\end{aligned} \tag{7}$$

where $f_i(t) \sim N(0, 5)$, $y_i(0) \sim U(100, 500)$, and $\varepsilon_i(t) \sim N(0, {\sigma_i}^2)$, $i = 1, 2, \ldots, 10$. During the simulation, the values of $y_i(t)$ across time t must be positive. The variance of noise ${\sigma_i}^2$ is determined by the variance of $y_i(t)$. The noise level is quantified by a signal-to-noise ratio, where ${\sigma_i}^2 = Var(y_i)/c$, and $c = 10$ or 4 for all i, representing data with signal noise ratio equal to 10 or 4 (denoted by SNR10 and SNR4, respectively). The model used in Eq. (7) is the dynamic of the 10-gene network with 18 true links at time t, thereby the gene expressions at time t are regulated by the latent factors at time t and the gene expressions at time $t-1$. The network roughly follows the sparse property of cis-regulatory networks stated in van Someren et al. (2002). The time-course data were simulated under four different conditions; sample sizes ($T = 50$ or 100) and noise levels (SNR10 or SNR4), where T is the number of time points.

There were 100 experiments were carried out for each simulation condition. For SNR4 group, the averaged true positive rates (TPRs), true negative rates (TNRs) and false positive rates (FPRs) yielded by the proposed algorithm when $T = 50$ (100) are 90.6% (97.1%), 93.2% (95.2%) and 6.8% (4.8%), respectively. For SNR10 group, the proposed algorithm results in averaged TPRs, TNRs, and FPRs about 76.8% (83.9%), 82.6% (89.4%), and 17.4% (10.6%), respectively for $T = 50$ (100). Thus, the simulation studies on the 10-gene network show that the proposed algorithm performs better when SNR is large, which is a reasonable result. In addition, the proposed algorithm is able to yield better result when sample size T is large since a larger sample size gives more power to estimate the model parameters.

6. Inferring genetic interactions in *S. cerevisiae*

6.1. Data preparation

In this study, we used the cell cycle cDNA microarray data in Spellman et al. (1998), sampled from yeast cultures synchronized by two independent methods: (-factor arrest (*alpha*) and elutriation (*elu*). In these data sets, there are, respectively, 18 and 14 time points with no replications. An abundance of number of time points is helpful in (more time points are helpful to) the analysis of genetic interactions of the whole genome. The time-course microarray data of all genes are divided into five groups: M/G_1, G_1, S, G_2, and M. The red (R) and green (G) fluorescence intensities are measured from the mRNA abundance in the experimental group and the control group, respectively. A full description of these data sets and complete normalized and raw data are available at http://cellcycle-www.stanford.edu.

Raw data of the *alpha* and *elu* data sets are freely available in Spellman et al. (1998), and complete normalized data of all data sets are available for public access on the web. In this study, the input of the proposed algorithm is the strength of the fluorescence of wild-type genes in the control group. Thus, we retrieve the fluorescence signals of each gene from the Cy3-Channel (Channel-1) microarray images. Afterward, a location and scale normalization process (Dudoit and Yang, 2002) is applied to the data to adjust the center and the spread of the distribution of the fluorescence signals obtained from the microarray images. Thus, we can obtain a hybrid microarray data with 32 time points. It is worth mentioning that there are two other data sets in Spellman et al. (1998) under the treatments of blocking a *cdc15* and *cdc28* temperature-sensitive mutants. However, the raw data of *cdc28* data set is unavailable, and the experimental conditions of the *cdc15* data set are too complicated. Thus, we will not use the *cdc15* nor the *cdc28* data set to evaluate the performance of the proposed algorithm.

6.2. Prediction of genetic interactions in yeast

Transcriptional regulators (TRs) that activate the transcriptions of DNAs are usually composed of a DNA-binding site and an activation domain. Transcription of DNA is initiated when transcriptional activator protein is bound to its specific DNA-binding sites. Therefore, transcriptional interactions are direct interactions. Because the microarray data retrieved from Spellman et al. (1998) was collected to identify the function of cell cycle-related genes, we focused on predicting cell cycle-related transcriptional interactions annotated in the YEASTRACT (Teixeira et al., 2006) and TRANSFAC (Matys et al., 2003) databases. YEASTRACT and TRANSFAC databases are accessible on the Internet and offer known associations between TFs and genes with biological evidences and/or literature support. We collected 118 and 64 transcriptional interactions from YEASTRACT and TRANSFAC databases, respectively, to form a *posteriori* knowledge data set to benchmark the proposed model. Besides the interactions collected from the databases, we also collected 109 transcriptional interactions by surveying published literature. The literature that we selected is related to the regulation of Cln-Cdc28 kinase. Cln-Cdc28 kinase is important because cyclin-dependent kinase Cdc28 is associated with the Cln3 cyclin and seems to be the most upstream activator (Cln3–Cdc28 complex) of two distinct sets of transcriptional complexes, SBF and MBF. Therefore, cyclin-dependent kinase Cdc28 may play an important role in the control of the cell cycle progression in buddy yeast (Dirick et al., 1995; Futcher, 1996; Levine et al., 1996; Stuart and Wittemberg, 1995; Tyres et al., 1993; Alberghina et al., 1998). Thus, the *a posteriori* knowledge data set is formed by 291 transcriptional interactions (with AT/RT annotations) from various resources, and should be adequate to avoid evaluation bias caused by lack of test data.

Gene expressions of 6116 genes in *S. cerevisiae* are used to fit the proposed model. First, we use the proposed gene activity extraction scheme to compute the activation levels of all genes across time. To obtain the gene activity, we need to set up two parameters, the mRNA self-degradation rate (λ) and the transcription rate of the gene (S, mRNA/h). To model the transcription, most previous studies assume that the production rate of new mRNA is a constant, which is approximately the same for all genes in the cell. However, to obtain more accurate information, parameters λ and S of each gene should be measured individually. Hence, we conducted a series of RNAi experiments that contain siRNA transfection protocol and nuclear run-on analysis to measure the mRNA degradation and production rates of each gene.

In the RNAi pathway for gene silencing, double-stranded RNA (dsRNA) enters the cell through the transcription of endogenous microRNAs (miRNAs), viral delivery, or endocytosis. After dsRNA enters the cell, it encounters the dicer enzyme complex. The dicer cuts the dsRNA into small fragments, called small interfering RNAs. Then, the RNA inference silencing complex (RISC) binds these siRNAs, and uses their hybridization ability to target and degrade longer complementary transcripts. Hence, with selection of efficient targeting sequence, the transcription of a specific gene is then inhibited. Followed by these procedures, RT-PCR experiment is conducted to measure total amount of RNA isolated from the

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

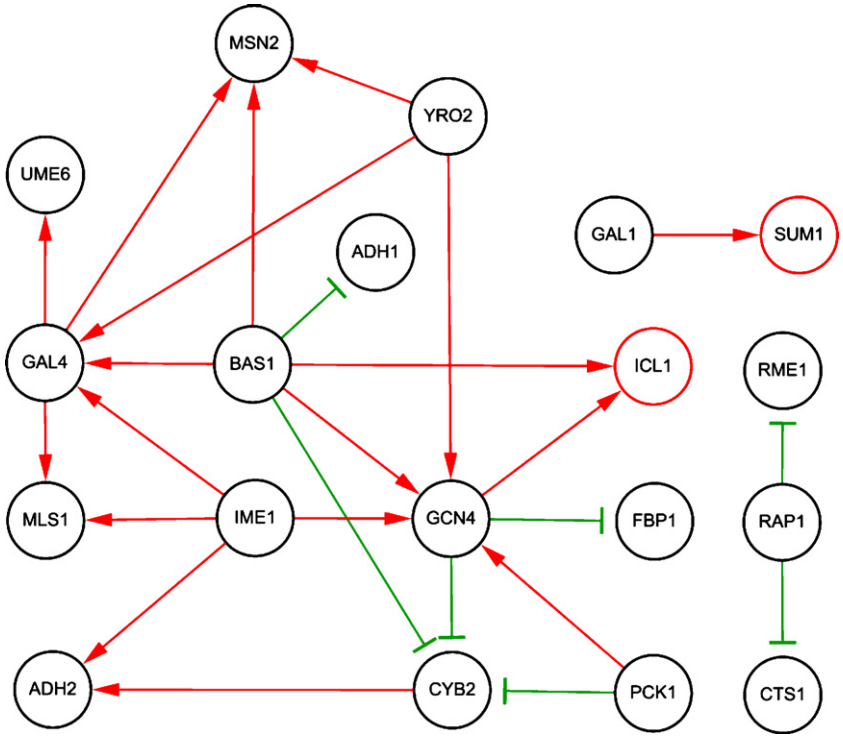

Fig. 5. Top-25 significant links predicted by the proposed model using *alpha* data set. We observed that BAS1, IME1, GCN4, and PCK1 are possible essential regulators that control several genes in the predicted network.

