Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 82702, 11 pages
doi:10.1155/2007/82702

Research Article

Comparison of Gene Regulatory Networks via Steady-State Trajectories

Marcel Brun,[1] Seungchan Kim,[1,2] Woonjung Choi,[3] and Edward R. Dougherty[1,4,5]

[1] *Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA*
[2] *School of Computing and Informatics, Ira A. Fulton School of Engineering, Arizona State University, Tempe, AZ 85287, USA*
[3] *Department of Mathematics and Statistics, College of Liberal Arts and Sciences, Arizona State University, Tempe, AZ 85287, USA*
[4] *Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA*
[5] *Cancer Genomics Laboratory, Department of Pathology, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA*

Received 31 July 2006; Accepted 24 February 2007

Recommended by Ahmed H. Tewfik

The modeling of genetic regulatory networks is becoming increasingly widespread in the study of biological systems. In the abstract, one would prefer quantitatively comprehensive models, such as a differential-equation model, to coarse models; however, in practice, detailed models require more accurate measurements for inference and more computational power to analyze than coarse-scale models. It is crucial to address the issue of model complexity in the framework of a basic scientific paradigm: the model should be of minimal complexity to provide the necessary predictive power. Addressing this issue requires a metric by which to compare networks. This paper proposes the use of a classical measure of difference between amplitude distributions for periodic signals to compare two networks according to the differences of their trajectories in the steady state. The metric is applicable to networks with both continuous and discrete values for both time and state, and it possesses the critical property that it allows the comparison of networks of different natures. We demonstrate application of the metric by comparing a continuous-valued reference network against simplified versions obtained via quantization.

1. INTRODUCTION

The modeling of genetic regulatory networks (GRNs) is becoming increasingly widespread for gaining insight into the underlying processes of living systems. The computational biology literature abounds in various network modeling approaches, all of which have particular goals, along with their strengths and weaknesses [1, 2]. They may be deterministic or stochastic. Network models have been studied to gain insight into various cellular properties, such as cellular state dynamics and transcriptional regulation [3–8], and to derive intervention strategies based on state-space dynamics [9, 10].

Complexity is a critical issue in the synthesis, analysis, and application of GRNs. In principle, one would prefer the construction and analysis of a quantitatively comprehensive model such as a differential equation-based model to a coarsely quantized discrete model; however, in practice, the situation does not always suffice to support such a model. Quantitatively detailed (fine-scale) models require significantly more complex mathematics and computational power for analysis and more accurate measurements for inference than coarse-scale models. The network complexity issue has similarities with the issue of classifier complexity [11]. One must decide whether to use a fine-scale or coarse-scale model [12]. The issue should be addressed in the framework of the standard engineering paradigm: the model should be of minimal complexity to solve the problem at hand.

To quantify network approximation and reduction, one would like a metric to compare networks. For instance, it may be beneficial for computational or inferential purposes to approximate a system by a discrete model instead of a continuous model. The goodness of the approximation is measured by a metric and the precise formulation of the properties will depend on the chosen metric.

Comparison of GRN models needs to be based on salient aspects of the models. One study used the L_1 norm between the steady-state distributions of different networks in the context of the reduction of probabilistic Boolean networks

[13]. Another study compared networks based on their topologies, that is, connectivity graphs [14]. This method suffers from the fact that networks with the same topology may possess very different dynamic behaviors. A third study involved a comprehensive comparison of continuous models based on their inferential power, prediction power, robustness, and consistency in the framework of simulations, where a network is used to generate gene expression data, which is then used to reconstruct the network [15]. A key drawback of most approaches is that the comparison is applicable only to networks with similar representations; it is difficult to compare networks of different natures, for instance, a differential-equation model to a Boolean model. A salient property of the metric proposed in this study is that it can compare networks of different natures in both value and time.

We propose a metric to compare deterministic GRNs via their steady-state behaviors. This is a reasonable approach because in the absence of external intervention, a cell operates mainly in its steady state, which characterizes its phenotype, that is, cell cycle, disease, cell differentiation, and so forth. [16–19]. A cell's phenotypic status is maintained through a variety of regulatory mechanisms. Disruption of this tight steady-state regulation may lead to an abnormal cellular status, for example, cancer. Studying steady-state behavior of a cellular system and its disruption can provide significant insight into cellular regulatory mechanisms underlying disease development.

We first introduce a metric to compare GRNs based on their steady-state behaviors, discuss its characteristics, and treat the empirical estimation of the metric. Then we provide a detailed application to quantization utilizing the mathematical framework of reference and projected networks. We close with some remarks on the efficacy of the proposed metric.

2. METRIC BETWEEN NETWORKS

In this section, we construct the distance metric between networks using a bottom-up approach. Following a description of how trajectories are decomposed into their transient and steady-state parts, we define a metric between two periodic or constant functions and then extend this definition to a more general family of functions that can be decomposed between transient and steady-state parts.

2.1. Steady-state trajectory

Given the understanding that biological networks exhibit steady-state behavior, we confine ourselves to networks exhibiting steady-state behavior. Moreover, since a cell uses nutrients such as amino acids and nucleotides in cytoplasm to synthesize various molecular components, that is, RNAs and proteins [18], and since there are only limited supplies of nutrients available, the amount of molecules present in a cell is bounded. Thus, the existence of steady-state behavior implies that each individual gene trajectory can be modeled as a bounded function $f(t)$ that can be decomposed into a transient trajectory plus a steady-state trajectory:

$$f(t) = f_{\text{tran}}(t) + f_{\text{ss}}(t), \tag{1}$$

where $\lim_{t\to\infty} f_{\text{tran}}(t) = 0$ and $f_{\text{ss}}(t)$ is either a periodic function or a constant function.

The limit condition on the transient part of the trajectory indicates that for large values of t, the trajectory is very close to its steady-state part. This can be expressed in the following manner: for any $\epsilon > 0$, there exists a time t_{ss} such that $|f(t) - f_{\text{ss}}(t)| < \epsilon$ for $t > t_{ss}$. This property is useful to identify $f_{\text{ss}}(t)$ from simulated data by finding an instant t_{ss} such that $f(t)$ is almost periodical or constant for $t > t_{ss}$.

