The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Delay Discrete Dynamical Models for Genetic Regulatory Networks

Hao Jiang[1] Wai-Ki Ching[1,*] Kiyoko F. Aoki-Kinoshita[2,†]
DianJing Guo[3,‡]

[1] Advanced Modeling and Applied Computing Laboratory
Department of Mathematics, The University of Hong Kong, Hong Kong, China
[2] Department of Bioinformatics, Faculty of Engineering, Soka University, Tokyo, Japan
[3] Department of Biology, The Chinese University of Hong Kong, Hong Kong

Abstract In this paper, we study the problem of constructing a regulatory network of yeast in oxidative stress process. Discrete Dynamic System (DDS) model has been introduced in describing Gene Regulatory Networks (GRNs). However, delay effect was not taken into consideration within the model. A Time-delay DDS model composed of linear difference equations is developed to represent temporal interactions among significantly expressed genes. Interpolation and re-sampling are imposed to equalize the non-uniformity of sampling time points. Statistical significance plays an active role in obtaining the optimal interaction matrix of GRNs. The constructed gene network using linear multiple regression has a very good match with the original data. Simulation results are given to demonstrate the effectiveness of our proposed model.

Keywords Gene Regulatory Networks; Linear Multiple Regression; Discrete Dynamic System Model; *k*-means Clustering

1 Introduction

The study of GRNs has drawn much attention in the recent years since it can help understand the function of genes. Even though it's a scientific challenge to understand the regulations of large groups of genes, insights have been gained from various mathematical formulations describing the dynamics of GRNs.

Compared with available knowledge, past models about the behavior of molecular and cellular systems seemed to be incomplete in that key numbers are unknown [3]. Numerous studies probably have relied on simulation, see for instance [8, 12]. Although thoughtful, these simulated networks appeared to be so insignificant that biologists perceived to guarantee following experimental endeavor. With the development of experimental techniques, rapid measurement of expression levels of genes became possible. A large amount of available gene expression data make formal mathematical methods more

[*] E-mail: wching@hkusua.hku.hk
[†] E-mail: kkiyoko@soka.ac.jp
[‡] E-mail: djguo@cuhk.edu.hk

and more popular when modeling the gene regulation processes. There have been a considerable number of models describing GRNs in the literature. Directed graphs could be viewed as the most straightforward way to model a GRN. Bower and Bolouri introduced some classic models of genetic networks [6]. A Bayesian network [5] depicts the gene regulatory process from a probability perspective. The dynamic Bayesian network, an extension of Bayesian networks can describe statistical temporal dependencies among genes. However, it does not explicitly describe temporal relations among genes in a functional form. In Boolean networks and probabilistic Boolean networks [13], each variable takes the value of either 0 or 1. This nature significantly limits its capacity to discriminate quantitative differences. The generalized logical method developed by Thomas and his collegues [11] can incorporate more than two levels for each variable. It is based on the Boolean networks and has undergone several extensions. Thanks to its pinpoint accuracy in describing gene expression level, modeling GRNs in a continuous version has become widespread. Ordinary Differential Equations (ODEs) [9] model a specific gene within the network by a differential equation formulating the rate of expression level. In particular, piecewise linear differential equations and qualitative differential equations have beneficial mathematical characteristics and thereby are capable of qualitative analysis. There appears to be a strong probability that their limited up scalability is a major difficulty in simulation. An ODE-based model offers a continuous description of the regulation process. However, the processes might be considered as being discrete. Taking this into account, together with the necessity for solving ODEs using computing technology, we are inspired to develop a discrete model for transcriptional regulation. Discrete Dynamical System (DDS) Model [10], a discrete version of ODEs, assists one understanding interactions among variables systematically. It has gained a solid foot in quantitative modeling of GRNs. The initial application of DDS model in mathematical biology is Verhulst equation, a single variable DDS model used for population dynamics. Using the least squares method for estimating system coefficients, the linear DDS model of mRNA expression levels proves to be biologically plausible. Based on the relatively coarse model, Song et al. [10] proposed a modified DDS model which avoids some problems in the previous one. It imposes log-time interpolation to equalize the non-uniformity in original time domain and assesses the statistical significance of equations for specific genes. The third innovation it makes is to use eigenvalue normalization to perform power stability to the whole system. Concisely speaking, a DDS model can be described as a series of difference equations:

$$\frac{g[t]-g[t-1]}{h} = Ag[t-1] + Be[t-1] + \varepsilon[t] \tag{1}$$

where $g[t] = (g_1[t], g_2[t], ...g_N[t])$ is a vector of expression levels at time t, N is the number of genes invloved. The entries of A, A_{ij} is the influence of gene j on gene i. The entries of B, B_{ik} is the influence of the kth stimulus on gene i. The term $\varepsilon[t]$ represents noise levels at time t. The classic estimation method is to utilize least squares.