experiment well. We can use this procedure to determine the time it takes for a gene's product to lose half of its quantity. Therefore, the time can be calculated by the aforementioned mRNA self-degradation rate (λ). A nuclear run-on analysis is conducted to identify the quantity of a gene's product that is being transcribed at a certain time. First, we lyse a number of cells to retrieve the nuclei. The transcription process is halted due to the nuclei were removed. We conducted RT-PCR experiment (with reverse transcription) to measure the initial number of mRNAs. Then, we allow transcription to finish and measure how many mRNAs are produced from the stalled RNA polymerase, and then we can measure the transcription rate of gene (S, mRNA/h) after the cells were lysed. By conducting the RNAi experiments, the mRNA production and degradation rates of 4817 and 5119 genes are measured. Thus, in this study we focus on 4817 genes in *S. cerevisiae* which have both λ and S available from the RNAi experiments. The complete list of production and degradation rates of all genes is available at http://bem.bime.ntu.edu.tw/clchuang/CNNM_Supp.pdf. An alternative way for the readers may be to set S as the average production rate of all genes, 7.37 mRNA/h, and λ as the average degradation rate of all genes, as 18.87.

We assume that the number of latent factors of the number of genes in a genome is 1%. Hence, we set the number of latent factors at the latent layer of the proposed model to 48. Scale factors $s_{i,j}$ are initialized by, and bias factors b_i are fixed at 0. Subsequently, we apply the proposed model to fit the activation levels of 4817 genes at the whole-genome scale using *alpha* and *elu* data sets.

Before we report the comparison results between the proposed algorithm and other alternatives, we need to define the performance metrics used in this study. There are 291 true links confirmed via database and literature. These true links signify not only the existence of the links, but also the interaction type (AT or RT) of the link. Hence, the definition of "correctly predicted link" is that an AT (RT) link is predicted to have a positive (negative) score; otherwise, the predicted link is regarded as "incorrectly predicted link". The true positive rate is the number of correctly predicted links divided by the number of true links (which is 291). In addition, we need to ensure that all comparisons are properly controlled for false positive rates. Thus, the prediction accuracy is the number of correctly predicted links divided by the number of significant links, and the significant links are determined by the control of false positive rates.

For the *alpha* data set, 98,087 out of 23,203,489 possible links are predicted to be significant interactions. The true positive rate yielded by the proposed model is 73.9% (215/291). There are 174 out of 291 links are identified as significant links by t-test at two-tailed significant level of 1%. Among the 174 significant links, 147 links are correctly predicted; thereby the prediction accuracy is 84.5% (147/174). Among 98,087 potential links, we select the top-25 significant links to construct a genetic regulatory network as shown in Fig. 5. Since it is not possible to know negative links, we show the plot of the true positive rate versus the prediction accuracy in Fig. 6 instead of the receiver operating characteristic (ROC) curve. We can see that most of the links are related to four genes, which are BAS1, IME1, GCN4, and PCK1. It is known that Bas1 the primary regulator and induces the expression of genes related to purine and histidine biosynthesis pathways (Daignan-Fornier and Fink, 1992). BAS1 is predicted to be a positive regulator for GCN4, and GCN4 has a similar function to regulate the expression of genes involved in purine biosynthesis and glycogen homeostasis (Natarajan et al., 2001). Moreover, GCN4 is also predicted to be a negative regulator of CYB2, and CYB2 is known to be repressed by glucose condition (Lodi and Guiard, 1991). IME1 is a regulator that activates the transcription of early meiotic genes through interaction with UME6 (Guttmann-Raviv et al., 2002). In the predicted network, IME1 is an indirect positive regulator to UME6, and UME6 is also reported as a key regulator of early meiotic genes (Vershon and Pierce, 2000).

Moreover, the proposed model has been applied to the *elu* data set to infer genetic interactions. Among 23,203,489 possible links, 79,102 are predicted to be significant interactions. By checking the prediction results against 291 transcriptional interactions in the *a posteriori* knowledge data set, the proposed model yields a 71.8% (209/291) true positive rate. There are 152 out of 291 links are iden-

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

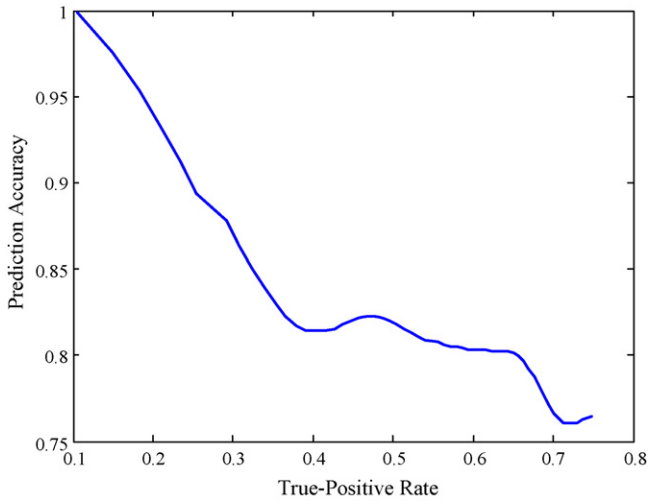

Fig. 6. Plot of true positive rate versus prediction accuracy of the proposed model using *alpha* data set.

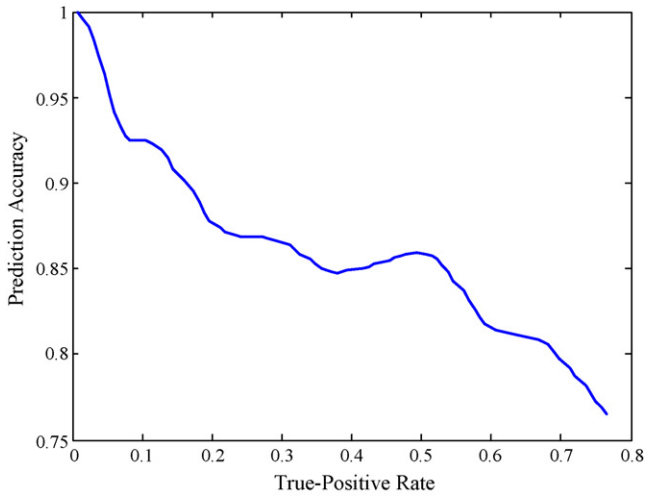

Fig. 8. Plot of true positive rate versus prediction accuracy of the proposed model using *elu* data set.

tified as significant links by *t*-test at two-tailed significant level of 1%. Among the 152 significant links, 133 links are correctly predicted; thereby prediction accuracy is around 87.5% (133/152). The top-30 significant links are selected to construct a genetic regulatory network shown in Fig. 7. The plot of the true positive rate versus the prediction accuracy is shown in Fig. 8. We can see that 8 out of 30 links are outgoing links emitted from SPO12, which implies that it plays an essential role in the regulation of cellular functions. SPO12

