A New Multiple Regression Approach for the Construction of Genetic Regulatory Networks*

Shu-Qin Zhang

School of Mathematical Sciences,

Fudan University, Shanghai, China

E-mail: zhangs@fudan.edu.cn

Wai-Ki Ching† Nam-Kiu Tsing Ho-Yin Leung

Advanced Modeling and Applied Computing Laboratory,

Department of Mathematics,The University of Hong Kong,

Pokfulam Road, Hong Kong

E-mail: {wching,nktsing,obliging}@hkusua.hku.hk

† Corresponding author.

Diane D. Guo

Department of Biology,

The Chinese University of Hong Kong,

Shatin, N.T., Hong Kong

E-mail:djguo@cuhk.edu.hk

Summary

Objective

The construction of genetic regulatory networks from time series gene expression data is an important research topic in bioinformatics as large amounts of quantitative gene expression data can be routinely generated nowadays. One of the main difficulties in building such genetic networks is that the data set has huge number of genes but small number of time points. In this paper, we propose a novel linear regression model for uncovering the relations among the genes.

Methods

The model is based on the multiple regression. It takes into account of the fact that the real biological networks have the scale-free property. Based on this property and the statistical tests, a filter can be constructed to filter some redundant interactions among the genes. By minimizing the distance between the observed data and the predicted data, the model can be finally constructed.

Results

*Preliminary version has presented in the International Conference on BioMedical Engineering and Informatics, 2008.

Numerical examples based on the yeast gene expression data are given to demonstrate our method. The proposed model can fit the data quite well. Some properties of the genes and the network are obtained. Among them, some are consistent with the experimental results.

Conclusions

In this paper, we proposed a new multiple regression approach to model the gene-gene interactions by taking into account the scale-free property. Numerical results show the effectiveness of our method. The comparison with some other models which didnot consider the scale-free property will be as one of our future research topics.

Keywords: gene regulatory network, multiple regression, power-law, statistical tests

1 Introduction

The development of microarray technologies has dramatically accelerated the exploration of living organisms at the genomic level. Huge amounts of quantitative gene expression data can be routinely generated nowadays. From such data, the regulatory interactions among different genes can be inferred with suitable methods. The difficulty in the inference process lies in the data dimensions: the huge number of genes and the small number of time points in the time series data and also the small number of different experiments for the steady-state data. Thus mathematical modeling and computational algorithms for inferring the relations are indispensable. In fact, many mathematical models and algorithms have been proposed for the inference

of gene networks [4, 18]. For the discrete gene expression data (expressed: 1, unexpressed: 0), many models including Boolean Network (BN) model, Probabilistic Boolean Network (PBN) model, multivariate Markov model etc. have been proposed [1, 2, 6, 16, 17, 19, 20]. For the continuous expression data, clustering algorithms, Bayesian networks and ordinary differential equations based methods have been proposed for the network inference [4].

We are particularly interested in the continuous gene expression data in this paper. Although clustering method is not a proper network inference method, it is still widely used in the case of large volume of data. The rationale behind clustering method is that genes in the same cluster are more likely to be functionally related to each other [10]. With such a method, the high dimensionality of genes can be reduced to many small clusters of genes. However, such methods may not give a reasonable explanation of the regulatory relations. Another widely used method is the Bayesian network [11, 24]. In a Bayesian network, the relationships between the genes are encoded as a directed acyclic graph, where the parents of a gene represent its regulators. The assumption behind Bayesian network is the Markovian assumption, which states that each gene is independent of its non-descendants given its parents. This excludes the case that a gene may regulate its parent and it is the major limitation of the Bayesian network approach. To overcome this limitation, dynamic Bayesian network has been developed to infer the interactions from the time-series data sets [24]. The Bayesian networks are probabilistic models and the inference of the network is an NP-hard problem.

Another class of methods for inferring the gene networks from the continuous data is the Ordinary Differential Equation (ODE) based algorithms. They are also developed

to study the gene-gene interactions [3, 8, 12, 22]. Such approaches can describe gene regulations and result in directed graphs and they can be applied to both steady-state and time-series expression profiles. The models can be applied to predict the behavior of the network under different conditions. To infer such models, different kinds of regression methods have been applied. The usual ODE model is:

$$\dot{x}_i(t) = \sum_{j=1}^{n} a_{ij} x_j(t) + b_i u(t), \tag{1}$$

where $i = 1, 2 \ldots, n, t = 1, 2, \ldots, m, n$ is the number of genes and m is the number of time points. Here $x_i(t)$ is the concentration of Transcript i at time point t and $\dot{x}_i(t)$ is the rate of change of concentration of Gene i at time t. The parameter a_{ij} represents the influence of Gene j on Gene i and b_i represents the effect of the external perturbation on x_i and $u(t)$ represents the external perturbation at time t. In Gardner *et al.* [12], the Network Identification by multiple Regression (NIR) was proposed to compute a_{ij} from the steady-state gene expression data ($\dot{x}(t) = 0$). It requires the knowledge about which genes have been directly perturbed in each perturbation experiment. And the number of input genes is determined by the users. The Mode-of-action by Network Identification (MNI) [9] is similar to the above method. The Time Series Network Identification (TSNI) algorithm [3] is proposed to identify the networks from the time series data. All these methods are restricted to the use of perturbations. When inferring the model, the $u(t)$ is assumed to be known, which is determined when generating the data sets. In van Someren *et al.* [21], a Least Absolute Regression Network Analysis (LARNA) method is proposed. They assume the following model:

$$X_{t+1} = AX_t + \epsilon, \quad t = 1, 2, \ldots, m. \tag{2}$$

Here ϵ is used to model the noise and each entry a_{ij} of A is used to model the influence of the expression of Gene j at time t on the Gene i at time $t+1$. To get a good estimate of A, instead of solving a Least Square (LS) problem which minimizes the errors between the observed data and the predicted data, a penalty term is added. This term is used to balance the data fit term and limit the connectivity among the genes, which only assumes the interactions among the genes are sparse.

The difficulty of applying the ODE based models lies in estimation of the interaction coefficients. It is well known that many real biological networks have the scale-free property (i.e. the degree approximately follows a power-law distribution) [5]. More precisely, it is observed that in a gene regulatory network, the out-degree distribution follows a power-law and the in-degree distribution follows Poisson distribution [14]. However, all the above approaches have not considered such distributions when inferring a genetic regulatory network. In this paper, we take into account such properties. We propose to use a linear model similar to the model in van Someren *et al.* [21] for modeling the relations among the genes by using the multiple regression method. The scale-free properties are employed in the design of a filter. Such a filter is applied to filter (remove) the small nonzero entries in matrix A so that the estimated gene-gene connections matrix A will have the property that the number of nonzero entries in each row follows the Poisson distribution and the the number of nonzero entries in each column follows the power-law. Two statistical tests: t-test and χ^2-test are applied to test the power-law distribution and the Poisson distribution respectively. The Least Square (LS) method with regularization is applied to get the estimate of the filter and the estimate of the matrix A based on the obtained filter.

The rest of the paper is organized as follows. In Section 2, we present the methodology which includes the proposed model and the model inference. In Section 3, numerical examples based on the yeast data are given to illustrate the method. Finally, conclusions are given in Section 4 to address further research issues.

2 Methodology

2.1 The Linear Model

In this subsection, we present the linear model. We assume the interactions among the genes can be described by the following linear model:

$$X_{t+1} = AX_t + \epsilon_t, \quad \text{for } 1 \leq t \leq m. \tag{3}$$

Here X_t is an $n \times 1$ vector describing the expression level of n different genes at time t. A is an $n \times n$ matrix, where each entry $a_{ij}, i = 1, 2, \cdots, n; j = 1, 2, \cdots, n$ of A models the regulatory ability of Gene j to Gene i. ϵ_t is used to model the noise at time t. Given the gene expression levels of the n genes at m time points, we get the following linear equations:

$$[X_m, X_{m-1}, \cdots, X_2] = A[X_{m-1}, X_{m-2}, \cdots, X_1] + [\epsilon_{m-1}, \epsilon_{m-2}, \cdots, \epsilon_1], \tag{4}$$

For the ease of discussion, we write the linear system (4) as:

$$Y = AX + \epsilon, \tag{5}$$

where

$$Y = [X_m, X_{m-1}, \cdots, X_2], X = [X_{m-1}, X_{m-2}, \cdots, X_1], \epsilon = [\epsilon_{m-1}, \epsilon_{m-2}, \cdots, \epsilon_1]. \tag{6}$$

We denote y_k^T to be the k-th row of Y and a_k^T to be the k-th row of A, where M^T denotes the matrix transpose of M. Then, by simple observation, we have

$$y_k^T = a_k^T X \quad \text{or} \quad y_k = X^T a_k. \tag{7}$$

The latter form now looks like the standard form of the multiple linear regression for the coefficients of a_k. We note that X^T is an $(m-1) \times n$ matrix where $m-1$ is often much smaller than n. This means that the normal regression does not work as the matrix

$$(X^T)^T X^T = X X^T$$

has rank smaller than or equal to $m-1$ and is therefore singular. To give a reasonable estimate of the matrix A, here we consider using the singular value decomposition of the matrix X^T [13], $X^T = U \Sigma V^T$, where U and V have orthonormal columns u_i (left singular vectors) and v_i (right singular vectors), and Σ is a diagonal matrix with diagonal entries $\sigma_i \geq 0$, which are assumed to be arranged in descending order. A large family of estimates a_k can be expressed as a linear combination of right singular vectors v_i,

$$a_k = \sum_{i=1}^{\mathrm{rank}(X^T)} f_i \times \frac{u_i^T y_k}{\sigma_i} \times v_i. \tag{8}$$

We note that for the least squares estimate $\|y_k - X^T a_k\|_2$, the filter factors are identically equal to one, $f_i = 1$, for all i. Expressing the least squares estimate in terms of the singular value decomposition makes manifest that errors of order ϵ in the y_k typically result in errors of order $\epsilon/\sigma_{\min}$ in the estimate a_k, where $\sigma_{\min}$ is the smallest nonzero singular value. Since the matrix X^T is rank deficient, the estimate with filter factors $f_i = 1$ is the least squares estimate with minimum norm $\|a_k\|_2$.