A deterministic gene regulatory network, whether it is represented by a set of differential equations or state transition equations, produces different dynamic behaviors, depending on the starting point. If ψ is a network with N genes and $\mathbf{x}_0$ is an initial state, then its *trajectory*,

$$\mathbf{f}_{(\psi,\mathbf{x}_0)}(t) = \left\{f^{(1)}_{(\psi,\mathbf{x}_0)}(t), \ldots, f^{(N)}_{(\psi,\mathbf{x}_0)}(t)\right\}, \tag{2}$$

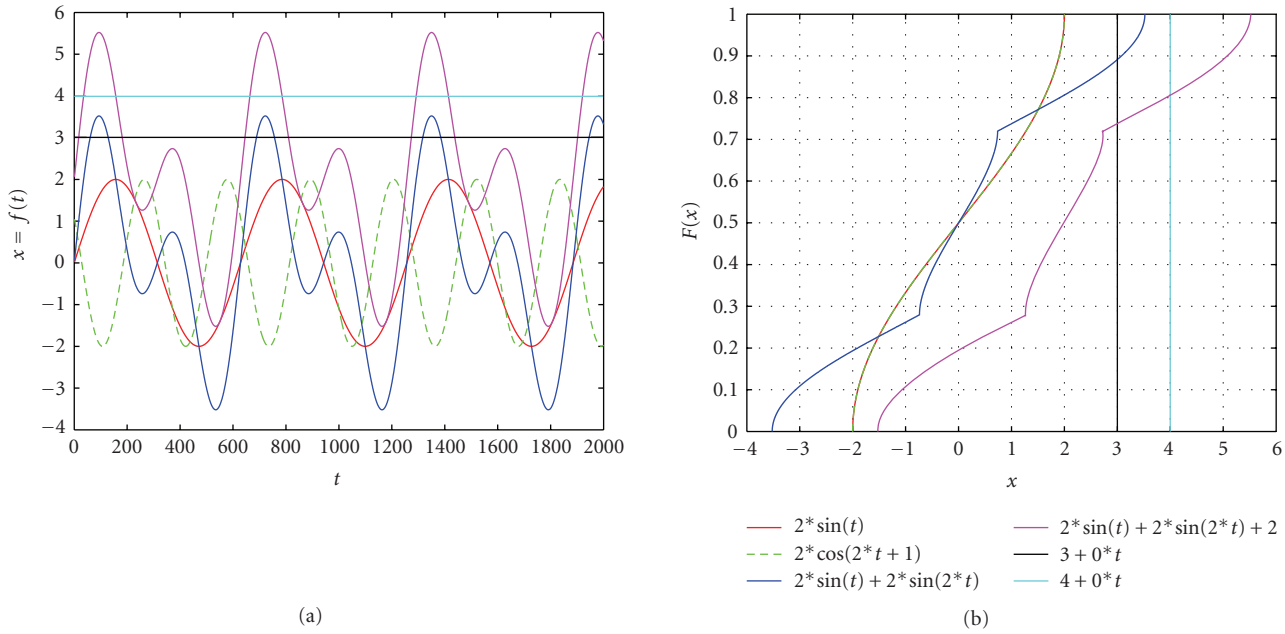
where $f^{(i)}_{(\psi,\mathbf{x}_0)}(t)$ is a trajectory for an individual gene (denoted by $f^{(i)}(t)$ or $f(t)$ where there is no ambiguity) generated by the dynamic behavior of the network ψ when starting at $\mathbf{x}_0$. For a differential-equation model, the trajectory $\mathbf{f}_{(\psi,\mathbf{x}_0)}(t)$ can be obtained as a solution of a system of differential equations; for a discrete model, it can be obtained by iterating the system's transition equations. Trajectories may be continuous-time functions or discrete-time functions, depending on the model.

The decomposition of (1) applies to $\mathbf{f}_{(\psi,\mathbf{x}_0)}(t)$ via its application to the individual trajectories $f^{(i)}_{(\psi,\mathbf{x}_0)}(t)$. In the case of discrete-valued networks (with bounded values), the system must enter an attractor cycle or an attractor state at some time point t_{ss}. In the first case $\mathbf{f}_{(\psi,\mathbf{x}_0),\text{ss}}(t)$ is periodical, and in the second case it is constant. In both cases, $\mathbf{f}_{(\psi,\mathbf{x}_0),\text{tran}}(t) = \mathbf{0}$ for $t \geq t_{ss}$.

2.2. Distance based on the amplitude cumulative distribution

Different metrics have been proposed to compare two real-valued trajectories $f(t)$ and $g(t)$, including the correlation $\langle f, g\rangle$, the cross-correlation $\Gamma_{f,g}(\tau)$, the cross-spectral density $p_{f,g}(\omega)$, the difference between their amplitude cumulative distributions $F(x) = p_f(x)$ and $G(x) = p_g(x)$, and the difference between their statistical moments [20]. Each has its benefits and drawbacks depending on one's purpose. In this paper, we propose using the difference between the amplitude cumulative distributions of the steady-state trajectories.

Let $f_{\text{ss}}(t)$ and $g_{\text{ss}}(t)$ be two measurable functions that are either periodical or constant, representing the steady-state parts of two functions, $f(t)$ and $g(t)$, respectively. Our goal is to define a metric (distance) between them by using the



(a) (b)

Figure 1: Example of (a) periodical and constant functions $f(t)$ and (b) their amplitude cumulative distributions $F(x)$.

amplitude cumulative distribution (ACD), which measures the probability density of a function [20].

If $f_{ss}(t)$ is periodic with period $t_p > 0$, its cumulative densityfunction $F(x)$ over $\mathbb{R}$ is defined by

$$F(x) = \lambda\left(\frac{\mathcal{M}(x)}{t_p}\right), \tag{3}$$

where $\lambda(A)$ isthe Lebesgue measure of the set A and

$$\mathcal{M}(x) = \{t_s \le t < t_e \mid f_{ss}(t) \le x\}, \tag{4}$$

where $t_e = t_s + t_p$, for any point t_s.

If f_{ss} is constant, given by $f_{ss}(t) = a$ for any t, then we define $F(x)$ as a unit step function located at $x = a$. Figure 1 shows an example of some periodical functions and their amplitude cumulative distributions.

Given two steady-state trajectories, $f_{ss}(t)$ and $g_{ss}(t)$, and their respective amplitude cumulative distributions, $F(x)$ and $G(x)$, we define the distance between f_{ss} and g_{ss} as the distance between the distributions

$$d_{ss}(f_{ss}, g_{ss}) = \|F - G\| \tag{5}$$

for some suitable norm $\|\cdot\|$. Examples of norms include L_∞, defined by the supremum of their differences,

$$d_{L_\infty}(f,g) = \sup_{0\le x\le\infty} |F(x) - G(x)|, \tag{6}$$

and L_1 defined by the area of the absolute value of their difference,

$$d_{L_1}(f,g) = \int_{0\le x<\infty} |F(x) - G(x)|\, dx. \tag{7}$$

In both cases, we apply the biological constraint that the amplitudes are nonnegative.