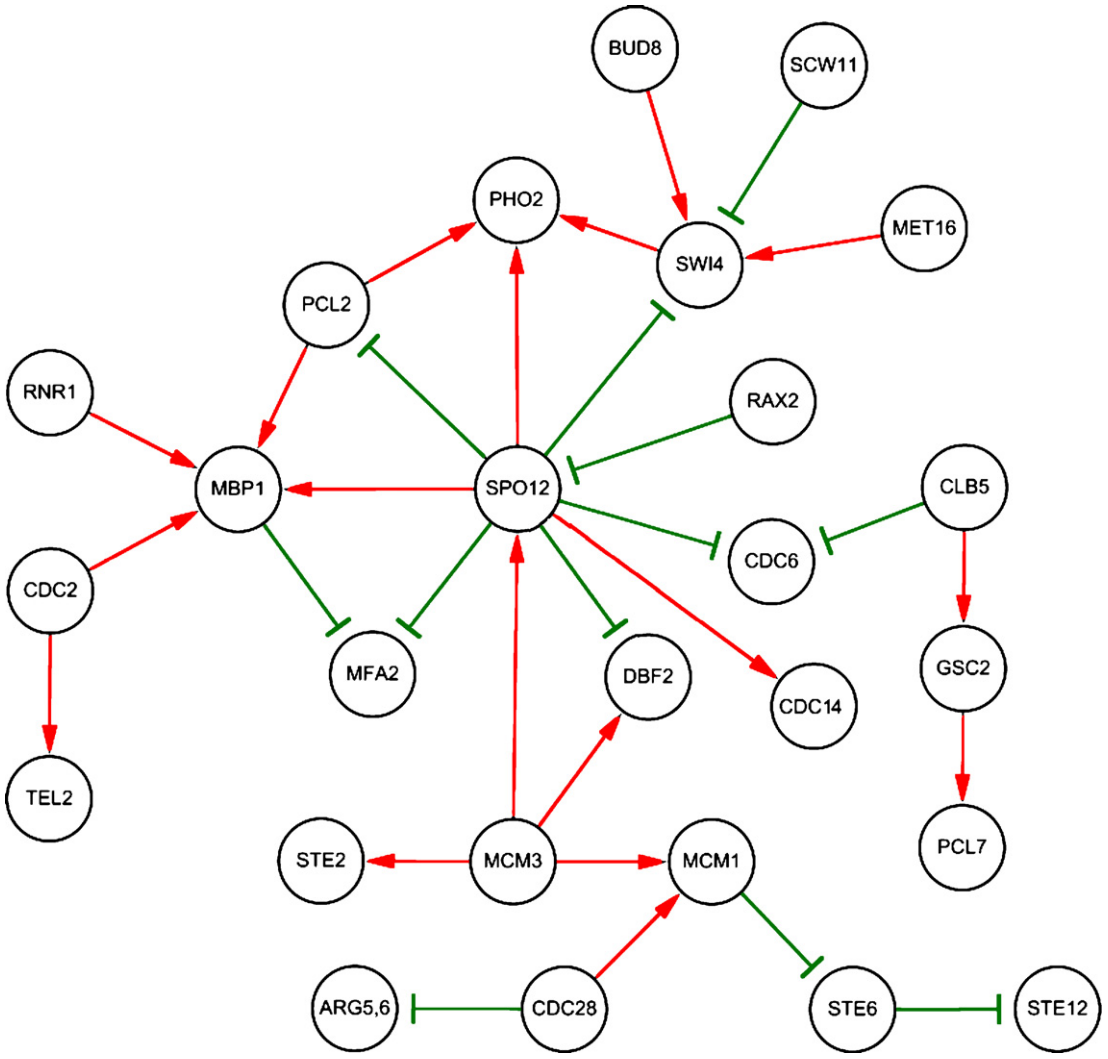

Fig. 7. Top-30 significant links predicted by the proposed model using *elu* data set. Several interactions are confirmed with biological evidences.

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

Table 1
Summary of the true positive rates (TPRs) and prediction accuracy (PreAcc) yielded by LPC, EB-GMMs, PARE, DBN, VBEM, MAPEM and the proposed model using *alpha* and *elu* data sets in Spellman et al. (1998).

Method	Data set			
	alpha		*elu*	
	Correctly predicted links	TPR[a] (PreAcc)[b]	Correctly predicted links	TPR[a] (PreAcc)[b]
LPC	139/291 (23/41)	47.8% (56.1%)	148/291 (28/47)	50.9% (59.6%)
EB-GGMs	151/291 (34/54)	51.9% (62.9%)	157/291 (43/64)	54.0% (67.2%)
PARE	197/291 (152/214)	67.7% (71.0%)	193/291 (171/228)	66.3% (75.0%)
DBN	12/291 (0/3)	4.1% (0.0%)	14/291 (1/5)	4.8% (20.0%)
VBEM	62/291 (46/102)	21.3% (45.1%)	65/291 (53/119)	22.3% (44.5%)
MAPEM	134/291 (32/58)	46.0% (55.1%)	141/291 (33/55)	48.5% (60.0%)
The proposed model	215/291 (147/174)	73.9% (84.5%)	209/291 (133/152)	71.8% (87.5%)

[a] TPR is the number of correctly predicted links divided by the number of true links.
[b] PreAcc is the number of correctly predicted links divided by the number of significant links.

has been implicated in the positive regulation of meiosis (Buonomo et al., 2003). In addition, SPO12 has been identified to be a positive regulator of the exit from mitosis (Chaves and Blobel, 2001). SPO12 and DBF2 are both reported to be related to the regulation of the mitotic cell cycle in buddy yeast (Toyn and Johnston, 1993, 1994). Extra copies of SPO12 can suppress DBF2, and the acquisition of an extra chromosome VIII, which carries the SPO12 locus, will also suppress DBF2 (Toyn and Johnston, 1993). SPO12 was reported to be involved in the activation and release of Cdc14 from the nucleolus during early anaphase (Asakawaa and Toh-e, 2002; Stegmeier et al., 2002, 2004). In the predicted network, MCM1 is predicted to be a suppressor of STE6, and STE6 is predicted to be a suppressor of STE12. Thus, we can interpret these interactions as an indication that MCM1 is an indirect positive regulator of STE12. In Bruhn and Sprague (1994), MCM1 is reported to mediate the interaction with both alpha-1 and STE12 for cell-type-specific gene regulation and DNA binding.

The predictions using the *alpha* and *elu* data sets were compared with independently published data as mentioned above. There was good agreement between the published and the predicted genetic interactions, which reinforce the possibility of applying the proposed model to predict pathways of the biological processes.

For the purpose of comparing the proposed model with existing methods, a lagged Pearson correlation (Schmitt et al., 2004), an empirical Bayesian-based Gaussian graph model (Schafer and Strimmer, 2005), a pattern recognition approach (Chuang et al., 2008), a dynamic Bayesian network (Perrin et al., 2003), a variational Bayes EM (Beal et al., 2005), and a state-space model (Rangel et al., 2004a) were applied to infer the interactions in the *a posteriori* knowledge data set. The model formulations of the lagged Pearson correlation (LPC), the empirical Bayesian approach-based Gaussian graph models (EB-GMM), the pattern recognition approach (PARE), the dynamic Bayesian network (DBN), the variational Bayes EM (VBEM), and the state-space model (MAPEM) are based on the log-ratio of the gene expression levels. We retrieved the normalized gene expression data of *alpha* and *elu* data sets from Spellman et al. (1998), and applied LPC, EB-GMMs, PARE, DBN, VBEM, and MAPEM to infer genetic interactions. For the *alpha* data set, the true positive rates yielded by LPC, EB-GMMs, PARE, DBN, VBEM and MAPEM are 48% (139/291), 52% (151/291), 68% (197/291), 5% (12/291), 21% (62/291), and 46% (134/291), respectively. For the *elu* data set, the true positive rates yielded by LPC, EB-GMMs, PARE, DBN, VBEM and MAPEM are 51% (148/291), 54% (157/291), 66% (193/291), 5% (14/291), 22% (65/291), and 49% (141/291), respectively. We have applied all approaches to the *alpha* and *elu* data set with their controls of false positive rates. In *alpha* data set, the prediction accuracies yielded by LPC, EB-GMM, PARE, DBN, VBEM and MAPEM, and the proposed method are 56% (23/41), 63% (34/54), 71% (152/214), 0% (0/3), 45% (46/102), 55% (32/58), and 85% (147/174), respectively. In addition, in *elu* data set, the prediction accuracies yielded by LPC, EB-GMM, PARE, DBN, VBEM and MAPEM, and the proposed method are 60% (28/47), 67% (43/64), 75% (171/228), 20% (1/5), 45% (53/119), 60% (33/55), and 88% (133/152), respectively. The results show that the performance of the proposed model is better than that of the other six algorithms. The evaluated results are summarized in Table 1 for comparison.

7. Predict pathways of pulmonary disease in *H. sapiens*

7.1. Experiment conditions and data preparation

Mycobacterium tuberculosis is a global infectious disease with approximately one-third of the world population having been exposed to it. The disease kills more than two million people annually. *M. tuberculosis* (MTB) is a major cause of human tuberculosis. Numerous studies have focused on the immune response to MTB. During the early stages of human tuberculosis, MTB induces an immune response which subsequently leads to the development of pulmonary granulomas (Flynn and Chan, 2001). Pulmonary granulomas consist of macrophages, T cells, B cells, and fibroblasts. Recent research reveals that fibroblasts are essential in secreting chemokine for modulating the inflammatory response to MTB infection and influencing the survival of MTB within the macrophages (O'Kane et al., 2007). Furthermore, fibroblasts involve in the regulation of granuloma formation during MTB infection (O'Kane et al., 2007, 2008). Despite the potentially crucial role of fibroblasts in MTB infection, knowledge of the detailed MTB-regulated mechanism in fibroblasts, especially its relation to MTB secreted protein, is still very limited. The CFP-10/ESAT-6 protein (CFES) is very famous secreted proteins from MTB and has been shown to elicit an immune response in the host organism (Abramo et al., 2006; Meher et al., 2007). However, the role and function of CFES in fibroblasts is not clear. To characterize the regulatory mechanisms of CFES, we have carried out whole-genome microarray experiments to identify potentially crucial pathways in the CFES-treated human diploid fibroblasts (WI-38) cells.