If the matrix X^T has small singular values, regularization methods stabilize the least squares estimates by filtering out the contributions of right singular vectors v_i that are associated with the small singular values σ_i. Thus we can consider minimizing the function

$$\|y_k - X^T a_k\|_2^2 + \lambda^2 \|a_k\|_2^2, \tag{9}$$

which is the Tikhonov regularization [15]. Usually λ is taken to be between 0 and 1. The filter factors are [15]

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2} \tag{10}$$

This filter function decays smoothly from $f_i \approx 1$ for $\sigma_i \gg \lambda$ to $f_i \approx 0$ for $\sigma_i \ll \lambda$; i.e., right singular vectors with singular values smaller than λ are effectively filtered out.

After solving the above minimization problems depending on the singular values of the matrix X^T , we can get the estimate of the matrix A denoted as $A^{(1)}$. However, the matrix $A^{(1)}$ can be dense and this is not consistent with the properties of a biological network since the actual underlying gene network should not have many nonzero entries. To get a consistent estimate of the matrix A, some nonzero entries need to be filtered. We call this process a "filtering process". We fix a certain percentage of nonzero entries based on the relative magnitude of $A^{(1)}$. The actual procedure is that we first normalize the rows of $A^{(1)}$, i.e., we subtract each entry by the mean of the corresponding row and then divide each entry by the standard deviation of the corresponding row. We denote $A^{(2)}$ to be the normalized form of $A^{(1)}$. Here when the magnitude of the entries of $A^{(2)}$ is below a certain threshold, it is regarded as zero. Otherwise, it is regarded as one. All these nonzero entries reflect the connectivity among the genes. To determine the threshold, we use some statistical methods based on the

properties of the gene regulatory network. Starting from zero, we choose the minimum percentage which can make the matrix A have the given probability distributions. The details of the filter design will be addressed in the following subsection. Now, a filter $A^{(3)}$ of size $n \times n$ is obtained. The matrix $A^{(3)}$ reflects the connections among the genes. To get the regulatory abilities among the genes, we need to solve the minimization problem again based on the filter. Given the position of all the nonzero entries in the matrix A which is same as that in the matrix $A^{(3)}$, we need to solve the original optimization problem again to obtain a new estimation, we denote this solution $A^{(4)}$, which is the final solution of A.

2.2 The Filter Design

To design the filter, we note that in a genetic network, there are important properties of the in-degree and out-degree of a gene. It is well-known that the in-degree follows the Poisson distribution while the out-degree follows the power-law, i.e., the out-degree to some negative power. These provide useful criteria for determining the percentages of nonzero entries in the matrix $A^{(2)}$.

Since the in-degree distribution follows the Poisson distribution:

$$f(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{11}$$

to test whether the distribution of the in-degree of the filter follows it or not, we carry out the χ^2 Goodness-of-fit test. We choose 5 bins for the test according to the following rules:

(i) if the floor of the mean number of in-degree, denoted by k, is greater than 2, the

bins $k-2$, $k-1$, k, $k+1$ and the remaining possibilities (0 to $k-3$ and $\geq k+2$) are chosen.

(ii) otherwise the bins 0, 1, 2, 3 and ≥ 4 are chosen.

The test statistic is the following:

$$T_1 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} \tag{12}$$

which follows the χ^2 distribution with degree of freedom $4 - 1 = 3$ as the parameter λ has to be estimated from the data. In Equation (12), O_i is the observed frequency in Bin i and E_i is the expected frequency in Bin i. The p-value is calculated with the formula $\Pr\{X \geq T_1\}$. The bigger the p-value, the less likely that the in-degree follows the Poisson distribution.

Similarly we can test the out-degree distribution and the statistical test is more straightforward. We first take logarithm on both the frequency and out-degree. We then perform a simple linear regression analysis on the transformed data. Then we test for the null hypothesis that the slope (β) is zero, i.e., the transformed data has no linear relationship, consequently the original data does not follow the power-law. We remark that other statistical method such as the Coefficient of Determination (R^2) can also be used. The test statistics in our case here is

$$T_2 = \frac{\widehat{\beta}}{\mathrm{SE}(\widehat{\beta})} \tag{13}$$

where

$$\mathrm{SE}(\widehat{\beta}) = \sqrt{\frac{\sum_{i=1}^{s}(y_i - \widehat{y_i})^2}{(s-2)\sum_{i=1}^{s}(x_i - \overline{x})^2}}.$$

Here $\widehat{\beta}$ is the estimated slope using regression analysis, s is the number of data points, $\widehat{y_i}$ is the estimate of y_i using the regressed linear relation and the data x_i and $\overline{x}$ (the

mean of the data x_i). The test statistics follows Student-t distribution with a degree of freedom $s-2$. The p-value of this test is $2\Pr\{X \geq T_2\}$. The smaller the p-value, the more likely that the out-degree of the filter follows power-law.

