HHS Public Access

Author manuscript

Proc Int Joint Conf Bioinform Syst Biol Intell Comput. Author manuscript; available in PMC 2016 December 09.

Published in final edited form as:

Proc Int Joint Conf Bioinform Syst Biol Intell Comput. 2009 August ; 2009: 472–478. doi:10.1109/IJCBS.2009.10.

Statistical Analysis of Discrete Dynamical System Models for Biological Networks

Zhengyu Ouyang and

Department of Computer Science, New Mexico State University, Las Cruces, NM, U.S.A

Mingzhou (Joe) Song

Department of Computer Science, New Mexico State University, Las Cruces, NM, U.S.A

Zhengyu Ouyang: oyoung@nmsu.edu; Mingzhou (Joe) Song: joemsong@cs.nmsu.edu

Abstract

Very few data-driven methods for dynamic biological networks reconstruction from gene expression data evaluate the statistical significance of a model. A hypothesis testing procedure examining the goodness of fit of trajectory-based modeling is designed, in contrast to transition-based model fitting. The former has substantially reduced the modeling error. Simulation studies on the residual between noisy observations and true system dynamics suggest the use of the statistical hypothesis testing, so that one can evaluate how significantly a model is supported by the observed data under certain noise distribution. This method can also evaluate the dynamic model for each individual gene. Through a biochemical reaction model in the yeast pheromone pathway the effectiveness of the proposed evaluation procedure is demonstrated.

1. Introduction

Statistical significance evaluation of dynamical system models for biological networks is crucial before a reconstructed model is delivered to a biologist. In typical system identification problems in control theory where sample size is much larger than dimensionality, statistical significance is often automatically assumed; while in dynamic biological network modeling, one often faces small sample size but much higher dimensionality and has to answer the question whether a model is significantly justified by the data. The modeling demands have been driven by high throughput instrumentation at molecular level, which can provide snapshots of thousands of biochemical concentrations within a cell. The snapshots can be acquired over time to understand the underlying system mechanism in reaction to external stimuli [1]. The vast magnitude of such experimental data enables potential discoveries about gene interactions in a biological network. Not surprisingly, various models have been designed for dynamic biological networks [2], such as Boolean networks (BNs) and their generalizations, dynamic Bayesian networks (DBNs) and differential equations networks (DENs).

Boolean networks are the simplest dynamical system model and were used for modeling gene regulatory networks in 1960's [3]. Probabilistic Boolean networks (PBNs) [4], extending the rule-based properties of BNs, are robust under uncertainty and represent dynamics in the probabilistic context of Markov chains.

Bayesian networks, a graph-based model of joint probability distributions, are effective at capturing conditional independence among variables [5]. DBNs [6] enable modeling of the dynamical aspect of biological interactions. As data-driven DBN modeling is NP -hard, DBN modeling is under constant improvement [7]; and DBN has been shown to subsume certain families of PBNs [8].

Difference and differential equation networks are also dynamical system models [9]. DENs allow detailed description of network dynamics, by explicitly modeling the concentration changes of molecules over time. The data may contain thousands of genes taken over fewer than tens of time-steps. Thus the modeling may be subject to the curse of dimensionality, though one may expand the sample sizes by adding noisy duplicates [10] or combining time-course datasets [11].

Data-driven methods used by PBNs or DBNs for reconstructing biological networks are transition-based under Markovian assumption, and they optimally fit transitions from one time point to the next. This does not necessarily lead to minimization of residuals from model prediction to the corresponding observation. Additionally, statistical significance evaluation of neither the entire nor partial network has been systematically studied in the past. Hereby, we focus on the evaluation procedure of data-driven methods for reconstruction of biological networks in the framework of discrete dynamical system (DDS) modeling. After optimizing transition-based fitting [12] to obtain an initial estimate of a DDS model, we continue to optimize the system parameters to fit the original observations. We then analyze the residuals and design a hypothesis testing procedure which determines the significance of a model as supported by data. The effectiveness of the procedure will be illustrated by a real biochemical reaction network model for the yeast pheromone pathway [13].