The following is the information on the experimental conditions underlying each microarray sample. The WI-38 cells were cultured in Modified Eagle Medium (Gibco) containing 10% fetal clf serum (FCS), 2 mM L-glutamine, 1 mM sodium pyruvate, 100 μg/mL streptomycin, 0.025 μg/mL amphotericin B, and 100 U/mL penicillin at 37 °C with 5% CO_2. The bacterial vector pET29b-CFES was created by cloning CFP-10 and ESAT-6 from the H37Rv strain of MTB and was transformed into *Escherichia coli* BL21 (DE3). IPTG was used to express abundant amounts of recombinant CFP-10/ESAT-6 protein (CFES). Next, we purified CFES protein using affinity chromatography with nickel ion characteristic and dialysis. The Bio-Rad DC Protein Assay Kit was used to measure CFES concentrations. The purity of the CFES protein was controlled at a level above 95% as assessed by the densitometry of SDS-PAGE gels. In addition, MALTI-

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

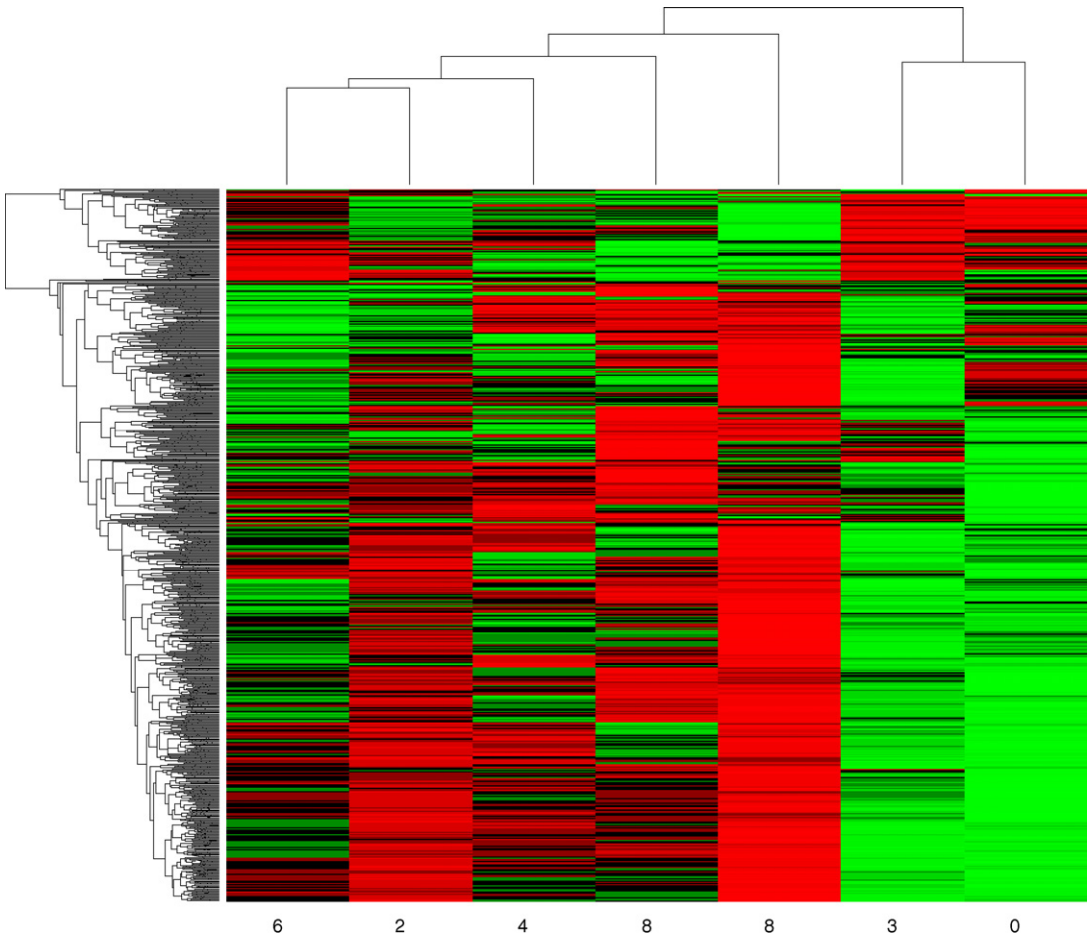

Fig. 9. Visualization of the CFES-treated WI-38 cell microarray data. The microarray data consists of 8 time points (7 in visualization plot) with 3 replicates.

TOF mass spectroscopy was used to confirm that the CFES protein was composed of CFP-10 and ESAT-6. WI-38 cells were arrested at the G0/G1 phase after treatment with serum free medium for 24 h, treated with CFES (12.5 μM), and incubated for 0, 3, 8, 16, 24, 32, 40, and 48 h. Total RNA was extracted with Trizol reagent, according to the manufacturer's protocol (Invitrogen) and with RNeasy Mini Kit (Qiagen). Purified RNA was quantified at OD260 with an ND-1000 spectrophotometer (Nanodrop), and determine the amounts of the mRNA was determined with Agilent Bioanalyzer 2100 (Agilent). Total RNA was amplified with a Fluorescent Linear Amplification Kit (Agilent). Next, cRNA was hybridized onto human whole-genome oligo microarrays (Agilent), according to the manufacturer's protocols. A total of 43,931 probe sets on the arrays were analyzed. Feature Extraction Software (Agilent) was used to analyze the scanned images and to estimate the differential gene expression by calculating the statistical confidences. We selected 41,675 probe sets by rank-consistency-filtering using the LOWESS method. These probe sets were aligned to 7526 genes and exons with 8 time points and 3 replicates based on genomic coordinates provided for each probe set from the Agilent microarray specification. The visualization of the CFES-treated WI-38 cell microarray data can be obtained by subtracting the first time point of each gene from the following ones, as shown in Fig. 9. The microarray data is freely available at http://bem.bime.ntu.edu.tw/clchuang/CNNM_MTB_Data_Raw.zip.

7.2. Prediction of interactions in human disease pathways

By surveying previous published studies, we suggest the readers to set S to the average production rate of all genes, 10 (mRNA/h), and λ to the average degradation rate of all genes, 120 (Chen et al., 2004; Yang et al., 2003). Then, we applied the proposed model to fit the microarray data, and 181,797 out of 56,640,676 possible links were predicted to be significant interactions. There are fourteen human cancer pathways in the KEGG DISEASE database (Kanehisa et al., 2006). We checked the predictions against the disease pathways annotated in the KEGG database. We found that the prediction results yielded by the proposed model were highly consistent with the pathways of small cell lung cancer, non-small cell lung cancer, acute myeloid leukemia, and cell adhesion molecular disorder. The matching results of the predicted and annotated interactions in each disease pathway are summarized in Table 2.

Table 2
Summary of the true positive rates of the proposed model by checking the predicted links against the known interactions in human diseases.

Pathway	Name	Predicted links	Correct predictions	Prediction accuracy
hsa04514	Cell adhesion molecules (CAMs)	7	5	71.4%
hsa05221	Acute myeloid leukemia	51	44	86.3%
hsa05222	Small cell lung cancer	17	14	82.4%
hsa05223	Non-small cell lung cancer	20	19	95.0%