3 Numerical Examples

In this section, we will demonstrate the procedures of our proposed algorithm. There are 384 genes in the data set which are measured at 17 time points during two cell cycles from yeast. All the genes are identified based on their peak times of five phases of the cell cycle and annotated. The levels of each gene were standardized to enhance the performance of model-based methods. The whole data set [23] can be downloaded at '*http://faculty.washington.edu/kayee/ model/*'. The description of the genes can be found at: '*http://genomics.stanford.edu*'.

With the singular vector decomposition to the matrix X^T, we can get its singular values. The minimum singular value is 2.3557. Thus we may use the least square estimate to estimate the matrix A. To get the estimate of the percentage γ of nonzero entries, the statistical tests addressed in the previous section are used. There are a number of thresholds (percentages of nonzero entries) that fulfill the in-degree and out-degree requirements under a significance level α of 5%, i.e., p-value of goodness-of-fit test is higher than α and that of the t-distribution test is lower than α. Starting from 0, we set the step size to be 0.001, and iteratively to find the minimum value that can make the gene connections have the Poisson distribution and power-law distribution. The minimum value here we obtained is 0.124. We remark that this does not mean

that when the percentage is greater than 0.124, all the networks obtained will have the scale-free property. We tested the percentage from 0 to 0.5, and found that when the percentage is greater than 0.124, in most cases(about 82%), the network has the scale-free property. Since the gene regulatory network should be very sparse [21], we take the value of γ to be 0.124. To see the sparsity patterns of the matrix A, we present the figures for $\gamma = 0.124$ and 0.180 which can also make the system have the property in Fig. 1. The patterns of these two cases look similar to each other.

We focus on the case $\gamma = 0.124$ from now onwards. Fig. 2 and Fig. 3 show the in-degree and out-degree distributions respectively. The in-degree follows the Poisson distribution and the out-degree follows the power-law. Table 1 and Table 2 show ten genes with the largest out-degree and the largest in-degree. The phase where the genes are found and their functions are also given, which are taken from:'*http://genomics.stanford.edu/yeast_cell_cycle/functional_categories.html*'. We also listed the explanations of these genes that can be found from GO in Table 3 and Table 4. From the out-degree distribution, we observe that there are a few genes having important influences on many other genes. From Table 3, we can see although some genes (YDR033w, YLR297w, YKL066w, YBR073w) with comparatively large out-degree are unknown now. These genes should influence many others genes and should be paid more attentions in the later studies. From Table 4, the function of Gene YLR236c, which has the largest in-degree is unknown either. The function of this gene may be studied starting from all its regulatory genes.

One of the main aims in modeling the gene regulatory network is to predict the gene activities. To illustrate the effectiveness of our proposed method, we set the initial

state of the whole system to be the state at time point 1, and then use our model to predict the states at all the other 16 time points. We expect that our model can fit the data very well and this is in fact the situation. We tested all the 384 genes with the obtained model and it can fit all the data quite well. Here, we only select two genes to show the results. Fig. 4 shows the predicted behavior and the errors for Gene YEL018w and Gene YLR376c, which have the largest in-degree and the smallest in-degree respectively. Thus, we can predict the behavior of all the genes in the long-term with the proposed model.

In the paper [7], cell cycle-dependent periodicity was found for 416 transcripts. We compared all the 384 genes with the genes listed in Table 1. of the paper [7], and found there are 205 common genes. We predicted the evolution of all the genes with the obtained model. Among all the 205 genes, 204 genes show the periodicity dependent on the cell cycle, which are consistent with the results in [7]. Only one gene YNR016c does not show the periodicity in the simulation. We show the simulation results for the ten genes with the largest out-degree and in-degree in the Fig. 5 and Fig. 6. Fig. 7 and Fig. 8 show the long term behavior of the above genes. The transcription level will approach zero as the time increases.

Finally, some genes may form a closed sub-network (evolution of this system can be determined by the genes in this network). We propose the following Algorithm (A) to identify a sub-network with a specified gene i_0. The input of the algorithm is: the filter $A^{(3)}$ and the specified gene number i_0. The output of the algorithm is: the genes which can construct a sub-network including the gene i_0. Using Algorithm (A), we found that given any gene in the data set, the smallest sub-network contains 239 genes.