2. Transition- versus trajectory-based DDS modeling

Many data-driven complex dynamical system modeling algorithms are based on the Markovian assumption leading to the reduction of the DDS model fitting problem to transition-based [12] model fitting. Since real systems are not ideally Markovian, a transition-based approach does not immediately lead to a DDS model that fits best to observed trajectories. However, directly fitting a DDS model to the observed trajectories, instead of the transitions, is a high dimensional hybrid numerical and combinatorial optimization problem and cannot be readily solved. In our approach, we first use transition-based model fitting to obtain quantitative relationships in a biological network. This first step mostly identifies non-zero coefficients in a system. Then we perform trajectory-based DDS model fitting to minimize residuals by optimizing the non-zero coefficients. The first-order linear DDS model [12] is defined as

$$\frac{g[t]-g[t-1]}{h}=A'g[t-1]+B'e[t-1]+\varepsilon[t] \quad (1)$$

where $\mathbf{g}[t]$ represents the concentration value of all N nodes at time t , t is the observation index, $t = 1, \dots, T-1$, T is the total number observation time points, h is the actual time interval between t and $t-1$, $\mathbf{e}[t]$ is the strength of K external signals at time t , A' is the system matrix ($N \times N$) representing the regulatory network, non-zero element a'_{ij} in A' means node i is regulated by node j , B' is the influence matrix ($N \times K$), and $\boldsymbol{\varepsilon}[t]$ is a vector of noise to all variable at time t .

We also define the first-order quadratic DDS model, for each variable $i = 1, \dots, N$, by

$$\frac{g_i[t] - g_i[t-1]}{h} = \mathbf{a}_i^T \mathbf{g}[t-1] + \mathbf{g}^T[t-1] Q_i \mathbf{g}[t-1] + \mathbf{b}_i^T \mathbf{e}[t-1] + \varepsilon[t] \quad (2)$$

where Q_i is the $N \times N$ quadratic coefficient matrix for gene i , and all other coefficients and variables are the same as the first-order linear DDS model.

We illustrate the transition- and trajectory-based fitting using the first-order linear DDS model. The fitting principle is the same for the quadratic DDS model, even though the actual computation might be intensified due to the complexity introduced by the quadratic terms.

2.1. Transition-based DDS model fitting

In transition-based DDS modeling, we solve A' and B' row-by-row which leads to N multiple linear regression problems by combinatorial least squares as

$$\arg \min_{\mathbf{a}'_i, \mathbf{b}'_i} \sum_{t=1}^T (\mathbf{a}'_i{}^T \mathbf{g}[t-1] + \mathbf{b}'_i{}^T \mathbf{e}[t-1] - y_i[t])^2 \quad (3)$$

where $y_i[t] = (g_i[t] - g_i[t-1])/h$, $\mathbf{a}'_i{}^T$ and $\mathbf{b}'_i{}^T$ are the i -th row in A' and B' , respectively. The solution details can be found in [12]. Biologically, it is unlikely for node i to be controlled by all other nodes. Thus, we limit the maximum number of non-zero coefficients to a given number. For each set of feasible non-zero coefficients, we compute the p -value of the corresponding multiple linear regression. Then we solve for $\mathbf{a}'_i$ and $\mathbf{b}'_i$ by minimizing the p -value of the multiple linear regression.

2.2. Trajectory-based DDS model fitting

While transition-based fitting provides an initial solution for A and B , trajectory-based fitting-minimizes the residuals between the predictions and the observations. The prediction of a linear DDS model is given by

$$\hat{\mathbf{g}}[t] = \begin{cases} \mathbf{g}[0] & \text{if } t=0; \\ (A\hat{\mathbf{g}}[t-1] + B\mathbf{e}[t-1])h + \hat{\mathbf{g}}[t-1] & \text{if } t>0; \end{cases} \quad (4)$$

The objective of trajectory-based fitting is to

$$\arg \min_{A,B} \left(\sum_{t=0}^T \|\mathbf{g}[t] - \hat{\mathbf{g}}[t]\|^2 \right) \quad (5)$$

Note that the minimization will only be applied to the nonzero coefficients in A' and B' obtained from transition-based fitting and the t as mentioned before is observation index, which is from 0 to $T-1$. We used the Levenberg-Marquardt algorithm to do the unconstrained linear least-square optimization [14]. Figure 1 shows the observations and predictions of two models for the first node, out of a total of nine nodes. One prediction was made by the model obtained with transition-based fitting; the other with the trajectory-based fitting. The initial states used in the predictions of the two models are the same. This toy-example has indicated that the trajectory-based solution is much superior to the transition-based solution, as the former is closer to the mean of the trajectory - the best one can do theoretically. Because we can only assume the mean of the replicas is the truth from the data.