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

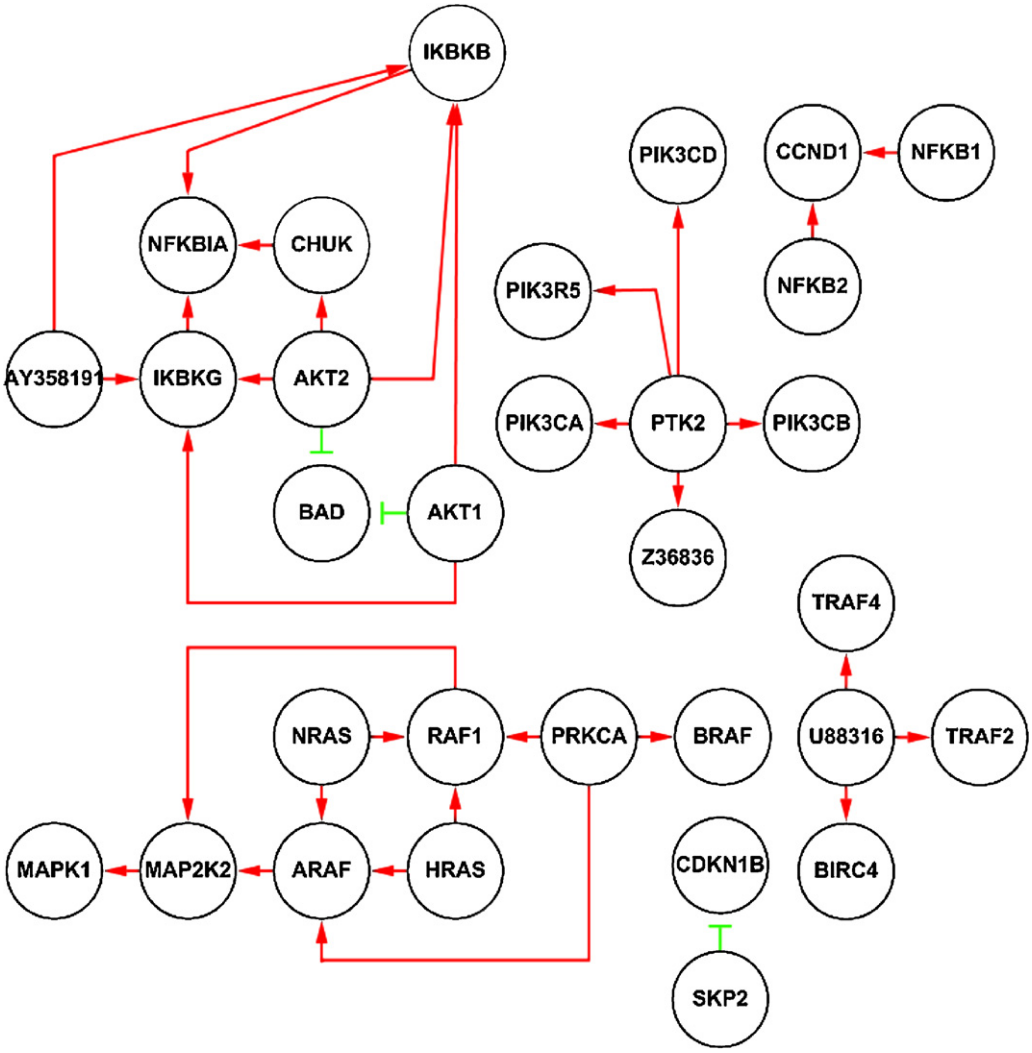

Fig. 10. Regulatory network constructed by the true positive links predicted by the proposed model in the pathways of small cell lung cancer and non-small cell lung cancer.

Several long-term clinical studies have shown that a large percentage of MTB patients show evidence of permanent obstruction or restrictive impairment of lung function (Willcox and Ferguson, 1989; Krishna et al., 1977; Vargha, 1983). For MTB patients who had damaged lung function, the risk of triggering the development of lung cancer is more than twice that of people with healthy lung (Sin et al., 2004; Man et al., 2003). By reviewing the medical records of patients with coexisting MTB and lung cancer, it has been reported that the incidence of lung cancer is significantly higher in MTB patients (Chen et al., 1996). Recent reports have shown that CFES protein released from MTB can have directly influence susceptibility to lung immune system (Abramo et al., 2006; Meher et al., 2007). Furthermore, fibroblasts play an essential role in modulating inflammatory response to MTB infection (O'Kane et al., 2007). It has been reported that inflammatory response-altering interleukin 1 (IL1) gene variations may raise the risk to develop lung cancer (Engel et al., 2007). Thus, we turned our focus to the disease pathways of small cell lung cancer and non-small cell lung cancer. The prediction result given by the proposed model contains 37 positive prediction interactions related to the pathways of lung cancer. By checking the known interaction annotated in KEGG database, 33 (89.1%) among them are true positives. These true positive links are shown in Fig. 10 as a gene regulatory network. The functions of 31 genes that involve in the predicted network are summarized in Table 3. A high percentage of genes were identified to involve in essential functions of MAPK, cell death, and apoptosis. There are 9 genes found to be oncogenes that may play a role in the malignant transformation to acquire the properties of cancer. Furthermore, 18 genes were found to involve in more than 5 cancer pathways. Thus, the prediction result yielded by the proposed model suggests that CFES protein released from MTB may be a potential factor to initiate the pathogenesis of lung cancer. With previous clinical studies mentioned in above state the associations between MTB and lung cancer, it would be worth for us to conduct a more comprehensive investigation into the relation between CFES protein and lung cancer at the molecular level.

As summarized in Table 2, it can be seen that the result given by the proposed model is highly consistent with interactions in the pathway of acute myeloid leukemia (AML). There are a total of 51 interactions that have been identified as positive links. By checking the result against the KEGG database, 44 (86.3%) among them were true positives in the pathway of AML. We formed these true positive links into a regulatory network as shown in Fig. 11. The predicted network encompasses with 36 genes summarized in Table 4. We can see that a large number of genes involve in functions of cell death, apoptosis, MAPK, and immune system response. In addition, 31 genes are identified to be oncogenes, and 15 genes involve in more than 5 cancer pathways. The predictions suggest that CFES protein released from MTB may be a potential factor in the genesis of AML. Such inference is valid only if the MTB has invaded bone

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

Table 3
Summary of functions of genes predicted to involve in the pathways of lung cancer.

Gene Symbol	GO database			KEGG or BIOCARTA pathway					Gene Symbol	GO database			KEGG or BIOCARTA pathway				
	①	②	③	④	⑤	⑥	⑦	⑧		①	②	③	④	⑤	⑥	⑦	⑧
AKT1								9	NFKB2								5
AKT2								9	NFKBIA								3
ARAF								9	NRAS								--
AY358191								--	PIK3CA								9
BAD								7	PIK3CB								9
BIRC4								1	PIK3CD								9
BRAF								10	PIK3R5								9
CCND1								10	PRKCA								1
CDKN1B								3	PTK2								1
CHUK								5	RAF1								9
HRAS								--	SKP2								1
IKBKB								5	TRAF2								1
IKBKG								5	TRAF4								1
MAP2K2								8	U88316								--
MAPK1								10	Z36836								--
NFKB1								5									

① MAPK, ② Cell death, ③ Immune system, ④ Apoptosis, ⑤ Immune system, ⑥ MAPK signaling, ⑦ Oncogene, ⑧ Number of cancer pathways involved.

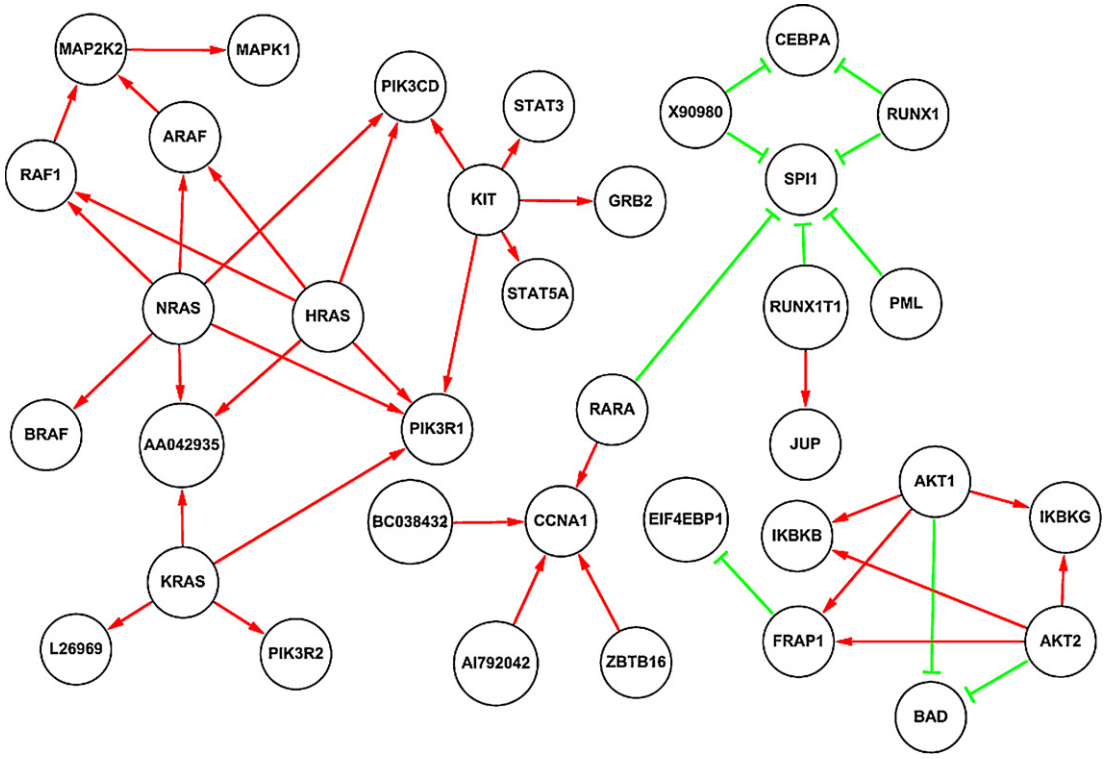

Fig. 11. Regulatory network constructed by the true positive links predicted by the proposed model in the pathway of acute myeloid leukemia.

