Evaluation in Dynamic System Modeling of Gene Networks

Zhengyu Ouyang[1] and Joe Song[2]

Short Abstract **—Hypothesis testing is introduced to evaluate data-driven dynamic system modeling of gene networks. Simulation studies of the residuals between noisy observations and true system dynamics suggests the use of statistical hypothesis testing to evaluate how significant a model is supported by observed data, when the noise distribution is taken into account. This method can also evaluate the fitting of each individual gene. The statistical power of the method is demonstrated through simulated and real biological dynamic systems.**

Keywords **— Dynamic system model evaluation, gene regulatory networks**

I. Introduction

HIGH throughput systematic and dynamic measurements of gene networks become available due to rapid advancement in biotechnology. The information enables potential discoveries about gene regulations, but it is also challenging to gain biological insights completely by intuitive and manual inspection of the data without a system modeling perspective. Data-driven dynamic system modeling is such a method to analyze systematic and dynamic gene expression data [1].

Several data-driven gene network models have been developed, including Boolean networks, dynamic Bayesian networks, and systems of ordinary differential equations [2]. However, there is still a paucity of literature on how to evaluate dynamic system models. Some basic performance evaluation methods [3] involve correlation or difference between the observation and the model prediction. These approaches often use how well a model fits the data to evaluate the model but would typically not provide a statistical assessment, and thus not be able to make a sensible evaluation of the model in the presence of noise. A potential problem is that a model can fit the data so well as to overfit, so that previously unseen data can be far away from the model prediction; on the other hand, an imperfectly fitted model may indicate that the observations are too noisy for the modeling to uncover any significant systematic behavior in a gene network. Therefore, without a statistical assessment, it is difficult to use model fitting to conclude that a model has captured significant biological interactions among genes.

Hypothesis testing has been utilized to evaluate dynamic system models [4]. We believe it facilitates a theoretically sound framework for model evaluation, but the noise analysis is often challenging and there is still large room to improve the statistical power of the tests.

II. Method And Results

A first-order linear discrete dynamic system model [4] is used in noise analysis and the hypothesis testing design.

We will demonstrate the residual distribution under the assumption that an observation contains two kinds of noise: the biological noise which is difference between individuals, and the measurement noise due to instruments.

The null hypothesis assumes no change of gene expression. The alternative hypothesis assumes a discrete dynamic system model. The test statistic is defined as

$$\Lambda = \frac{(RSS_0/\sigma_0^2 - RSS_1/\sigma_1^2)/p}{RSS_1/(\sigma_1^2 \times (N-p-1))}, \tag{1}$$

where p is the number of parameters, N is the sample size, RSS_0 and RSS_1 are the residual under the null and alternative hypotheses, and σ_0^2 and σ_1^2 are the residual variance which represent the noise. The distribution of Λ are simulated with different strength of noise. The test shows better statistical power than an obvious alternative similar to the *F*-test.

We evaluate the dynamic system modeling on gene expression in the cell cycle of yeast, as well as on a real biological reaction network.

III. Conclusion

The *p*-value of a hypothesis test quantifies the significance of a model as supported by the data. If the model is significant, it implies that either part of or the entire model is sufficiently supported by the data. If the significance of the model is not obvious, we inspect the distribution of residuals. Normality of the residuals can suggest that the data is noisy and further improvement of the model is unlikely. If the residuals are not normally distributed, the complexity of the model needs to be increased.

References

[1] Janes KA and Yaffe MB (Nov 2006), Data-driven modelling of signal-transduction networks, Nat Rev Mol Cell Biol 7(11), pp.820-8.

[2] Michael E. Driscolla and Timothy S. Gardner (March 2006), Identification and control of gene networks in living organisms via supervised and unsupervised learning, Journal of Process Control, Vol.16(3), pp. 303-311.

[3] Willmott, C.J. (1982): Some Comments on the Evaluation of Model Performance. *Bull. Amer. Meteor. Soc.*, 63, pp.1309–1313.

[4] Song, M. and Liu, Z. (2007). A linear discrete dynamic system model for temporal gene interaction and regulatory network influence in response to bioethanol conversion inhibitor HMF for ethanologenic yeast. Lecture Notes in Bioinformatics. 4532:77-95.

Acknowledgements: Grants from NSF, NIH, and USDA.

[1,2]Department of Computer Science, New Mexico State University, Las Cruces, NM 88003. E-mail: [1]oyoung@nmsu.edu, [2] joemsong@cs.nmsu.edu