3. Statistical analysis of DDS modeling error

Considering underfitting or overfitting problems, we will propose a statistical method to analyze the fitting results after introducing the properties of modeling error. When we formulate the DDS reconstruction problem in the framework of statistical hypothesis testing, the null hypothesis states that there is no interaction between any nodes. The alternative hypothesis states that there are some interactions (directed edges) among some nodes. Thus, when a non-singleton DDS model is reconstructed, the test statistic used in the reconstruction based on observed data must reject the null hypothesis with statistical significance. The test statistic follows certain distribution determined by the noise in the observed dynamical data. We assume that the trajectory data contains two types of additive noise: *biologicalnoise* due to random differences between individuals, and *measurementnoise* due to instrumentation. Let vector $\mathbf{x}[t]$ represents the ground-truth state at time t , the observation vector is defined as

$$\mathbf{o}[t] = \mathbf{x}_b[t] + \varepsilon_m[t]$$

where vector $\varepsilon_m[t]$ is measurement noise at time t , and $\mathbf{x}_b[t]$ is the sum of $\mathbf{x}[t]$ and biological noise

$$\mathbf{x}_b[t] = \begin{cases} \mathbf{x}[0] + \varepsilon_b[0] & \text{if } t=0; \\ f(\mathbf{x}_b[t-1], \mathbf{x}_b[t-2], \dots, \mathbf{x}_b[0]) + \varepsilon_b[t] & \text{if } t>0; \end{cases} \quad (6)$$

where vector $\varepsilon_b[t]$ is the biological noise at time t , and f is function defining the relationship among nodes.

3.1. The residual distribution under null hypothesis of no interaction

Under the null hypothesis, as shown in Eq. (6), if the output of function f is constant, there is no relationship among any nodes. The system dynamics are caused simply by noise as

$$\mathbf{o}[t] = \mathbf{x} + \varepsilon_b[t] + \varepsilon_m[t] \quad (7)$$

where $\mathbf{x}$ is time-invariant. If we assume noise ε_b and ε_m are independent and zero-mean normally distributed random vectors, it follows that the residual $(\mathbf{x} - \mathbf{o}[t])$ is also normally distributed. Simulation has verified that the residual under the noise assumptions made here does not increase over time.

3.2. The residual distribution under alternative hypothesis of interaction

We believe that additional complexity for a model cannot improve the fitting when the residual is already normally distributed with zero mean. Under the first-order linear DDS model, we have

$$\mathbf{o}[t] = \mathbf{x}_b[t] + \varepsilon_m[t] \quad (8)$$

where

$$\mathbf{x}_b[t] = \begin{cases} \mathbf{x}[0] & \text{if } t=0; \\ (A\mathbf{x}_b[t-1] + B\mathbf{e}[t-1])h + \mathbf{x}_b[t-1] + \varepsilon_b[t] & \text{if } t>0; \end{cases}$$

We define the residuals at time t as

$$\mathbf{r}[t] = \mathbf{r}_b[t] + \varepsilon_m[t] \quad (9)$$

where $\mathbf{r}_b$ represents the residuals caused by biological noise ($t > 0$)

$$\begin{aligned}
\mathbf{r}_b[t] &= \mathbf{x}_b[t] - \mathbf{x}[t] \\
&= (A\mathbf{x}[t-1] + B\mathbf{e}[t-1])h + \mathbf{x}[t-1] - \mathbf{x}[t] + A\mathbf{r}_b[t-1]h + \mathbf{r}_b[t-1] + \varepsilon_b[t] \\
&= \begin{cases} \varepsilon_b[0] & \text{if } t=0; \\ A\mathbf{r}_b[t-1]h + \mathbf{r}_b[t-1] + \varepsilon_b[t] & \text{if } t>0; \end{cases}
\end{aligned}$$

While shown above theoretically, our simulation has also confirmed that the residual under the alternative hypothesis of interaction increases over time.