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

Table 4
Summary of functions of genes predicted to involve in the pathways of acute myeloid leukemia.

Gene Symbol	GO database			KEGG or BIOCARTA pathway					Gene Symbol	GO database			KEGG or BIOCARTA pathway				
	①	②	③	④	⑤	⑥	⑦	⑧		①	②	③	④	⑤	⑥	⑦	⑧
AA042935								--	KRAS								10
AI792042								--	L26969								--
AKT1								9	MAP2K2								8
AKT2								9	MAPK1								10
ARAF								9	NRAS								--
BAD								7	PIK3CD								9
BC038432								--	PIK3R1								9
BRAF								10	PIK3R2								9
CCNA1								1	PML								1
CEBPA								1	RAF1								9
EIF4EBP1								1	RARA								1
FRAP1								2	RUNX1								2
GRB2								7	RUNX1T1								1
HRAS								--	SPI1								1
IKBKB								5	STAT3								2
IKBKG								5	STAT5A								2
JUP								1	X90980								--
KIT								1	ZBTB16								1

① MAPK, ② Cell death, ③ Immune system, ④ Apoptosis, ⑤ Immune system, ⑥ MAPK signaling, ⑦ Oncogene, ⑧ Number of cancer pathways involved.

marrow, since only bone marrow necrosis may induce AML. In order to verify the inference of the proposed model, clinical evidence that addresses the association between MTB and AML is required, and will now be discussed. In earlier clinical studies (Chandra et al., 1986; Lowther, 1959; Twomey and Leavell, 1965), a variety of haematological alternations ranging from various cytopenias to leukemoid reactions and frank leukemia in association with MTB were reported. In Sen et al. (1991), the relationship between developed MTB and myelodysplastic syndrome has been confirmed, where the myelodysplastic syndrome is known to represent a preleukemic state in the stem cells and evolves into the pathogenesis of AML (Francis and Hoffbrand, 1985). In Olaniyi and Aken'Ova (2003), it was reported that MTB patients can develop leukocytosis. In addition, MTB can present with bone marrow necrosis to induce AML (Kumar et al., 2005; Paydas et al., 2002; Al-Anazi et al., 2007), and then cause fatality. These clinical studies further support the inference given by the proposed model in the relation between CFES protein released from MTB and AML.

Finally, we looked into the association between cell adhesion molecules and MTB. There are a total of 11 links were predicted to be positive interactions, and 8 (72.7%) out of them were true positives annotated in the KEGG database. The predictions yielded by the proposed model can be interpreted as the fact that CFES protein released from MTB causing inflammatory response in lung, and these reactions are a part of the pathway regulated by cell adhesion molecules. Many studies have indicated the importance of cell adhesion molecules in the pathogenesis of various inflammatory lung diseases (Mukae et al., 2003), which includes inflammatory and immune responses caused by MTB infection. In Lai et al. (1993), it has been reported that cell adhesion molecules may be involved in the pathogenesis of MTB, and that some adhesion molecules could be used as markers to diagnose the disease activity. Thus, the clinical evidence mentioned above further supports the inferences yielded by the proposed model that cell adhesion molecules play a role in inflammatory and immune responses against CFES protein released from infectious MTB.

The consistencies between prediction and known human disease pathways reinforce the possibility of using the proposed model to predict pathways of complex human diseases.

8. Conclusions

The proposed nonlinear model first calculates the activation levels of genes using a robust correlation estimator and the gene activity extraction scheme. Next, it uses a recurrent network to model the mechanism of gene regulation. After the nonlinear fitting process, the parameters of the tuned network can be used to predict the functional linkage between genes. The proposed model was applied to simulated data set and microarray data sets in *S. cerevisiae* and *H. sapiens*. For *S. cerevisiae*, the proposed model was applied to *alpha* and *elu* date sets in Spellman et al. (1998). The true positive rates of the proposed model were 73.9–71.8% in the *alpha* and *elu* data set, respectively. The performance of the proposed model was evaluated and compared with the lagged Pearson correlation, Bayesian-based Gaussian graph model, the pattern recognition, the dynamic Bayesian network, the variational Bayes EM, and the state-space model using the same microarray data sets. The results showed that the proposed model outperformed the other six algorithms. Furthermore, many predicted interactions were shown to agree with known interactions. These consistencies indicate that

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

the proposed model has the potential to predict genetic interactions.

Furthermore, the proposed model was applied to time-course gene expression data of CFES-treated human diploid fibroblasts (WI-38) cells. By matching the predicted interactions against known interactions annotated in the KEGG database, the proposed model yielded high true positive rates in four human disease pathways. These pathways are considered to be highly related to the CFES-treated WI-38 cells, which are small cell lung cancer, non-small cell lung cancer, acute myeloid leukemia, and cell adhesion molecular disorder. The true positive rates of the proposed model in the prediction of the interactions in human diseases ranged from 71.4% to 95%. The prediction results imply that CFES-treated WI-38 cells might have activated these human diseases due to the treatments used in the experiment. By surveying previous studies and reports, many clinical researchers confirmed the inference made by the proposed model for many clinical medical records. Thus, the proposed model may be a feasible tool to predict interactions in complex human diseases. The predicted interactions in the pathways need to be further investigated in order to unravel the underlying regulatory mechanisms of complex diseases, which we leave for future research.

Acknowledgements

Prof. Jiang is grateful for the financial support from the President of National Taiwan University under contract no. 97R0533-2. The authors would like to thank the National Science Council of Taiwan, ROC for financially supporting this research under contract no. NSC 97-2218-E-002-006 (for Prof. Jiang), NSC95-2221-E-002-029-MY2 (for Prof. Chen) and NSC95-2320-B-182-032-MY2 (for Prof. Chan). We are grateful to Guest Editor and three anonymous referees for their invaluable suggestions to improve the paper.

References

Abramo, C., Meijgaarden, K.E., Garcia, D., Franken, K.L., Klein, M.R., Kolk, A.J., Oliveira, S.C., Ottenhoff, T.H., Teixeira, H.C., 2006. Monokine induced by interferon gamma and IFN-γ response to a fusion protein of *Mycobacterium tuberculosis* ESAT-6 and CFP-10 in Brazilian tuberculosis patients. Microbes Infect. 8, 45–51.

Al-Anazi, K.A., Al-Jasser, A.M., Evans, D.A.P., 2007. Infections caused by mycobacterium tuberculosis in patients with hematological disorders and in recipients of hematopoietic stem cell transplant, a twelve year retrospective study. Ann. Clin. Microbiol. Antimicrob. 6, 16–22.

Alberghina, L., Smeraldi, C., Ranzi, B.M., Porro, D., 1998. Control by nutrients of growth and cell cycle progression in budding yeast. J. Bacteriol. 180, 3864–3872.

Asakawaa, K., Toh-e, A., 2002. A defect of Kap104 alleviates the requirement of mitotic exit network gene functions in *Saccharomyces cerevisiae*. Genetics 162, 1545–1556.

Beal, M.J., Falciani, F., Ghahramani, Z., Rangel, C., Wild, D.L., 2005. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioinformatics 21, 349–356.

Biedl, T., Brejov, B., Demaine, E.D., Hamel, A.M., Vinar, T., 2001. Optimal arrangement of leaves in the tree representing hierarchical clustering of gene expression data. Technical Report 2001–2014. Dept. Comput. Sci., Univ. Waterloo.

Bruhn, L., Sprague Jr., G.F., 1994. MCM1 point mutants deficient in expression of alpha-specific genes: residues important for interaction with alpha 1. Mol. Cell. Biol. 14, 2534–2544.

Buonomo, S.B., Rabitsch, K.P., Fuchs, J., Gruber, S., Sullivan, M., Uhlmann, F., Petronczki, M., Tóth, A., Nasmyth, K., 2003. Division of the nucleolus and its release of CDC14 during anaphase of meiosis I depends on separase, SPO12, and SLK19. Dev. Cell 4, 727–739.