The L_1 norm is well suited to the steady-state behavior because in the case of constant functions $f(t) = a$ and $g(t) = b$, their distributions are unit steps functions at $x = a$ and $x = b$, respectively, so that $d_{L_1}(f,g) = |a - b|$, the distance, in amplitude, between the two functions. Hence, we can interpret the distance $d_{L_1}(f,g)$ as an extension of the distance, in amplitude, between two constant signals, to the general case of periodic functions, taking into consideration the differences in their shapes.

2.3. Network metric

Once a distance between their steady-state trajectories is defined, we can extend this distance to two trajectories $f(t)$ and $g(t)$ by

$$d_{tr}(f,g) = d_{ss}(f_{ss}, g_{ss}), \tag{8}$$

where d_{ss} is defined by (5).

The next step is to define the distance between two multivariate trajectories $\mathbf{f}(t)$ and $\mathbf{g}(t)$ by

$$d_{tr}(\mathbf{f},\mathbf{g}) = \frac{1}{N}\sum_{i=1}^{N} d_{tr}(f^{(i)}, g^{(i)}), \tag{9}$$

where $f^{(i)}(t)$ and $g^{(i)}(t)$ are the component trajectories of $\mathbf{f}(t)$ and $\mathbf{g}(t)$, respectively. Owing to the manner in which a norm is used to define d_{ss}, in conjunction with the manner in which d_{tr} is constructed from d_{ss}, the triangle inequality

$$d_{tr}(\mathbf{f},\mathbf{h}) \le d_{tr}(\mathbf{f},\mathbf{g}) + d_{tr}(\mathbf{g},\mathbf{h}) \tag{10}$$

holds, and d_{tr} is a metric.

The last step is to define the *metric between two networks* as the expected distance between the trajectories over all possible initial states. For networks ψ_1 and ψ_2, we define

$$d(\psi_1, \psi_2) = E_S[d_{\text{tr}}(\mathbf{f}_{(\psi_1,\mathbf{x}_0)}, \mathbf{f}_{(\psi_2,\mathbf{x}_0)})], \tag{11}$$

where the expectation is taken with respect to the space S of initial states.

The use of a metric, in particular, the triangle inequality, is essential for the problem of estimating complex networks by using simpler models. This is akin to the pattern recognition problem of estimating a complex classifier via a constrained classifier to mitigate the data requirement. In this situation, there is a complex model that represents a broad family of networks and a simpler model that represents a smaller class of networks. Given a reference network from the complex model and a sampled trajectory from it, we want to estimate the optimal constrained network. We can identify the optimal constrained network, that is, projected network, as the one that best approximates the complex one, and the goal of the inference process should be to obtain a network close to the optimal constrained network. Let ψ be a reference network (e.g., a continuous-valued ODE-based network), let $P(\psi)$ be the optimal constrained network (e.g., a discrete-valued network), and let $\overline{\omega}$ be an estimator of $P(\psi)$ estimated from data sampled from ψ. Then

$$d(\overline{\omega}, \psi) \le d(\overline{\omega}, P(\psi)) + d(P(\psi), \psi), \tag{12}$$

where the following distances have natural interpretations:

(i) $d(\overline{\omega}, \psi)$ is the *overall distance* and quantifies the approximation of the reference network by the estimated optimal constrained network;

(ii) $d(\overline{\omega}, P(\psi))$ is the *estimation distance* for the constrained network and quantifies the inference of the optimal constrained network;

(iii) $d(P(\psi), \psi)$ is the *projection distance* and quantifies how well the optimal constrained network approximates the reference network.

This structure is analogous to the classical constrained regression problem, where constraints are used to facilitate better inference via reduction of the estimation error (so long as this reduction exceeds the projection error) [11]. In the case of networks, the constraint problem becomes one of finding a projection mapping for models representing biological processes for which the loss defined by $d(P(\psi), \psi)$ may be maintained within manageable bounds so that with good inference techniques, the estimation error defined by $d(\overline{\omega}, P(\psi))$ will be minimized.

2.4. Estimation of the amplitude cumulative distribution

The amplitude cumulative distribution of a trajectory can be estimated by simulating the trajectory and then estimating the ACD from the trajectory. Assuming that the steady-state trajectory $f_{\text{ss}}(t)$ is periodic with period t_p, we can analyze $f_{\text{ss}}(t)$ between two points, t_s and $t_e = t_s + t_p$. For a continuous function $f_{\text{ss}}(t)$, we assume that any amplitude value x is visited only a finite number of times by $f_{\text{ss}}(t)$ in a period $t_s \le t < t_e$. In accordance with (3), we define the cumulative distribution

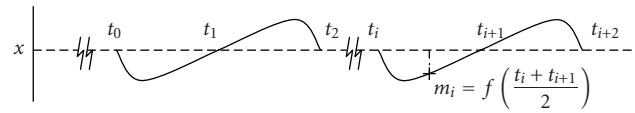


Figure 2: Example of determination of values m_i.

$$F(x) = \frac{\lambda(\{t_s \le t \le t_e \mid f_{\text{ss}}(t) \le x\})}{t_p}. \tag{13}$$

To calculate $F(x)$ from a sampled trajectory, for each value x, let S_x be the set of points where $f_{\text{ss}}(t) = x$:

$$S_x = \{t_s \le t \le t_e \mid f_{\text{ss}}(t) = x\} \cup \{t_s, t_e\}. \tag{14}$$

The set S_x is finite. Let $n = |S_x|$ denote the number of elements $t_0, \ldots, t_{n-1}$. These can be sorted so that $t_s = t_0 < t_1 < t_2 < \cdots < t_{n-1} = t_e$. Now we define the set m_i, $i = 0, \ldots, n-2$, of intermediate values between two consecutive points where $f_{\text{ss}}(t)$ crosses x (see Figure 2) by

$$m_i = f_{\text{ss}}\left(\frac{t_i + t_{i+1}}{2}\right). \tag{15}$$

Let I_x be a set of the indices of points t_i such that the function $f(t)$ is below x in the interval $[t_i, t_{i+1}]$,

$$I_x = \{0 \le i \le n-2 \mid m_i \le x\}. \tag{16}$$

Finally, the cumulative distribution $F(x)$, defined by the measure of the set $\{t_s \le t \le t_e \mid f(t) \le x\}$, can be computed as the sum of the lengths of the intervals where $f(t) \le x$:

$$F(x) = \frac{\sum_{i \in I_x} (t_{i+1} - t_i)}{t_p}. \tag{17}$$

The estimation of $F(x)$ from a finite set $\{a_1, \ldots, a_m\}$ representing the function $f(t)$ at points $t_1, \ldots, t_m$ reduces to estimating the values in (17):

$$\widetilde{F}(x) = \frac{|\{1 \le i \le m \mid a_i \le x\}|}{m} \tag{18}$$

at the points a_i, $i = 1, \ldots, m$.