3.3. The covariance structure of residual under interaction assumption

Under the first-order linear DDS model, the residuals of different nodes become more entangled with each other over time. By Eq. (9), the residuals of different nodes are not independent under the alternative hypothesis. For example, a DDS model of five variables (A, B, C, D, E) has a system matrix

$$\begin{bmatrix} 0.4464735 & 0.2889718 & 0.0000000 & 0.0000000 & 0.5342276 \\ 0.0000000 & -0.5024268 & 0.6015901 & 0.0000000 & 0.4422450 \\ 0.0000000 & 0.0000000 & 1.1119174 & 1.363949 & 0.8825299 \\ 0.4140338 & 0.0000000 & 0.0000000 & 0.978484 & 0.0000000 \\ 0.4210915 & 0.0000000 & 0.0000000 & 1.090330 & 0.4532926 \end{bmatrix}$$

Figure 2 shows that the covariance structure of the variables becomes intertwined as the biological noise propagates over time and radiates towards other nodes. For instance, noise added to different nodes is independent at the beginning (top), but the residual of node A is affected by the residual of node A, B and E at previous time point (bottom left) and node A, B, C, D and E at two-time-point back (bottom right). Thus the residual of each node is affected by residuals of other nodes in the past.

4. Hypothesis testing based on comparative residuals

How would one know if the model has underfitting or overfitting? we propose a statistical strategy to answer this question. We are interested in if consistent behaviors in the observations can be captured by a model with statistical significance. According to the previous residual analysis, we design a hypothesis testing somehow similar to the F -test. Let $\hat{\mathbf{x}}[t]$ be the fitted value for $\mathbf{x}[t]$.

4.1. The Λ test statistic

Under the null hypothesis of no interaction, the prediction of the null model and the sum of residual squares for the null model are

$$\hat{\mathbf{x}}[t] = \frac{1}{T+1} \sum_{t=0}^T \mathbf{o}[t] \quad \text{and} \quad RSS_0 = \sum_{t=0}^T \|\mathbf{o}[t] - \hat{\mathbf{x}}[t]\|^2$$

Under the alternative hypothesis of interaction, the residuals of each node are accumulated and dependent on each other. By Eq.(4) and let $\mathbf{x}[t] \triangleq \hat{\mathbf{g}}[t]$, the sum of residual squares for the alternative model is

$$RSS_1 = \sum_{t=0}^T \|\mathbf{o}[t] - \hat{\mathbf{x}}[t]\|^2$$

Applying the relative residual reduction ratio principle in the F -test widely used in linear regression, we create the Λ test statistic

$$\Lambda = \frac{(RSS_0 - RSS_1)/p}{RSS_1/(N-p-1)} \quad (10)$$

where p is the maximum number of parents allowed for each node, controlling the complexity of the alternative model. In this definition, the test statistic will be large when reduction in residual by the interaction model is sufficiently greater than the model residual, modulus the model degrees of freedoms.

4.2. Simulation study of the distribution of Λ

As observations at different time points are generally not independent in a dynamical system, Λ does not necessarily follow an F -distribution. Our simulation study has also shown p -value from Kolmogorov-Smirnov test of Λ -distribution and F -distribution is less than 2.2×10^{-16} which rejecting that Λ has an F -distribution. Since the Λ distribution is a theoretical open problem, we can estimate the statistical significance or p -value by simulation.

We permute the observation data by time sequence for each node, respectively, to destroy all dynamical interactions, while keeping the marginal distribution of each node unchanged from the original. After DDS reconstruction on the permuted data sets, is obtained for each node of each data set by Eq. (10). The number of the permuted data sets is typically at least 400. The null distribution of Λ is thus obtained. It can be used to calculate the corresponding p_{Λ} -value.

5. A recommended procedure for evaluating the significance of DDS models

Based on the above analysis, we propose a procedure to evaluate the modeling outcome of each node in a DDS model. This node-by-node procedure is especially useful when various noises unevenly influence nodes in a network. Based on the understanding that if residuals are normally distributed, a model has reached the maximum performance, we recommend the following procedure for evaluating the modeling for each node:

1. Permute observations many times $> (400)$ and obtain model predictions;
2. Collect Λ for permuted data sets to form null distribution;
3. Reconstruction a DDS model from the original observations and obtain model prediction;
4. Calculate Λ , using the model from Step 3, for each node by Eq. (10) and the corresponding p -value; and
5. Analyze each node for its residual distribution.

According to the Λ statistic and the distribution of residuals obtained from the above procedure, four evaluation conclusions regarding each node can be made as shown in Table 1. And the normal distribution of the residual can be checked by the Shapiro test. This table gives rise to an iterative modeling procedure with increased model complexity.