A statistically significant DDS model which consists of linear difference equations can be utilized to infer transcriptional regulations. It also accelerates the characterization of gene interactions. The estimation of parameters is easy to implement and the constructed model itself reveals some biological meaning. Although considerable research has been devoted to model GRNs, this needs enlarging to a broader sense such as modeling the genetic network via a discrete dynamic system model with time-delay in that the rate of expression levels at time t does not only depend on expression levels on t, but also it is influenced by time $t-1$. If non time-delay model can reveal some biological sense, then the model taking into account the delay effect probably can provide more sound biological meaning.

The remainder of the paper is structured as follows. Section 2 presents the proposed Delay DDS model. In Section 3, we utilize the model to construct GRNs of the yeast. Results and Discussion are presented in Section 4. Finally, conclusions and potential future work are given in the last Section.

2 DDS Model with Delay Effect

The model developed by Song et al. [10] can be deemed as a power-stable significant DDS model. However, simulation of the model shows that the results without stabilization turned out to be more reasonable during initial period before nonlinear effect takes place. The result seemed to be distorted by the procedure of being stabilized. One interpretation might be due to the small number of points needing to be predicted. On the other hand, the major aim of stabilization is to ensure the stability of the system as time goes to infinity, but the DDS model proposed here indeed is to predict the expression levels of the network before perturbation takes place, namely two hours in total. It possibly poses no need to stabilize the system. In the stabilisation of the DDS model, W was replaced by W_s defined as $\frac{1}{\rho(W)}W$ if $\rho(W) > 1$, where $W = hA + I, W = V \wedge V^{-1}$, $\rho(W) = max\{|\lambda| : \lambda \in \lambda(W)\}$. A was stablized as A_s: $\frac{1}{h}[\frac{hA+I}{\rho(hA+I)} - I]$ if $\rho(W) > 1$. The final form of the model was retrieved as:

$$\frac{g[t]-g[t-1]}{h} = \begin{cases} \{\frac{1}{h}[\frac{hA+I}{\rho(hA+I)} - I]g[t-1] + Be[t-1]\} + \varepsilon[t] & \text{if } \rho(W) > 1 \\ \{Ag[t-1] + Be[t-1]\} + \varepsilon[t] & \text{otherwise} \end{cases} \tag{2}$$

Therefore, if $\rho(W) > 1$, we have $g[t] = \frac{hA+I}{\max\{|\lambda|:\lambda\in\lambda(hA+I)\}} g[t-1] + hBe[t-1] + h\varepsilon[t]$. It does make sense in the long run as time goes by, but such procedures may probably be redundant in the initial phase with only few values to be predicted.

Nevertheless, in the model developed above, the change rate of the expression level merely relying on the current expression levels of the system tended to be biased. A considerable number of examples have shown that information transmission even near the speed of light may not be fast enough to ignore the effects of delays. There is a possibility that the delays might cause a completely different type of behavior that is absent in model without delay.[4]

Delays in various processes such as transcription, transportation and translation result in time delay in gene regulation processes. Bliss et al. [1] were some of the first to explicitly consider transcriptional and translational delays in their modeling of the tryptophan operon. A mixed integer linear programming framework [2] for inferring time delay in GRNs was described on account of the effect of time-delay. This key attribute of the regulatory structure is essential to ensure that the proposed inference model accurately captures the dynamics of the system.Time delays have been considerably used in biology, population dynamics for instance. They are ubiquitous in the biological sciences but are not always well-represented. In mathematical models, it may be more practicable to add the time-delay parameter. This seems to be more authentic and therefore may be more biologically plausible. Taking time-delay effect into consideration, one may extend this existing DDS model to a modified one. The Delay-DDS model can be written as

$$g[t] = g[t-1] + hAg[t-2] + hBe[t-1] + h\varepsilon[t] \tag{3}$$

and $t, t-1, t-2$ represent discrete time points, h represents time interval between two consecutive time points. The matrices A and B are to be estimated. This model describes an situation when the change rate of expression level of genes at time t not only relies on time t but also depends on $t-1$. Here we do not model multiple time-delay effect partly due to the confinement of sampling points in that even though interpolation is introduced, it has drawbacks in cementing measurement errors and inducing arbitrariness. [7]. Our aim is to use the minimum interpolation points required. Usually there are a lot more genes than time points in standard gene expression profiling. This makes the unique determination of gene interaction matrix of linear models impossible to realize. We are therefore encouraged to find the optimal one among those interaction matrices. The procedures will be discussed in details in the following sections.