Algorithm (A)

Initialize $V_{i_0} = (i_0)$;

For each node i in V_{i_0}, find the indices set V such that for all $j \in V$, $A_{i,j} \neq 0$;

If $j \in V_{i_0}$, return;

Else $V_{i_0} = V_{i_0} \bigcup V$; continue;

The connections among these genes can be found from the filter.

4 Conclusions

In this paper, we proposed a multiple regression model for the network inference of time series gene expression data. To infer the model, a filtering process is considered first. It is based on the properties of a real gene regulatory network. This process is to filter those redundant connections among the genes to get a good estimate of the network structure. A minimization process is considered to get the estimate of the influence coefficients of the gene relations. Since the number of time points is very small compared to the number of genes, the multiple regression can fit the data very well. Numerical examples based on the yeast data are applied to illustrate the method and the effectiveness of the method. Although the good fitness of our model to the data may result from the fact that there are more parameters than necessary, it also depends on the selection of the percentage of nonzeros γ.

For our future research, we'll compare our model with some proposed models such as the multiple regression method by Someren *et al.* [21], which did not consider the scale-free property of the network. More numerical experiments based on some other

practical data sets will be conducted to explore more interesting interactions among the genes. Our multiple regression method can be applied to the sub-network obtained from Algorithm A to get a more reliable network. The process can then be iterated until it converge to a fixed sub-network.

Acknowledgment: Research supported in part by HKRGC Grant No. 7017/07P, HKUCRGC Grants and HKU Strategy Research Theme fund on Computational Sciences.

Conflict of interest statement: No conflict of interests.

References

[1] T. Akutsu, S. Miyano and S. Kuhara, Inferring qualitative relations in genetic networks and metabolic pathways, *Bioinformatics* 16 (2000) 727-734.

[2] E. Boros, T. Ibaraki and K .Makino, Error-free and best-fit extensions of partially defined Boolean functions, *Information and Computation* 140 (1998) 254-283.

[3] M. Bansal, D. G. Giusy and D. di Bernardo, Inference of gene regulatory networks and compound mode of action from time course gene expression profiles, *Bioinformatics* 22 (2006) 815-822.

[4] M. Bansal, V. Belcastro, A. Ambesi-Impiombato and D. di Bernardo, How to infer gene networks from expression profiles, *Molecular Systems Biology* 3(2007), article number 78.

[5] A. L. Barabási and R. Albert. Emergence of scaling in random networks, *Science* 286 (1999) 509-512.

[6] W. Ching, E. Fung, M. Ng and T. Akutsu, On construction of stochastic genetic networks based on gene expression sequences, *International Journal of Neural Systems* 15 (2005) 297-310.

[7] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart and R. W. Davis, A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle, *Molecular Cell* 2 (1998) 65-73.

[8] P. D' haeseleer, X. Wen, S. Fuhrman and R. Somogyi, Linear modeling of mRNA expression levels during CNS development and injury, *Proceedings Symp Biocomput* (1999) 41-52.

[9] D. di Bernardo, M. J. Thompson, T. S. Gardner, S. E. Chobot, E. L. Eastwood, A. P. Wojtovich, S. J. Elliott, S. E. Schaus and J. J. Collins, Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks, *Nat. Biotechnol.* 23:3 (2005) 377-383.

[10] M. Eisen, P. Spellman, P. Brown and D. Botsterin, Clustering analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA* 95 (1998) 14863-14868.

[11] N. Friedman and G. Elidan. Bayesian network software libB 2.1. *http://www.cs.huji.ac.il/labs/compbio/LibB/*

[12] T. S. Gardner, D. di Bernardo, D. Lorenz and J. J. Collins, Inferring genetic networks and identifying compound mode of action via expression profiling, *Science* 301 (2003) 102-105.

[13] G. Golub and C. van Loan, *Matrix Computations* (The John Hopkins University Press, Baltimore,1993).

[14] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès, Topological and causal structure of the yeast transcriptional regulatory network, *Nature Genetics* 31 (2002) 60-63.

[15] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion* (Society for Industrial and Applied Mathematics, Philadelphia, 1998).

[16] O. Hirose, N. Nariai, Y. Tamada, H. Bannai, S. Imoto and S. Miyano, Estimating gene networks from expression data and binding location data via Boolean networks, *Lecture Note in Computer Scineces* 3480 (2006) 349-356.

[17] T. E. Ideker, V. Thorsson and R. M. Karp, Discovery of regulatory interactions through perturbation: Inference and experimental design, *Proc. Pacific Symp. Biocomputing* 5 (2000) 302-313.

[18] H. de Jong, Modeling and simulation of genetic regulatory systems: A Literature Review, *J. Comput. Biol.* 9 (2002) 69-103.

[19] K. Noda, A. Shinohara, M. Takeda, S. Matsumoto, S. Miyano and S. Kuhara, Finding genetic network from experiments by weighted network model, *Genome Informatics* 9 (1998) 141-150.