6. DDS modeling of an artificial discrete dynamical system

An artificial DDS was utilized to demonstrate how to use the proposed model evaluation procedure.

The artificial DDS model includes 32 nodes, with $(32 \times 32 + 32) = 1,056$ coefficients in the first-order linear DDS, and $((32+1) \times 32 + 32) \times 32 + 32 = 17,952$ coefficients in the first-order quadratic DDS. The relationships between nodes are described as difference equations. Some nodes have only linear terms, the others have quadratic terms. Noise was added to simulated trajectories with a signal to noise ratio (SNR) of 20dB. We generated 2 trajectories, each containing 6 time points, starting from different initial states.

After the transition-based fitting Eq. (3) and trajectory-based fitting Eq. (5), we first reconstructed a linear DDS model from the noisy data. The ground-truth, observations and linear DDS predictions are shown in Fig. 3.

According to Eq. (10), we can calculate Λ and find its distribution for the linear DDS model based on the 2 trajectories, as well as the p_{Λ} -value and $p_{shapiro}$ -value from Shapiro test for normality of residual. From Table 2, the $p_{2,\Lambda}$ -value (> 0.05) indicates that the DDS model is insignificant. $p_{2,shapiro}$ -value (> 0.05) states that the residuals are already normally distributed, suggesting the modeling performance cannot be improved by increasing replica to reduce noise. This analysis concludes that the system dynamics are not adequately represented by the 2 trajectories and it is necessary to use additional initial states to probe the underlying system. This leads to increasing the number of trajectories from 2 to 16, starting from 16 different initial states. The statistics of modeling on the enlarged data set is also shown in Table 2. The *zero* $p_{16,\Lambda}$ -values indicate substantial improvement of the model fitting for the nodes shown.

After increasing the number of trajectories, the model is significant for the data set, Table 3. But the $p_{shapiro}$ -value ($\ll 0.05$) suggests non-normality of the residua and thus a bias could exist due to the insufficient complexity of the DDS model. This pushed the modeling to

proceed from the linear to the quadratic. After we applied the quadratic model to the 16 trajectory data set, we obtained further modeling improvement as shown in Table 4.

7. DDS modeling of the yeast pheromone pathway

We applied our model evaluation procedure on a real biochemical reaction network model for the yeast pheromone pathway [13] to illustrate the effectiveness of the statistical procedure. The biochemical reaction network is composed of about 35 nonlinear ODEs for the proteins and enzymes involved. Each protein constitutes a node in the network. The set of ODEs, encoded in the SBML, is indexed as BIOMD0000000032 from the BioModels database [15]. We used the MATLAB SBML Toolbox [16] to simulate noisy trajectories. We added biological noise with strength shown in Table 5 and measurement noise ($\sigma_m = 0.2$) to the original ODEs. Then we performed simulation of the ODE model starting from an initial state, and collected data between 0 and 4 minutes with 10 time steps in 3 trajectories, shown as circles in Fig. 5. After we reconstructed a linear DDS model from the simulated data from the yeast pheromone pathway, the residual for most nodes are not normally distributed. Then we moved up to use the first-order quadratic DDS model, Eq. (2).

We estimated the first three modules - the receptor activation, the G protein cycle, the complex formation involving Ste5 - in the yeast pheromone pathway, from the simulated trajectories (Fig. 5). The maximum number of parents is 4, but the product of two molecules in the quadratic model was counted as only one parent. Figure 5 shows the observations and the model predictions of some sample proteins made by both the transition- and the trajectory-based modeling.

Table 5 displays the statistics of the predicted DDS model and the p -values of the normality tests for residuals. Normality is rejected when $p\text{-value} < 0.05$. From the table, we have the following four modeling groups: First, large Λ with normal residuals, including Ste2, GaGTP, Gbc, Ste5, Ste11, Ste7, Fus3, and complexC; second, large Λ with non-normal residuals, including Ste2a, and complexB; third, small Λ with normal residuals, including complexA, and complexD; and forth, small Λ with non-normal residuals, including alpha, Gabc, and GaGDP. While some nodes, such as GaGTP, Gbc, Ste2a and complexB, have good fit, some other nodes are not satisfactorily modeled. There maybe at least two reasons for the latter: the network simulation is incomplete in covering various aspects of the system dynamics; and some nodes have highly linearly correlated dynamics even under no noise, so that the parent selection performance may be negatively influenced. The strategy to overcome the two disadvantages is to perform simulation from various initial states and apply variable clustering before DDS modeling.