3 Constructing GRNs of Yeast Using Delay DDS Model

We apply the Delay DDS model to the time-course microarray measurements of relative expression levels among genes in yeast(Saccharomyces cerevisiae) during early exposure to 150mM cumene hydroperoxide treatment. The data was normalized using (Robust Multichip Average)RMA algorithm implemented in a R package named Affy from the Bioconductor website (http://www.bioconductor.org) and listed are the genes that show significant changes comparing to mock-treated controls. Data can be accessed at (http://hkumath.hku.hk/ wkc/papers/all-de-yeast1.xls)

3.1 Data Preprocessing

The data is obtained from [14]. They were used as the log-transformed expression data to correct system bias. Due to the large amount of gene data, lots of genes may share similarities in performance which indicate their related expression patterns. They can then be treated as a single gene with one as a representative. This can largely reduce the computational cost while obtaining the major features of the network as well. Various algorithms exist for clustering in the literature such as “Hierarchical Clustering”, “Partitional Clustering” and “Spectral Clustering”. Here we employ the “k-means clustering” which is a kind of “Partitional Clustering” methods for its simplicity and fast speed in processing large datasets.

3.2 Interpolation

Since there are only 5 data points and they are sampled at $0min$, $15min$, $30min$, $60min$, $120min$, it would result in data over-fitting when implementing delay-DDS model merely by this information. To overcome the deficit, data interpolation is used. Here we adopt cubic-spline interpolation at time $45min, 75min, 90min, 105min$ in order to ensure the equality of time distance.

3.3 Multiple Linear Regression with Statistical Significance

With the complementary interpolated data, linear multiple regression is enforced to reconstruct the GRNs. Sparseness in GRN allows for the small number of non-zero parameters in each equation for every specific gene. In this Delay-DDS model, two is the maximum number of non-zero parameters in the influencing number of genes. For each difference equation, p-value in the F-test is utilized to measure the statistical significance. The overall p-value of the model can be expressed as: $p - value = 1 - \prod_{i=1}^{N}(1 - p_i)$, p_i is the significance value of the F-test during fitting of a linear model for gene i. Considering the maximum number of influence genes 2, there are C_N^2 possibilities of traversal, where p_i is chosen as the minimum of those p_is among the C_N^2 possibilities.

4 Results and Discussions

In the following, we will present some results illustrating the effectiveness of our Delay-DDS model. There are 4 models for comparison in total.

Model 1: Discrete Dynamic System Model.
The model is of the form in Equation1. The parameters of the model have been explained previously in section 1. In the process of parameter estimation, data is used in the log-form to save system bias. Log-time transformation is introduced to equalize the non-uninformity of time distances. The logarithm transform on time is defined by $t' = log(t + t_0)$, where t' is the time variable in the log-time domain. Selection of the constant t_0 is determined by the extent of equalization between the consecutive pair of time points after the log-time transform. Linear-multiple regression is performed to construct the model.

Model 2: Piecewise DDS Model.
The model is an improvement of Model 1. As in the log-time transformation process, the information at $0min$ seems to be lost. This may pose some unnecessary error when predicting the value at other time points. This piecewise model is an extension of Model 1 in that it first use Model 1 to predict the value at $15min, 30min, 60min, 120min$. Then it proceeds to use the predicted value of $15min, 30min, 45min, 60min$, together with the original value at $0min$. With the 5 values, Model 1 is implemented again to predict the values at $15min, 30min$ in the context of no log-time transformation. Because here time distance is equal,there is no need to use log-time transform.

Model 3: Nondelay DDS Model.
The model unlike the previous two doesn't use log-time transformation,instead, it uses cubic interpolation to equalize the non-uniformity of time sampling. The form of Non-delay DDS Model is the same as Model 1. Cubic interpolation is utilized at time $45min$, $75min$, $90min$, $105min$. In this situation will there be 9 data points in total. With the 9 values, Nondelay DDS model can be performed to predict the model parameters using linear multiple regression.

Model 4: Delay DDS Model.
This model is slightly different from Nondelay DDS Model which has been forumlated in Equation 3. This model takes account of the effect in previous discrete time points. But the procedure of parameter estimation is the same as Model 3.