[20] I. Shmulevich, A. Saarinen, O. Yli-Harja and J. Astola, Inference of genetic regulatory networks under the best-fit extension paradigm, *Computational and Statistical Approaches To Genomics* (W. Zhang and I. Shmulevich, Eds. Boston, MA: Kluwer. 2002).

[21] E. P. van Someren, B. L. T. Vaes, W. T. Steegenga, A. M. Sijbers, K. J. Dechering and M. J. Reinders, Least Absolute Regression Network Analysis of the Murine Osteoblast Differentiation Network, *Bioinformatics* 22:4 (2006) 477-484.

[22] J. Tegner, M. K. Yeung, J. Hasty and J. J. Collins, Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling, *Proc Natl Acad Sci USA* 100:10 (2003) 5944-5949.

[23] K. Yeung and W. Ruzzo, An empirical study on principal component analysis for clustering gene expression data, *Bioinformatics* 17 (2001) 763-774.

[24] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink and E. D. Javis, Advances to Bayesian network inference for generating causal networks from observeational biological data, *Bioinformatics* 20 (2004) 3594-3603.

Figure captions:

Figure. 1: Sparsity of matrix A for $\gamma = .124$, and $.180$. 'nz' is the number of nonzero entries in the matrix A. For these two cases, matrix A has the same pattern.

Figure. 2: In-degree distributions for all the 384 genes when $\gamma = .124$. In-degree follows the Poisson distribution.

Figure. 3: Out-degree distributions for all the 384 genes when $\gamma = .124$. Out-degree follows the power-law.

Figure. 4: Prediction results: the left two figures show the prediction results and the right two figures are the prediction errors (predicted-observed). The model can fit the given experimental data well.

Figure. 5: The periodicity of the ten genes with the largest out-degree.

Figure. 6: The periodicity of the ten genes with the largest in-degree.

Figure. 7: The long term behavior of the ten genes with the largest out-degree.

Figure. 8: The long term behavior of the ten genes with the largest in-degree.

Table 1: Ten genes with the largest out-degree. Explanations are taken from: http://genomics.stanford.edu/ yeast_cell_cycle/functional_categories.html

Out-degree	Name of Gene	Phase	Function Explanation
274	YBL002w	S	DNA replication
274	YPL187w	Late G1	mating pathway
271	YDR033w	Early G1	unknown function
266	YLR297w	M	unknown function
262	YLR254c	Early G1	unknown function, hypothetical protein
261	YKL066w	S	unknown function, hypothetical protein
257	YPL256c	Late G1	cell cycle regulators
255	YBL003c	S	DNA replication
253	YBR073w	Late G1	miscellaneous
251	YPL127c	Late G1	transcription, unknown/complex phenotype

Table 2: Ten genes with the largest in-degree. Explanations are taken from: http://genomics.stanford.edu/ yeast_cell_cycle/functional_categories.html

In-degree	Name of Gene	Phase	Function Explanation
61	YEL018w	S	unknown function, weak similarity to Rad50p
61	YLR236c	Late G1	unknown function, hypothetical protein
59	YKR001c	S	biosynthesis
59	YKL165c	Late G1	biosynthesis
59	YPL209c	Late G1	chromosome, nuclear segregation
59	YLR015w	Early G1	unknown function,hypothetical protein
59	YKL163w	Early G1	unknown function, PIR3 protein with internal repeats
58	YLR228c	S	unknown function
58	YDL095w	S	biosynthesis
58	YDL124w	Late G1	unknown function

Table 3: GO terms for the ten largest out-degree genes taken from [14]

Name of Gene	GO terms
YBL002w	chromatin assembly or disassembly
YPL187w	pheromone-dependent signal transduction during conjugation with cellular fusion
YLR254c	nuclear migration, microtubule-mediated
YPL256c	re-entry into mitotic cell cycle after pheromone arrest regulation of cyclin--dependent protein kinase activity
YBL003c	chromatin assembly or disassembly, DNA repair
YPL127c	negative regulation of DNA recombination regulation of transcription, DNA-dependent

Table 4: GO terms of the ten largest in-degree genes taken from [14]

Name of Gene	GO terms
YEL018w	DNA repair
YKR001c	actin cytoskeleton organization and biogenesis peroxisome organization and biogenesis protein retention in Golgi (IMP) protein targeting to vacuole, vacuolar transport
YKL165c	ATP transport, GPI anchor biosynthetic process
YPL209c	attachment of spindle microtubules to kinetochore chromosome segregation, meiotic sister chromatid segregation mitotic spindle disassembly, regulation of cytokinesis
YLR015w	chromatin silencing at telomere, histone methylation telomere maintenance, transcription
YKL163w	cell wall organization and biogenesis
YLR228c	sterol biosynthetic process
YDL095w	protein amino acid O-linked glycosylation
YDL124w	metabolic process