Chandra, J., Marwaha, R.K., Marwaha, N., Kumar, A., Mohanty, O., 1986. Pancytopenia and leukemoid reactions in tuberculosis. Indian J. Paediatr. 53, 665–667.

Chaves, S.R., Blobel, G., 2001. Nuclear import of Spo12p, a protein essential for meiosis. Biol. Chem. 276, 17712–17717.

Chen, H., Pan, Y.X., Dudenhausen, E.E., Kilberg, M.S., 2004. Amino acid deprivation induces the transcription rate of the human asparagine synthetase gene through a timed program of expression and promoter binding of nutrient-responsive basic region/leucine zipper transcription factors as well as localized histone acetylation. J. Biol. Chem. 279, 50829–50839.

Chen, L., Aihara, K., 2002. Stability of genetic regulatory networks with time delay. IEEE Trans. Circuits Syst.-I 49, 602–608.

Chen, Y.M., Chao, J.Y., Tsai, C.M., Lee, P.Y., Perng, R.P., 1996. Shortened survival of lung cancer patients initially presenting with pulmonary tuberculosis. Jpn. J. Clin. Oncol. 26, 322–327.

Chuang, C.L., Chen, C.M., Shieh, G.S., Jiang, J.A., 2007. GeneCFE-ANFIS: a neuro-fuzzy inference system to infer gene–gene interactions based on recognition of microarray gene expression patterns. Biomed. Eng.: Appl. Basis Commun. 19, 71–78.

Chuang, C.L., Jen, C.H., Chen, C.M., Shieh, G.S., 2008. A pattern recognition approach to infer time-lagged genetic interactions. Bioinformatics 24, 1183–1190.

Daignan-Fornier, B., Fink, G.R., 1992. Coregulation of purine and histidine biosynthesis by the transcriptional activators BAS1 and BAS2. Proc. Natl. Acad. Sci. U.S.A. 89, 6746–6750.

Dirick, L., Bohm, T., Nasmyth, K., 1995. Roles and regulation of Cln-Cdc28 kinases at the start of the cell cycle of *Saccharomyces cerevisiae*. EMBO J. 14, 4803–4813.

Dobra, A., Hans, C., Jones, B., Nevins, J.R., West, M., 2004. Sparse graphical models for exploring gene expression data. J. Multivariate Anal. 90, 196–212.

Dojer, N., Gambin, A., Mizera, A., et al., 2006. Applying dynamic Bayesian networks to perturbed gene expression data. BMC Bioinform. 7, 249–259.

Dudoit, S., Yang, Y.H., 2002. Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In: Parmigiani, G., Garrett, E.S., Irizarry, R.A., ZegerF S.L. (Eds.), The Analysis of Gene Expression Data: Methods and Software. Springer, New York.

Eisen, M.B., Spellman, P.T., et al., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U.S.A. 95, 14863–14869.

Engel, E.A., Wu, X., Gu, J., Dong, Q., Liu, J., Spitz, M.R., 2007. Systematic evaluation of genetic variants in the inflammation pathway and risk of lung cancer. Cancer Res. 67, 6520–6527.

Flynn, J.L., Chan, J., 2001. Immunology of tuberculosis. Annu. Rev. Immunol. 19, 93–129.

Fournier, T., et al., 2009. Stochastic models and numerical algorithms for a class of regulatory gene networks. Bull. Math. Biol., arXiv:0810.0239v1.

Francis, G.E., Hoffbrand, A.V., 1985. The myelodysplastic syndromes and preleukaemia. In: Hoffbrand, A.V. (Ed.), Recent Advances in Haematology No. 4. Churchill Livingstone, New York.

Friedman, N., et al., 2000. Using Bayesian networks to analyze expression data. J. Comp. Biol. 7, 601–620.

Friedman, N., 2004. Inferring cellular networks using probabilistic graphical models. Science 303, 799–805.

Futcher, B., 1996. Cyclins and the wiring of the yeast cell cycle. Yeast 12, 1635–1646.

Getz, G., Levine, E., et al., 2000. Coupled two-way clustering analysis of gene microarray data. Proc. Natl. Acad. Sci. U.S.A. 97, 12079–12084.

Goutsias, J., 2005. Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems. J. Chem. Phys. 122, 184102.

Guttmann-Raviv, N., et al., 2002. Ime2, a meiosis-specific kinase in yeast, is required for destabilization of its transcriptional activator, Ime1. Mol. Cell. Biol. 22, 2047–2056.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al., 2004. Transcriptional regulatory code of a eukaryotic genome. Nature 431, 99–104.

Haseltine, E.L., Rawlings, J.B., 2002. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. J. Chem. Phys. 117, 6959–6969.

Huber, P.J., 1981. Robust Statistics. John Wiley & Sons.

Hughes, T.R., et al., 2000. Functional discovery via a compendium of expression profiles. Cell 102, 109–126.

Husmeier, D., 2003. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 19, 2271–2282.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., et al., 2002. Revealing modular organization in the yeast transcriptional network. Nat. Genet. 31, 370–377.

Imoto, S., Goto, T., Miyano, S., 2002a. Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. Pac. Symp. Biocomput. 7, 175–186.

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., Miyano, S., 2002b. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. J. Bioinform. Comput. Biol. 1, 231–252.

de Jong, H., 2002. Modeling and simulation of genetic regulatory systems: a literature review. J. Comp. Biol. 9, 69–105.

Kumar, P.V., Monabati, A., Kadivar, R., Soleimanpour, H., 2005. Peripheral blood and marrow findings in disseminated bacilli Calmette-Guerin infection. J. Pediatr. Hematol. Oncol. 27, 97–99.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354–357.

Kim, S., Imoto, S., Miyano, S., 2004. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. Biosystems 75, 57–65.

Kimura, S., Idel, K., Kashihara, A., Kano, M., Hatakeyama, M., Masui, R., Nakagawa, N., Yokoyama, S., Kuramitsu, S., Konagaya, A., 2005. Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. Bioinformatics 21, 1154–1163.

Krause, E.F., 1986. Taxicab Geometry: An Adventure in Non-Euclidean Geometry. Dover Publications.

Krishna, K., Bond, S., Artvinli, M., et al., 1977. Pulmonary function in treated tuberculosis; a long term follow-up. Am. Rev. Respir. Dis. 115, 402.

Lähdesmäki, H., Hautaniemi, S., Shmulevich, I., Yli-Harja, O., 2006. Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. Signal Process. 86, 814–834.