Since for "k-means clustering", different centroids were chosen initially which may lead to different kinds of clustering results. Hence, we perform more than 10 times of the clustering program to get a relatively significant clustering result. In the dataset, we have 4118 genes in total. Thanks to the expression similarities in genes, we allow the maximum of clustering number to be 200. Clustering result will be adopted when each cluster has more than 20 members. Finally the most frequent clustering number 43 was selected when running the clustering program and simultaneously the one that located at the center of the cluster was picked out as representative. With the existing 43 groups, 4 different models were imposed on them to predict the values at the sampled time points. For the noise $\varepsilon[t]$, we all model it Gaussian. The results were illustrated in the following figures. We can have a clear understanding on the superiority of the delay-DDS model from the simulation results.

Figure 1 displays the prediction performance in cluster 1 in two models: DDS model without normalization and DDS model with normalization. Here, normalization means the stabilization procedure as described in Section 2. DDS model without normalization is the DDS model without stabilization. Data(y-axis) is measured as log-transformed fold change in expression measurement relative to the untreated control. '1-2–NonNormalized' depicts the prediction performance of DDS model without stabilization, the prediction error of the model was described by '2-2–NonNormalized'; '1-3–Normalized' depicts the prediction performance of DDS model stabilized, with '2-3–Normalized' shows the prediction error of the stabilized DDS model. From the figure we can know that Non-Normalized model predicts the value with error of $1e-2$, while Normalized model tends to distort original values and yield huge errors. It clearly explains that model without normalization is much better than the one normalized. Normalization is to predict the long time behavior while here we only need to predict few data points in the initial phase. This motivates us to develop the model without normalization.

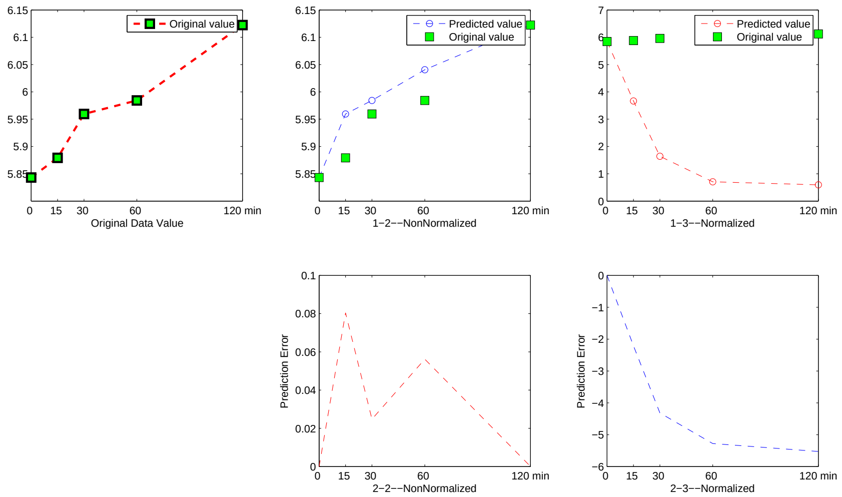

Figure 1: Comparison of DDS Model without stablization with DDS Model

Figure 2 is the comparison result of four different models in predicting performance of cluster 1. Data(y-axis) is measured as log-transformed fold change in expression mea-