8. Conclusion

Based on the analysis on the residual between model predictions and observations, a data-driven model evaluation procedure based on a statistical hypothesis testing was introduced. This procedure can evaluate a dynamical system model by answering the question whether an entire model or a part of it is supported significantly by the data. One can either further

improve the model by increasing its complexity, or decide that the data are too noisy to draw any conclusion and further data collection is entailed. Only after this iterative model development procedure, the model for an entire biological network or a partial network can be transferred to a wet lab for biological validation.

References

1. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*. Oct; 1997 278(5338):680–686. [PubMed: 9381177]
2. de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*. Jan; 2002 9(1):67–103. [PubMed: 11911796]
3. Kauffman S. Homeostasis and differentiation in random genetic control networks. *Nature*. Oct. 1969 224:177–178. [PubMed: 5343519]
4. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002; 18(2):261–274. [PubMed: 11847074]
5. Friedman N, Linial M, Nachman I, Pe'er D. Using bayesian networks to analyze expression data. *Journal of Computational Biology*. 2000; 7(3–4):601–620. [PubMed: 11108481]
6. Murphy, K.; Mian, S. Tech Rep. University of California at Berkeley; Berkeley, CA 94720: 1999. Modelling gene expression data using dynamic bayesian networks.
7. Zou M, Conzen SD. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*. 2005; 21(1):71–79. [PubMed: 15308537]
8. Lähdesmäki H, Hautaniemi S, Shmulevich I, Yli-Harja O. Relationships between probabilistic boolean networks and dynamic bayesian networks as models of gene regulatory networks. *Signal Processing*. Apr; 2006 86(4):814–834. [PubMed: 17415411]
9. Chen, T.; He, H.; Church, G. Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing*; 1999. p. 29–40.
10. van Someren, E.; Wessels, L.; Reinders, M.; Backer, E. Robust genetic network modeling by adding noisy data. *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*; June 2001;
11. Wang Y, Joshi T, Zhang X-S, Xu D, Chen L. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*. 2006; 22(19):2413–2420. [PubMed: 16864593]
12. Song, M.; Liu, L. LNCS. Vol. 4532. Springer; 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; p. 77–95.
13. Kofahl B, Klipp E. Modelling the dynamics of the yeast pheromone pathway. *Yeast*. 2004; 21(10): 831–850. [PubMed: 15300679]
14. Lourakis, M. levmar: Levenberg-marquardt nonlinear least squares algorithms in C/C++. Jul. 2004 <http://www.ics.forth.gr/~lourakis/levmar/>
15. Snoep, J.L.; Dharuri, H. Biomd0000000032-kofahl2004 pheromone. Sep. 2005 <http://www.ebi.ac.uk/biomodels/>
16. Keating SM, Bornstein BJ, Finney A, Hucka M. Sbmtoolbox: an sbml toolbox for matlab users. *Bioinformatics*. 2006; 22(10):1275–1277. [PubMed: 16574696]

Figure 1.
A toy example comparing predictions of models reconstructed using transition- versus trajectory-based fitting. The circles represent the observed trajectory

Figure 3.

Ground-truth, noisy observations, and model predictions of the reconstructed DDS models. Black: ground-truth, red: noisy observations, green: linear DDS predictions, blue: quadratic DDS predictions.

Figure 4.

A portion of the yeast pheromone pathway [13] used in validating our statistical model evaluation procedure.

Figure 5.

Observations and predictions made by transition- versus trajectory-based DDS modeling for several proteins in the yeast pheromone pathway.

Table 1

Model evaluation conclusions.

Residuals of each node	$p_A < 0.05$	$p_A = 0.05$
Normal distribution ($p_{shapiro} = 0.05$)	The model is significant.	The model is NOT significant, due to the noise. Collection of additional
Non-normal distribution ($p_{shapiro} < 0.05$)	The model is significant, but still can be improved	The model is NOT significant, due to the model solution or model complexity

Table 2
The model is not significant due to strong noise. A node in red stands for correct parent selection.