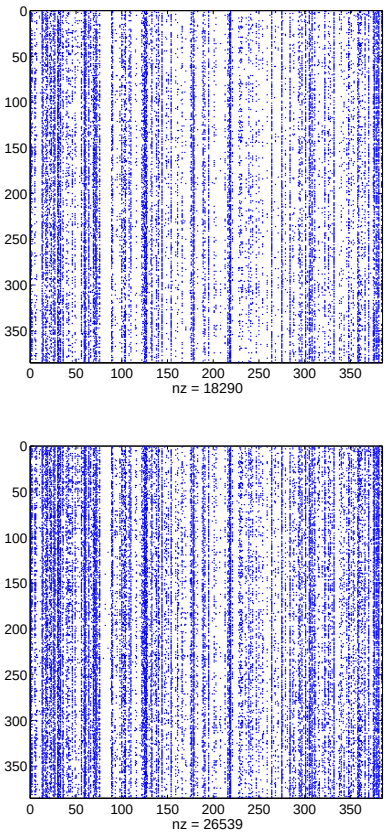

Figure 1: Sparsity of matrix A for $\gamma = .124$, and $.180$

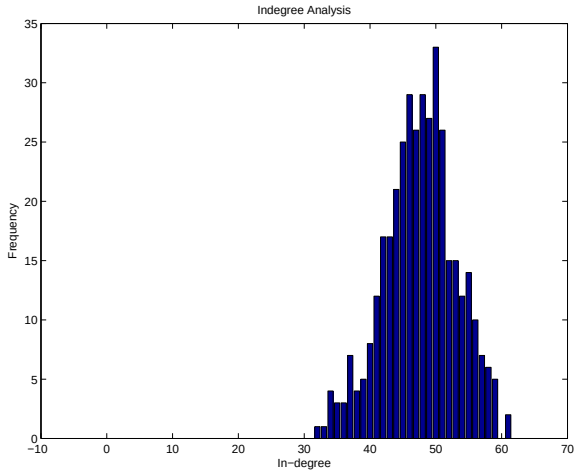

Figure 2: In-degree distributions for all the 384 genes when $\gamma = .124$

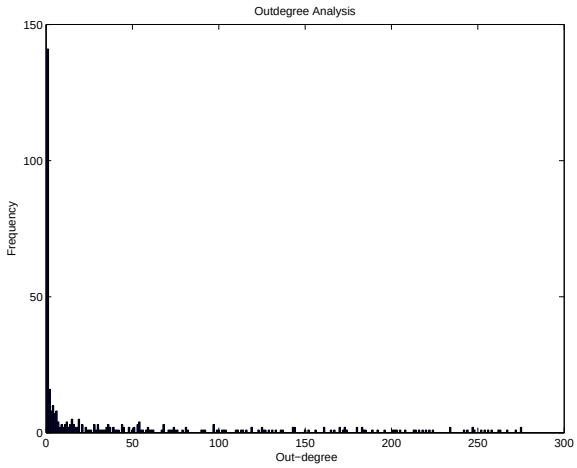

Figure 3: Out-degree distributions for all the 384 genes when $\gamma = .124$

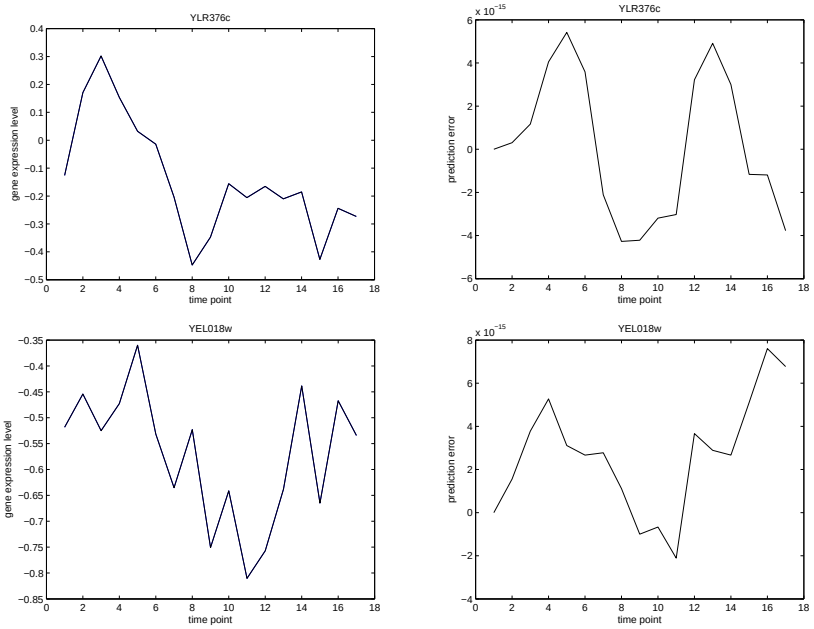

Figure 4: Prediction results: the left two figures show the prediction results and the right two figures are the prediction errors (predicted-observed)