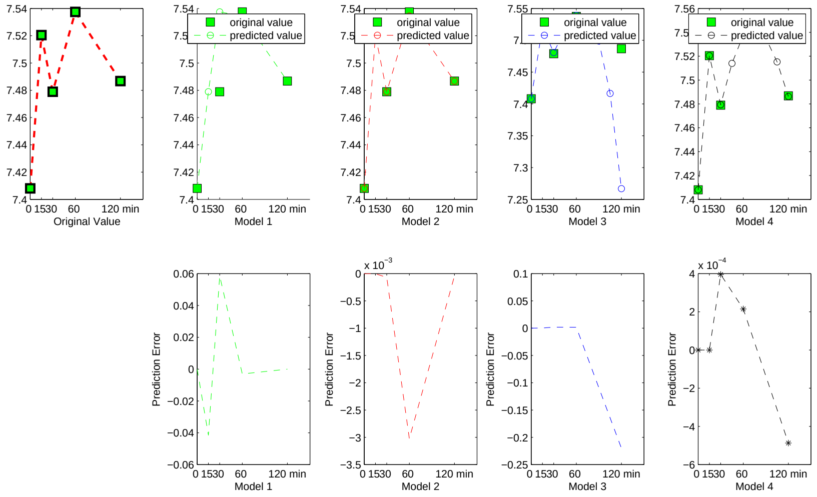

Figure 2: Comparison of 4 models

surement relative to the untreated control. From Figure 1, conclusions can be made that models without normalization exhibit superiority and we will then focus on them. Therefore, the four models are all non-normalized. The first row is the prediction result, with the second row the error of prediction results versus the original values. Model 1,2,3 and 4 exactly correspond to the 4 models described in Section 4. The first figure in row 1 is the original value, with the second to fifth figure showing the prediction performance of Model 1 to Model 4 respectively; the figures in row 2 illustrate the corresponding prediction errors of the 4 models. Delay DDS model(Model 4) exhibits superiority among the four models. It predicts the value with minimal error at about $1e-4$, while piecewise model(Model 2) ranks 2 at predicting with error at about $1e-3$, which shows better performance than original DDS model(Model 1). This is consistent with the fact that piecewise model is an improvement of original DDS model. Even though interpolation is used in Nondelay DDS model(Model 3), the prediction result is not satisfactory, the error of which is always maximum among the four, which shows inferiority to original DDS model. This also explains the positive effect of log-time transformation takes in model construction. In a word, the performance of the four models encourages us to take into account of the effect of time-delay in modeling.

5 Concluding Remarks

This paper mainly deals with the statistical properties of the data set while the biological meaning from this model is not discussed. Therefore some of the procedures in dealing with the data may lack biological stringency. For instance, we just use “k-means” in clustering the data while this method just cluster the data in a mathematical way without considering any prior biological information.

The model is easy to implement and matches the data well. Considering the performance of Delay DDS model versus DDS model without delay effect, the model with delay significantly outperforms. This might imply time lag does exist in oxidatative stress process of yeast. From the good performance of the delay-DDS model we may use it to extract some biological information hidden inside. It may also aid in intuitive understanding of the mechanisms which would give suggestions to biologists on discovering unannotated functions of genes in GRNs. These would be our future research extensions.

Acknowledgements

Research supported in part by HKGRF Grant No. 7017/07P, HKUCRGC Grants, HKU Strategy Research Theme fund on Computational Sciences, National Natural Science Foundation of China Grant No. 10971075 and Guangdong Provincial Natural Science Grant No. 9151063101000021.

References

[1] R.D.Bliss, R.P.Painter, A.G.Marr. Role of feedback inhibition in stabilizing the classical operon. J. Theor. Biol. 97 (1982) 177-193.

[2] M.S.Dasika, A.Gupta, C.D.Maranas. A mixed integer linear programming (MILP) framework for infering time delay in gene regulatory networks. Pac. Sym. Biocomput. 9 (2004) 474-485.

[3] D. Endy, R. Brent. Modeling cellular behavior. Nature. 409(6818) (2001) 391-395.

[4] J.F.Feng, J.Jost and M.P.Qian. Networks: From biology to theory. Eds. London: Springer; (2007).

[5] N.Friedman, M.Linial, I.Nachman, D.Pe'er D. Using Bayesian networks to analyze expression data. J. Comput. Biol. 7(3-4) (2000) 601-620.

[6] M.Gibson and E.Mjolsness. Modeling the activity of single genes. Bower JM ,Bolouri H,ed. computational modeling of genetic and biochemical networks. eds. Cambridge MA: MIT Press; (2001) chapter 1.

[7] R.Guthke, U.Moller, M.Hoffmann, F.Thies, S.Topfer. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection.Bioinformatics. 21(8) (2005) 1626-1634.

[8] F.Christian, Lehner, H.Patrick, O'Farrel. The roles of drosophila cyclins A and B in mitotic control. Cell. 61(3) (1990) 535-547.

[9] P.Smolen, D.Baxter, J.Byrne. Modeling transcriptional control in gene networks: Methods, recent results, and future directions. Bull. Math. Biol. 62(2) (2000) 247-292.

[10] M.Song, Z.OuYang, Z.Liu. Discrete dynamical system modeling for gene regulatory networks of 5-hxymethylfurfural tolerance for ethanologenic yeast. IET Syst. Biol. 3(3) (2009) 203-218.

[11] R.Thomas, R.d'Ari. Biological feedback. Boca Raton, FL: CRC Press; (1990) Chapter 8.

[12] J.John, Tyson. Modeling the cell division cycle: cdc2 and cyclin interactions. Proc. Natl. Acad. Sci. USA. 88(16) (1991) 7328-7332.

[13] S.Zhang. Mathematical models and algorithms for genetic regulatory networks [PhD's thesis]. Hong Kong: The University of Hong Kong; (2007).

[14] S.Zhang, W.Ching, N.Tsing, H.Leung, D.Guo. A new multiple regression approach for the construction of genetic regulatory networks. Journal of Artificial Intelligence in Medicine, 48 (2010) 153-160.