Node Name	<i>Regulator</i> ₂	<i>p</i> _{2, Λ} -value	<i>p</i> _{2, <i>shapio</i>} -value	<i>Regulator</i> ₁₆	<i>p</i> _{16, Λ} -value
Node0	Node0, Node2, Node26,	0.547	0.576	Node18, Node28,	0
Node2	Node5,	0.187	0.473	Node5,	0
Node11	Node0, Node11, Node28,	0.182	0.983	Node15, Node16,	0
Node13	Node9, Node19, Node27,	0.227	0.110	Node2, Node7,	0
Node20	Node8, Node15, Node21,	0.145	0.747	Node25,	0
Node31	Node11, Node15, Node17,	0.155	0.823	Node23,	0

Table 3

The linear DDS model is not significant due to insufficient model complexity. A node in red stands for correct parent selection.

Node Name	<i>Regulator_{linear}</i>	<i>P_{lin}</i> , Δ	<i>P_{lin,shapiro}</i>
Node4	Node7, Node19, Node21	0	0.002
Node6	Node16, Node24, Node26,	0	4.38e-07
Node15	Node21, Node30,	0	4.12qe-06
Node17	Node18, Node24, Node26,	0	4.45e-05
Node18	Node3, Node20, Node24,	0	0.002
Node23	Node6, Node10,	0	0.015
Node28	Node19,	0	0.0004
Node29	Node0, Node4, Node5,	0	4.12e-04

Table 4

The significant quadratic DDS model. A node in red stands for correct parent selection.

Node Name	<i>Regulator_{quadratic}</i>	<i>P_{qua}</i> , Δ	<i>P_{qua,shapiro}</i>
Node4	Node19, Node21*Node14	0	0.026
Node6	Node26*Node24,	0	0.090
Node15	Node30, Node21*Node16,	0	0.095
Node17	Node16, Node18, Node26*Node24,	0	0.023
Node18	Node20, Node18*Node0, Node26*Node3,	0	0.008
Node23	Node5, Node10*Node6, Node23*Node23,	0	0.443
Node28	Node23*Node19, Node28*Node20, Node30*Node14,	0	0.01
Node29	Node11*Node3, Node23*Node6, Node29*Node29,	0	0.019

Table 5

Statistics of the reconstructed DDS model from simulated noisy data generated in the yeast pheromone pathway.

Node Name	σb	Λ	p -value	Regulator
alpha	0.2	391	0.044	Ste2,Ste2.alpha,complexB,complexB.alpha
Ste2	5	4718.5	0.339	Ste2.Ste2,complexA
Ste2a	0.5	8019.2	0.006	alpha.alpha,Ste2a,Ste7.Ste2,Ste7.Ste2a
Gabc	10	123.5	0.028	Ste2.alpha,GaGTP.Gabc,complexA.alpha,complexC.Gabc
GaGTP	1	65949.5	0.11	Ste2.alpha,GaGTP.alpha,Gbc.Ste2,complexB.Gbc
Gbc	0.4	9240	0.232	Gbc.Ste2,Ste11.Ste2a,complexB.alpha,complexB.Gbc
GaGDP	0.6	25.3	0.001	alpha,GaGDP.GaGDP,Ste7.GaGDP,complexA.Gabc
Ste5	0.12	19486	0.071	Gbc.Ste2a,Ste5.Gbc,complexB.Ste5,complexC.Gabc
Ste11	0.03	10583.8	0.077	GaGDP.alpha,Ste5.GaGTP,complexB.Gbc,complexC.alpha
Ste7	0.29	1849.5	0.675	GaGTP,GaGTP.Ste2,Ste7.Ste7,complexC.complexB
Fus3	0.25	47402.1	0.916	Ste2a.alpha,complexB.Gbc,complexB.complexA,complexC.Fus3
complexA	0.8	320.1	0.905	GaGTP.GaGTP,Ste5.Ste2a,Ste7.Ste11,complexC.Ste7
complexB	0.05	8705.9	0.047	Gbc.alpha,Gbc.GaGTP,complexB.Ste5,complexC.Gabc
complexC	0.32	25645.2	1	Gabc.Gabc,Gbc.GaGTP,complexB.Ste2,complexC.Gbc
complexD	0.6	1.8	0.239	GaGDP,GaGDP.Ste2,complexA.alpha,complexD.Fus3
Entire Network		269.1		