In the case of computing the distance between two functions $f(t)$ and $g(t)$, where the only information available consists of two samples, $\{a_1, \ldots, a_m\}$ and $\{b_1, \ldots, b_r\}$, for f and g, respectively, both cumulative distributions $\widetilde{F}(x)$ and $\widetilde{G}(x)$ need only be defined at the points in the set

$$S = \{a_1, \ldots, a_m\} \cup \{b_1, \ldots, b_r\}. \tag{19}$$

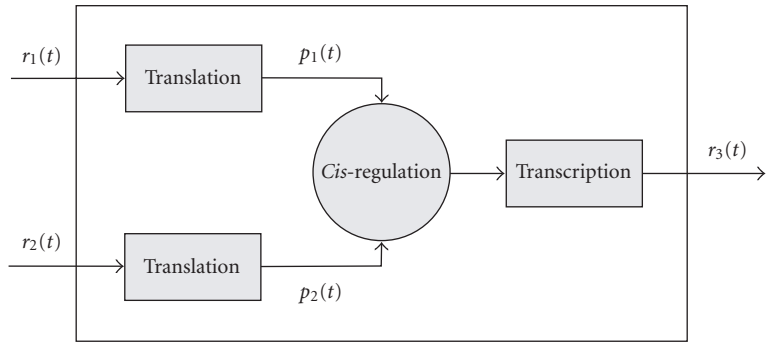


FIGURE 3: Block diagram of a model for transcriptional regulation.

In this case, if we sort the set S so that $0 = s_0 < s_2 < \cdots < s_k = T$ (with T being the upper limit for the amplitude values, and $k \leq r + m$), then (6) can be approximated by

$$\widetilde{d}_{L_\infty}(f,g) = \max_{0\leq i\leq k} |\widetilde{F}(s_i) - \widetilde{G}(s_i)| \tag{20}$$

and (7) can be approximated by

$$\widetilde{d}_{L_1}(f,g) = \sum_{0\leq i\leq k-1} (s_{i+1} - s_i)\,|\widetilde{F}(s_i) - \widetilde{G}(s_i)|. \tag{21}$$

3. APPLICATION TO QUANTIZATION

To illustrate application of the network metric, we will analyze how different degrees of quantization affect model accuracy. Quantization is an important issue in network modeling because it is imperative to balance the desire for fine description against the need for reduced complexity for both inference and computation. Since it is difficult, if not impossible, to directly evaluate the goodness of a model against a real biological system, we will study the problem using a standard engineering approach. First, an *in numero* reference network model or *system* is formulated. Then, a second network model with a different level of abstraction is introduced to approximate the reference system. The objective is to investigate how different levels of abstraction, quantization levels in this study, impact the accuracy of the model prediction. The first model is called the *reference model*. From it, reference networks will be instantiated with appropriate sets of model parameters. The model will be continuous-valued to approximate the reference system at its fullest closeness. The second model is called a *projected model*, and projected networks will be instantiated from it. This model will be a discrete-valued model at a given different level of quantization.

The ability of a projected network, an instance of the projected model, to approximate a reference network, an instance of the reference model, can be evaluated by comparing the trajectories generated from each network with different initial states and computing the distances between the networks as given by (11).

3.1. Reference model

The origin of our reference model is a differential-equation model that quantitatively represents transcription, translation, *cis*-regulation and chemical reactions [7, 15, 21]. Specifically, we consider a differential-equation model that approximates the process of transcription and translation for a set of genes and their associated proteins (as illustrated in Figure 3) [7].The model comprises the following differential equations:

$$\begin{aligned}
\frac{dp_i(t)}{dt} &= \lambda_i r_i(t - \tau_{p,i}) - \gamma_i p_i(t), \quad i \in \mathcal{G},\\
\frac{dr_i(t)}{dt} &= \kappa_i c_i(t - \tau_{r,i}) - \beta_i r_i(t), \quad i \in \mathcal{G},\\
c_i(t) &= \phi_i[p_j(t - \tau_{c,j}),\ j \in \mathcal{R}_i], \quad i \in \mathcal{G},
\end{aligned} \tag{22}$$

where r_i and p_i are the concentrations of mRNA and proteins induced by gene i, respectively, $c_i(t)$ is the fraction of DNA fragments committed to transcription of gene i, κ_i is the transcription rate of gene i, and $\tau_{p,i}$, $\tau_{r,i}$, and $\tau_{c,i}$ are the time delays for each process to start when the conditions are given. The most general form for the function ϕ_i is a real-valued (usually nonlinear) function with domain in $\mathbb{R}^{|\mathcal{R}_i|}$ and range in $\mathbb{R}$, $\phi_i : \mathbb{R}^{|\mathcal{R}_i|} \to \mathbb{R}$. The functions are defined by the equations

$$\begin{aligned}
\phi_i[p_j, j \in \mathcal{R}_i] = &\left[1 - \prod_{j\in\mathcal{R}_i^+} \rho(p_j, S_{ij}, \theta_{ij})\right]\\
&\times \prod_{j\in\mathcal{R}_i^-} \rho(p_j, S_{ij}, \theta_{ij}),\\
\rho(p, S, \theta) = &\frac{1}{(1+\theta p)^S},
\end{aligned} \tag{23}$$

where the parameters θ are the *affinity constants* and the parameters S_{ij} are the distinct sites for gene i where promoter j can bind. The functions depend on the discrete parameter S_{ij}, the number of binding sites for protein j on gene i, and θ_{ij}, the affinity constant between gene i and protein j.

A discrete-time model results from the preceding continuous-time model by discretizing the time t on intervals $n\delta t$, and the assumption that the fraction of DNA

Table 1: Parameter values used in simulations.