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013

Lähdesmäki, H., Shmulevich, I., 2008. Learning the structure of dynamic Bayesian networks from time series and steady state measurements. Mach. Learn. 71, 185–217.
Lai, C.K., Wong, K.C., Chan, C.H., Ho, S.S., Chung, S.Y., Haskard, D.O., Lai, K.N., 1993. Circulating adhesion molecules in tuberculosis. Clin. Exp. Immunol. 94, 522–526.
Lee, T.I., et al., 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298, 799–804.
Levine, K., Huang, K., Cross, F.R., 1996. *Saccharomyces cerevisiae* G1 cyclins differ in their intrinsic functional properties. Mol. Cell. Biol. 16, 5843–5853.
Liang, S., Fuhrman, S., Somogyi, R., 1998. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac. Symp. Biocomput. 3, 18–29.
Liebermeister, W., 2002. Linear modes of gene expression determined by independent component analysis. Bioinformatics 18, 51–60.
Lipan, O., Wong, W.H., 2005. The use of oscillatory signals in the study of genetic networks. Proc. Natl. Acad. Sci. U.S.A. 102, 7063–7068.
Lodi, T., Guiard, B., 1991. Complex transcriptional regulation of the *Saccharomyces cerevisiae* CYB2 gene encoding cytochrome b2: CYP1(HAP1) activator binds to the CYB2 upstream activation site UAS1-B2. Mol. Cell. Biol. 11, 3762–3772.
Lowther, C.P., 1959. Leukaemia and tuberculosis. Ann. Intern. Med. 5, 52–56.
Man, S.F.P., McAlister, F.A., Anthonisen, N.R., Sin, D.D., 2003. Contemporary management of chronic obstructive pulmonary disease (COPD). Clinical Applications. JAMA 290, 2313–2316.
Matys, V., et al., 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 31, 374–378.
Meher, A.K., Lella, R.K., Sharma, C., Arora, A., 2007. Analysis of complex formation and immune response of CFP-10 and ESAT-6 mutants. Vaccine 25, 6098–6106.
Michaelis, L., Menten, M.L., 1913. Die Kinetik der Invertinwirkung. Biochem. Z. 49, 334–336.
Mukae, H., Ashitani, J., Tokojima, M., Ihi, T., Kohno, S., Matsukura, S., 2003. Elevated levels of circulating adhesion molecules in patients with active pulmonary tuberculosis. Respirology 8, 326–331.
Natarajan, K., et al., 2001. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. Mol. Cell. Biol. 21, 4347–4368.
Nicholas, J.G., et al., 2006. A bottom-up approach to gene regulation. Nature 439, 856–860.
O'Kane, C.M., Boyle, J.J., Horncastle, D.E., Elkington, P.T., Friedland, J.S., 2007. Monocyte-dependent fibroblast CXCL8 secretion occurs in tuberculosis and limits survival of mycobacteria within macrophages. J. Immunol. 178, 3767–3776.
O'Kane, C.M., Elkington, P.T., Friedland, J.S., 2008. Monocyte-dependent oncostatin M and TNF-alpha synergize to stimulate unopposed matrix metalloproteinase-1/3 secretion from human lung fibroblasts in tuberculosis. Eur. J. Immunol. 28, 1321–1330.
Olaniyi, J.A., Aken'Ova, Y.A., 2003. Hematological profile of patients with pulmonary tuberculosis in Ibadan, Nigeria. Afr. J. Med. Med. Sci. 32, 239–242.
Paydas, S., Ergin, M., Baslamisli, F., Yavuz, S., Zorludemir, S., Sahin, B., Bolat, F.A., 2002. Bone marrow necrosis: clinicopathologic analysis of 20 cases and review of the literature. Am. J. Hematol. 70, 300–305.
Pedraza, J., van Oudenaarden, A., 2005. Noise propagation in gene networks. Science 307, 1965–1969.
Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., dAlché-Buc, F., 2003. Gene networks inference using dynamic Bayesian networks. Bioinformatics 19, ii138–ii148.
Pilpel, Y., Sudarsanam, P., Church, G.M., 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat. Genet., 153–159.
Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D.L., Falciani, F., 2004a. Modelling T-cell activation using gene expression profiling and state space models. Bioinformatics 20, 1361–1372.
Rangel, C., Angus, J., Ghahramani, Z., Wild, D.L., 2004b. Modeling genetic regulatory networks using gene expression profiling and state space models. In: Husmeier, D., Roberts, S., Dybowski, R. (Eds.), Probabilistic Modelling in Bioinformatics and Medical Informatics. Springer-Verlag, Berlin, pp. 269–293.
Rao, C.V., Arkin, A.P., 2003. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. J. Chem. Phys. 118, 4999–5010.
Raychaudhuri, S., Stuart, J.M., Altman, R.B., 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac. Symp. Biocomput. 2000, 455–466.
Rousseeuw, P.J., 1984. Least median of squares regression. J. Am. Stat. Assoc. 79, 871–880.
Rousseeuw, P.J., 1985. In: Grossman, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), Multivariate Estimation With High Breakdown Point. Mathematical Statistics and Applications. B, pp. 283–297.
Sabatti, C., James, G.M., 2005. Bayesian sparse hidden components analysis for transcription regulation networks. Bioinformatics 22, 736–746.
Schafer, J., Strimmer, K., 2005. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 21, 754–764.
Schmitt, W.A., et al., 2004. Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. Genome Res. 14, 1654–1663.
Segal, E., Shapira, M., Regev, A., Botstein, D., Koller, D., Friedman, N., 2003. Module networks: discovering regulatory modules and their condition specific regulators from gene expression data. Nat. Genet. 34, 166–176.
Sen, R., Yadav, S.S., Sen, J., Singh, S., Singh, H., 1991. Myelodysplastic syndrome and hypercoagulable state with tuberculosis—a case report. Indian J. Tuberc. 38, 29–32.
Simon, I., Fiser, A., Tusnady, G.E., 2001. Predicting protein conformation by statistical methods. Biochim. Biophys. Acta 1549, 123–136.
Sin, D.D., Golmohammadi, K., Jacobs, P., 2004. Cost-effectiveness of inhaled corticosteroids for chronic obstructive pulmonary disease according to disease severity. Am. J. Med. 116, 325–331.
Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell 9, 3273–3297.
Spiegel, M.R., 1992. Theory and Problems of Probability and Statistics. McGraw-Hill, New York, pp. 114–115.
Stegmeier, F., Huang, J., Rahal, R., Zmolik, J., Moazed, D., Amon, A., 2004. The replication fork block protein Fob1 functions as a negative regulator of the FEAR network. Curr. Biol. 14, 467–480.
Stegmeier, F., Visintin, R., Amon, A., 2002. Separase, polo kinase, the kinetochore protein Slk19, and Spo12 function in a network that controls Cdc14 localization during early anaphase. Cell 108, 207–220.
Stuart, D., Wittemberg, C., 1995. CLN3, not positive feedback, determines the timing of CLN2 transcription in cycling cells. Genes Dev. 9, 2780–2794.
Tamada, Y., Kim, S.Y., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., Miyano, S., 2003. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. Bioinformatics 19, ii227–ii236.
Teixeira, M.C., et al., 2006. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. Nucleic Acids Res. 34, D446–D451.
Toyn, J.H., Johnston, L.H., 1993. Spo12 is a limiting factor that interacts with the cell cycle protein kinases Dbf2 and Dbf20, which are involved in mitotic chromatid disjunction. Genetics 135, 963–971.
Toyn, J.H., Johnston, L.H., 1994. The dbf2 and dbf20 protein kinases of budding yeast are activated after the metaphase to anaphase cell cycle transition. EMBO J. 13, 1103–1113.
Törönen, P., Kolehmainen, M., Wong, G., Castrén, E., 1999. Analysis of gene expression data using self-organizing maps. FEBS Lett. 451, 142–146.
Tyres, M., Tokiwa, G., Futcher, B., 1993. Comparison of the *Saccharomyces cerevisiae* G1 cyclins: Cln3 may be an upstream activator of Cln1 Cln2 and other cyclins. EMBO J. 12, 1955–1968.
Twomey, T.J., Leavell, B.S., 1965. Leukemoid reactions to tuberculosis. Arch. Intern. Med. 116, 21–28.
van Someren, E.P., Wessels, L.F., Backer, E., Reinders, M.J., 2002. Genetic network modeling. Pharniacogenomics 3, 507–525.
Vargha, G., 1983. Fifteen year follow-up of lung function on obstructive and non-obstructive pulmonary tuberculosis. Acta. Med. Hung. 40, 271–276.
Vershon, A.K., Pierce, M., 2000. Transcriptional regulation of meiosis in yeast. Curr. Opin. Cell Biol. 12, 334–339.
Vu, T.T., Vohradsky, J., 2007. Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of *Saccharomyces cerevisiae*. Nucleic Acids Res. 35, 279–287.
Whittaker, J., 1990. Graphical Models in Applied Multivariate Statistics. Wiley, NY.
Willcox, P.A., Ferguson, A.D., 1989. Chronic obstructive airways disease following treated pulmonary tuberculosis. Respir. Med. 83, 195–198.
Wong, F., Carter, C.K., Kohn, R., 2003. Efficient estimation of covariance selection models. Biometrika 90, 809–830.
Wu, F., Zhang, W., Kusalik, A., 2004. Modeling gene expression from microarray expression data with state-space equations. Pac. Symp. Biocomput. 9, 581–592.
Yamaguchi, R., Higuchi, T., 2006. State-space approach with the maximum likelihood principle to identify the system generating time-course gene expression data of yeast. Int. J. Data Mining Bioinform. 1, 77–87.
Yamaguchi, R., Yoshida, R., Imoto, S., Higuchi, R., Miyano, S., 2007. Finding module-based gene networks with state-space models. IEEE Signal Process. Mag. 24, 37–46.
Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M., Darnell Jr., J.E., 2003. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Res. 13, 1863–1872.
Zhong, W., Sternberg, P.W., 2006. Genome-wide prediction of C. elegans genetic interactions. Science 311, 1481–1484.
Zweigenbaum, P., Demner-Fushman, D., Yu, H., Cohen, K.B., 2007. Frontiers of biomedical text mining: current progress. Brief Bioinform. 8, 358–375.

Please cite this article in press as: Chuang, C.-L., et al., A robust correlation estimator and nonlinear recurrent model to infer genetic interactions in *Saccharomyces cerevisiae* and pathways of pulmonary disease in *Homo sapiens*. BioSystems (2009), doi:10.1016/j.biosystems.2009.05.013