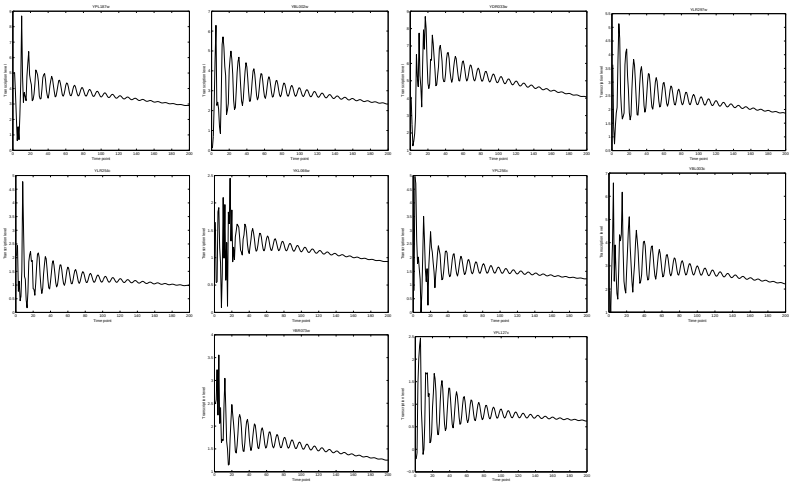

Figure 5: The periodicity of the ten genes with the largest out-degree

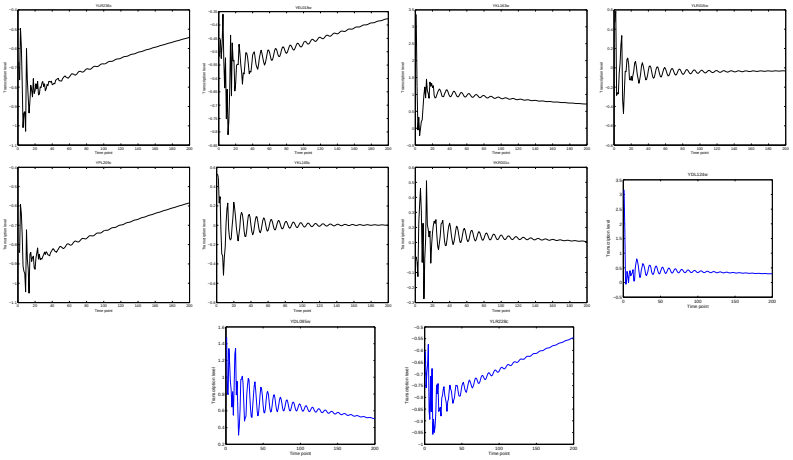

Figure 6: The periodicity of the ten genes with the largest in-degree

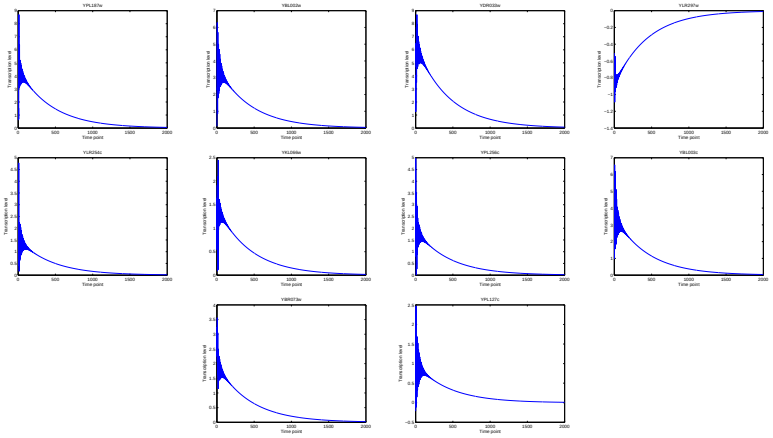

Figure 7: The long-term behavior of the ten genes with the largest out-degree

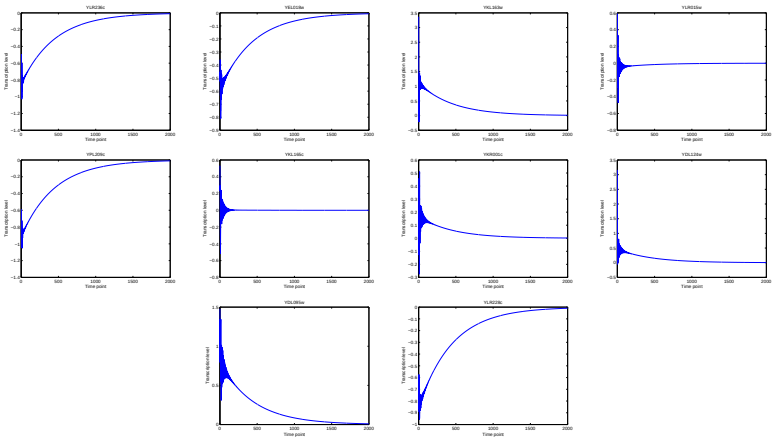

Figure 8: The long term behavior of the ten genes with the largest in-degree