Parameter	Value	Parameter	Value
Affinity constant	$\theta = 10^8\ \mathrm{M}^{-1}$	Number of binding sites	$S = 1$
mRNA and protein half-life	$\rho = 1200\ \mathrm{s}$ $\pi = 3600\ \mathrm{s}$	Transcription rates	$\kappa_1 = 0.001\ \mathrm{pMs}^{-1}$ $\kappa_2 = \kappa_3 = \kappa_4 = 0.05\ \mathrm{pMs}^{-1}$
Translation rate	$\lambda = 0.20\ \mathrm{s}^{-1}$	Time delays	$\tau_r = 2000\ \mathrm{s}$ $\tau_c = 200\ \mathrm{s}$ $\tau_p = 2400\ \mathrm{s}$

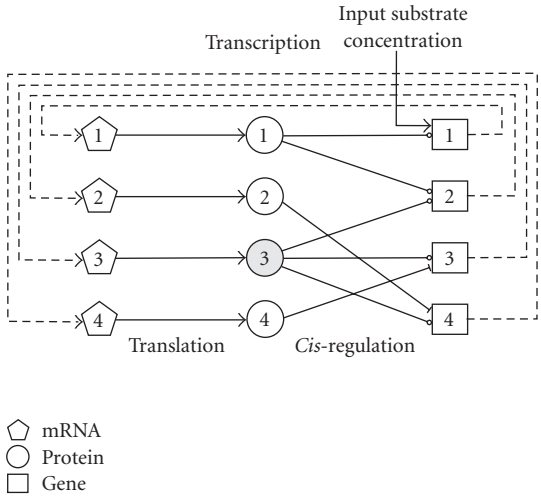


Figure 4: Example of a tRS of a hypothetical metabolic pathway that consists of four genes. In this figure, —∘ denotes an activator, whereas, ⊣ denotes a repressor.

fragments committed to transcription and concentration of mRNA remains constant in the time interval $[t - \delta t, t)$ [7]. In place of the differential equations for r_i, p_i, and c_i, at time $t = n\delta t$, we have the equations

$$\begin{aligned} r_i(n) &= e^{-\beta_i \delta t} r_i(n-1) + \kappa_i s(\beta_i, \delta t) c_i(n - n_{r,i} - 1), \\ p_i(n) &= e^{-\gamma_i \delta t} p_i(n-1) + \lambda_i s(\lambda_i, \delta t) r_i(n - n_{p,i} - 1), \\ c_i(n) &= \phi_i\big[p_j(n - n_{c,j}),\ j \in \mathcal{R}_i\big], \quad i \in \mathcal{G}, \end{aligned} \tag{24}$$

where $n_{r,i} = \tau_{r,i}/\delta t$, $n_{p,i} = \tau_{p,i}/\delta t$, $n_{c,j} = \tau_{c,j}/\delta t$, and

$$s(x, y) = \frac{1 - e^{-xy}}{x}. \tag{25}$$

This model, which will serve as our reference model, is called a (discrete) *transcriptional regulatory system* (*tRS*).

We generate networks using this model and a fixed set $\boldsymbol{\theta}$ of parameters. We call these networks *reference networks*. A reference network is identified by its set $\boldsymbol{\theta}$ of parameters,

$$\begin{aligned} \boldsymbol{\theta} = (&\alpha_1, \beta_1, \lambda_1, \gamma_1, \kappa_1, \tau_{p,1}, \tau_{r,1}, \tau_{c,1}, \phi_1, \mathcal{R}_1, \ldots, \alpha_N, \\ &\beta_N, \lambda_N, \gamma_N, \kappa_N, \tau_{p,N}, \tau_{r,N}, \tau_{c,N}, \phi_N, \mathcal{R}_N). \end{aligned} \tag{26}$$

3.2. Projected model

The next step is to reduce the reference network model to a projected network model. This is accomplished by applying constraints in the reference model. The application of constraints modifies the original model, thereby obtaining a simpler one. We focus on quantization of the gene expression levels (which are continuous-valued in the reference model) via uniform quantization, which is defined by a finite or denumerable set $\mathcal{L}$ of intervals, $L_1 = [0, \Delta_x)$, $L_2 = [\Delta_x, 2\Delta_x), \ldots$, $L_i = [(i-1)\Delta_x, i\Delta_x), \ldots$, and a mapping $\Pi_{\mathcal{L}} : \mathbb{R} \to \mathbb{R}$ such that $\Pi(x) = a_i$ for some collection of points $a_i \in L_i$.

The equations for r_i, p_i, and c_i (24) are replaced by

$$\overline{r}_i(n) = \Pi\big(e^{-\beta_i \delta t}\overline{r}_i(n-1) + \kappa_i s(\beta_i, \delta t)\overline{c}_i(n - n_{r,i} - 1)\big), \tag{27}$$

$$\overline{p}_i(n) = \Pi\big(e^{-\gamma_i \delta t}\overline{p}_i(n-1) + \lambda_i s(\lambda_i, \delta t)\overline{r}_i(n - n_{p,i} - 1)\big), \tag{28}$$

$$\overline{c}_i(n) = \phi_i\big[\overline{p}_j(n - n_{c,j}),\ j \in \mathcal{R}_i\big], \quad i \in \mathcal{G}. \tag{29}$$

Issues to be investigated include (1) how different quantization techniques (specification of the partition $\mathcal{L}$) affect the quality of the model; (2) which quantization technique (mapping Π) is the best for the model; and (3) the similarity of the attractors of the dynamical system defined by (27) and (28) to the steady state of the original system, as a function of Δ_x. We consider the first issue.

3.3. A hypothetical metabolic pathway

To illustrate the proposed metric in the framework of the reference and projected models, we compare two networks based on a hypothetical metabolic pathway. We first briefly describe the hypothetical metabolic pathway with necessary biochemical parameters to set up a reference system. Then, the simulation study shows the impacts of various quantization levels in both time and trajectory based on the proposed metric.

We consider a gene regulatory network consisting of four genes. A graphical representation of the system is depicted in Figure 4, where —∘ denotes an activator and ⊣ denotes a repressor. We assume that the GRN regulates a hypothetical pathway, which metabolizes an input substrate to an output product. This is done by means of enzymes whose transcriptional control is regulated by the protein produced from gene 3. Moreover, we assume that the effect of a higher input substrate concentration is to increase the transcription rate κ_1,

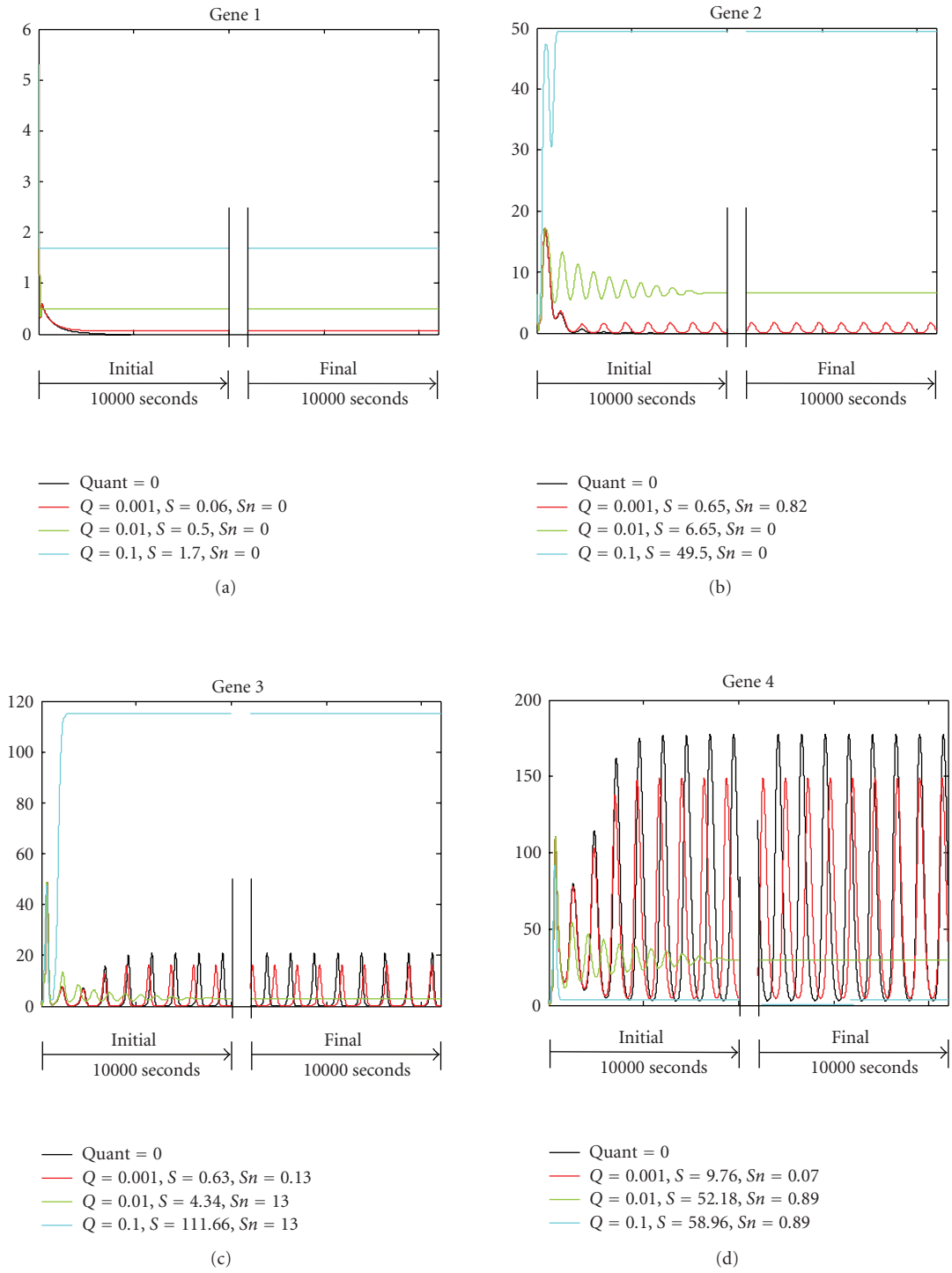


FIGURE 5: Example of trajectories from the first simulation of 4-gene network. Each figure shows the trajectory for one of the four genes, for several values of the level quantization Δ_x, represented by the lines $Q = 0$, $Q = 0.001$, $Q = 0.01$ and $Q = 0.1$ ($Q = 0$ represents the original network without quantization). The values S displayed in the graphs shows the distance computed between the trajectory and the one with Q=0. The vertical axis shows the concentration levels x in pM. The horizontal axis shows the time t in seconds.

whereas the effect of a lower substrate concentration is to reduce κ_1. Unless otherwise specified, the parameters are assumed to be gene-independent. These parameters are summarized in Table 1.

We assume that each *cis*-regulator is controlled by one module with four binding sites, and set $S = 4$, $\theta = 10^8\,\mathrm{M}^{-1}$, $\kappa_2 = \kappa_3 = \kappa_4 = 0.05\,\mathrm{pMs}^{-1}$, and $\lambda = 0.05\,\mathrm{s}^{-1}$. The value of the affinity constant θ corresponds to a binding free energy

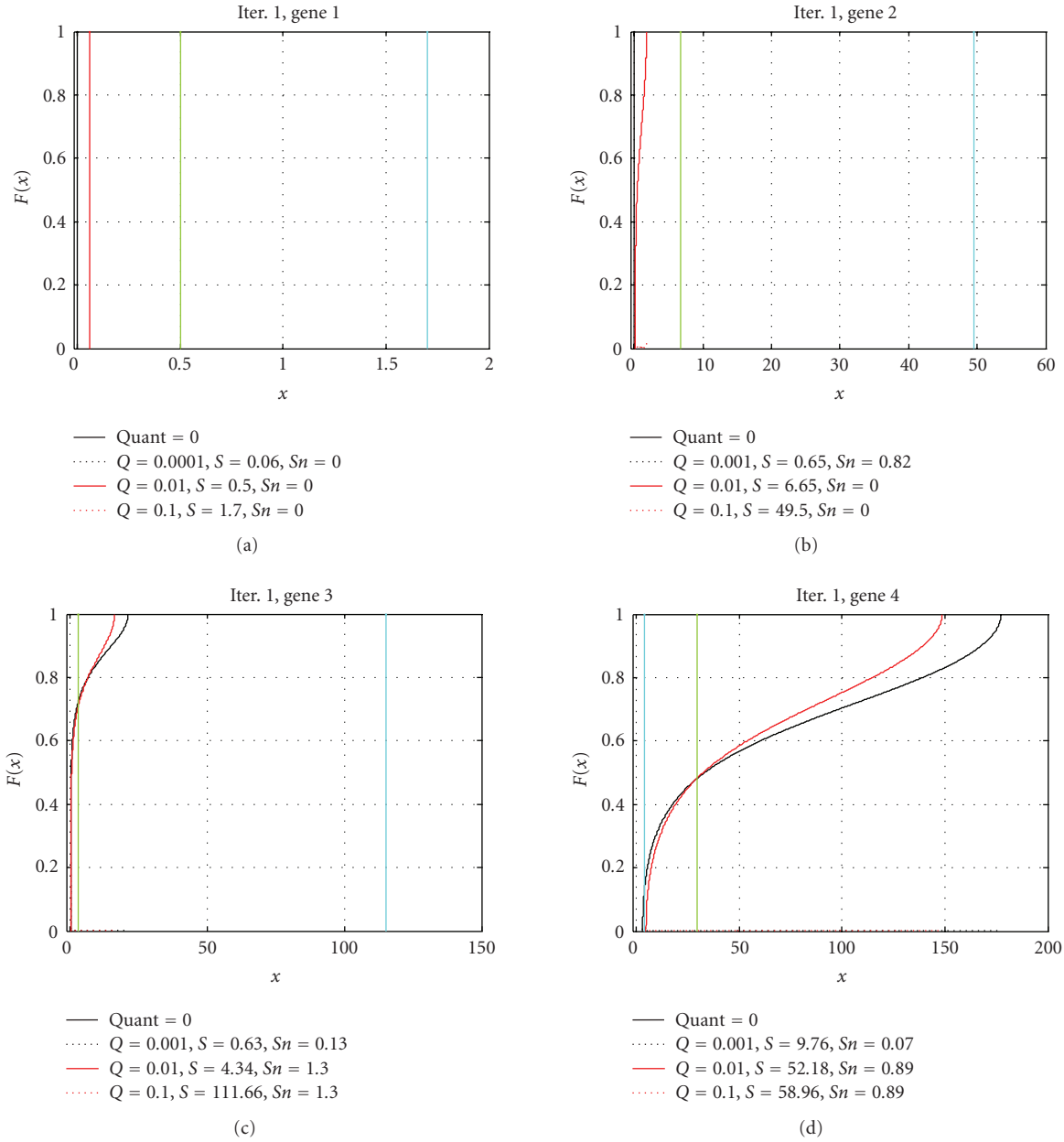


FIGURE 6: Example of estimated cumulative density function (CDF) from the first simulation of 4-gene network, computed from the trajectories in Figure 5. Each figure shows the CDF for one of the four genes, for several values of the level quantization Δ_x, represented by the lines $Q = 0$, $Q = 0.001$, $Q = 0.01$, and $Q = 0.1$ ($Q = 0$ represents the original network without quantization). The value S displayed in the graphs show the distance computed between the trajectory and the one with $Q = 0$. The vertical axis shows the cumulative distribution $F(x)$. The horizontal axis shows the concentration levels x in pM.

of $\Delta U = -11.35$ kcal/mol at temperature $T = 310.15°$K (or 37°C). The values of the transcription rates κ_2, κ_3, and κ_4 correspond to transcriptional machinery that, on the average, produces one mRNA molecule every 8 seconds. This value turns out to be typical for yeast cells [22]. We also assume that on the average, the volume of each cell in $\mathcal{C}$ equals 4 pL [18]. The translation rate λ is taken to be 10-fold larger than the rate of 0.3/minute for translation initiation observed *in vitro* using a semipurified rabbit reticulocyte system [23].

The degradation parameters β and γ are specified by means of the mRNA and protein half-life parameters ρ and π, respectively, which satisfy

$$e^{-\beta\rho} = \frac{1}{2}, \qquad e^{-\gamma\pi} = \frac{1}{2}. \tag{30}$$

In this case,

$$\beta = \frac{\ln 2}{\rho}, \qquad \gamma = \frac{\ln 2}{\pi}. \tag{31}$$

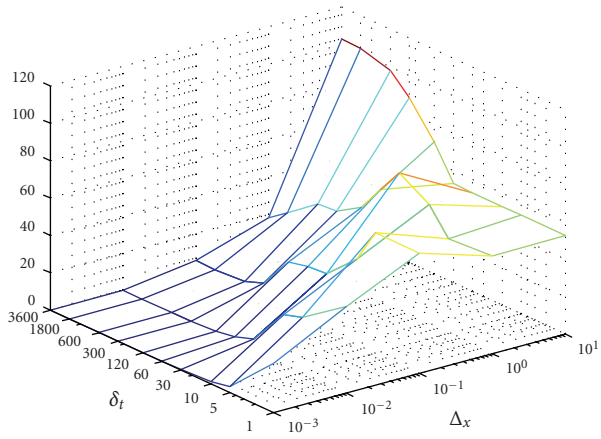
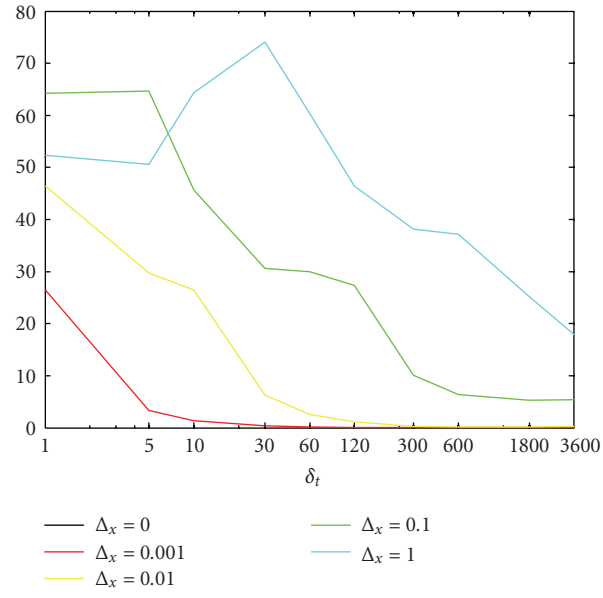


FIGURE 7: Results for the first simulation: the vertical axis shows the distance $\widetilde{d}_{L_1}(f_{(\Delta_x,\delta_t)}, f_{(\Delta_x=0,\delta_t)})$ as function of quantization levels for both the values (axis labeled "Δ_x") and the time (axis labeled "δ_t").



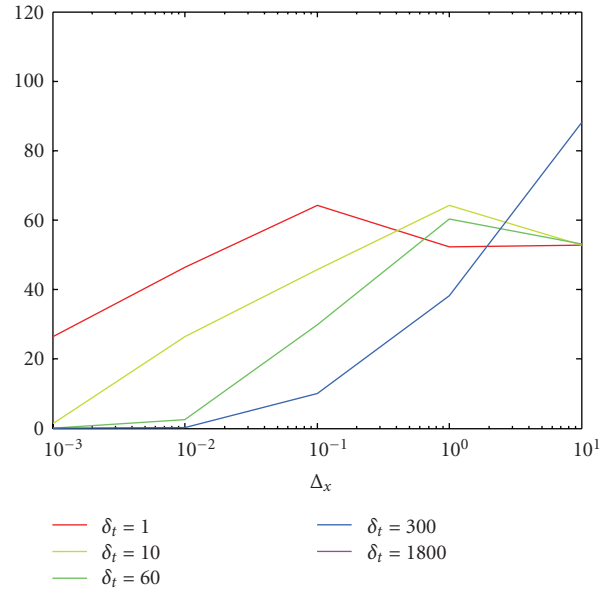
(a)

3.4. Results and discussion

It is expected that the finer the quantization is (smaller values of Δ_x), the more similar will be the projected networks to the reference networks. This similarity should be reflected by the trajectories as measured by the proposed metric. A straightforward simulation consists of the design of a reference network, the design of a projected network (for some value of Δ_x), the generation of several trajectories for both networks from randomly selected starting points, and the computation of the average distance between trajectories, using (9) and (21). Each process is repeated for different time intervals δt to study how the time intervals used in the simulation affect the analysis.

The firstsimulation is based on the same 4-gene model presented in [7]. We use 6 different quantization levels, $\Delta_x = 0, 0.001, 0.01, 0.1, 1$, and 10, where $\Delta_x = 0$ means no quantization, and designates the reference network. For each quantization level Δ_x and starting point x_0, we generate the simulated time series expression and compare it to the time-series generated with $\Delta_x = 0$ (the reference network), estimating the proposed metric using (21). The process is repeated using a total of 10 different time intervals, δ_t = 1 second, 5 seconds, 10 seconds, 30 seconds, 1 minute, 2 minutes, 5 minutes, 10 minutes, 30 minutes, and 1 hour. The simulation is repeated and the distances are averaged for 30 different starting points x_0.

Figures 5 and 6 show the trajectories and empirical cumulative density functions estimated from the simulated system as illustrated in the previous section. Several quantization levels are used in the simulation. The last graph in Figure 5 shows the mRNA concentration for the forth gene, over the 10 000 first seconds (transient) and over the last 10 000 seconds (steady-state). We can see that for quantizations 0 and 0.001, the steady-state solutions are periodic, and for quantizations 0.001 and 0.1, the solutions are constant. This is reflected by the associated plot of $F(x)$ in Figure 6.



(b)

FIGURE 8: Results for the first simulation: the vertical axis shows the distance $\widetilde{d}_{L_1}(f_{(\Delta_x,\delta_t)}, f_{(\Delta_x=0,\delta_t)})$ as function of quantization levels for both the values (labeled "Δ_x") and the time (labeled "δ_t"). Part (a) shows the distance as a function of Δ_x for several values of δ_t. Part (b) shows the distance as a function of δ_t for several values of Δ_x.

Figure 7 shows how strong quantization (high values of Δ_x) yields high distance, with the distance decreasing again when the time interval (δ_t) increases. The z-axis in the figure represents the distance $\widetilde{d}_{L_1}(f_{(\Delta_x,\delta_t)}, f_{(\Delta_x=0,\delta_t)})$.

In our second simulation, we use a different connectivity (all other kinetic parameters are unchanged), and we

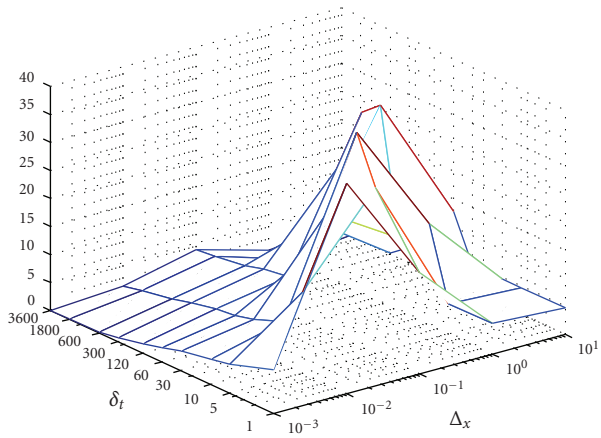


FIGURE 9: Results for the second simulation: the vertical axis shows the distance $\widetilde{d}_{L_1}(f_{(\Delta_x,\delta_t)}, f_{(\Delta_x=0,\delta_t)})$ as function of quantization levels for both the values (axis labeled "*Dx*") and the time (axis labeled "delta t").

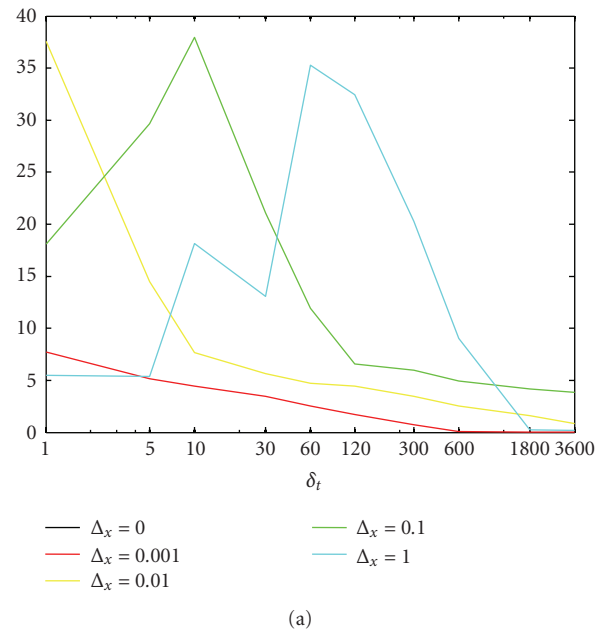
again use 10 different time intervals, $\delta_t = 1$ second, 5 seconds, 10 seconds, 30 seconds, 1 minute, 2 minutes, 5 minutes, 10 minutes, 30 minutes and 1 hour, and 6 different quantization levels, $\Delta_x = 0$, 0.001, 0.01, 0.1, 1, and 10. ($\Delta_x = 0$ meaning no quantization). The simulation is repeated and the distances are averaged for 30 different starting points. Analogous to the first simulation, Figure 9 shows how strong quantization (high values of Δ_x) yields high distance, which decreases when the time interval (δ_t) increases.

An important observation regarding Figures 8 and 10 is that the error decreases as δ_t increases. This is due to the fact that the coarser the amplitude quantization is, the more difficult it is for small time intervals to capture the dynamics of slowly changing sequences.

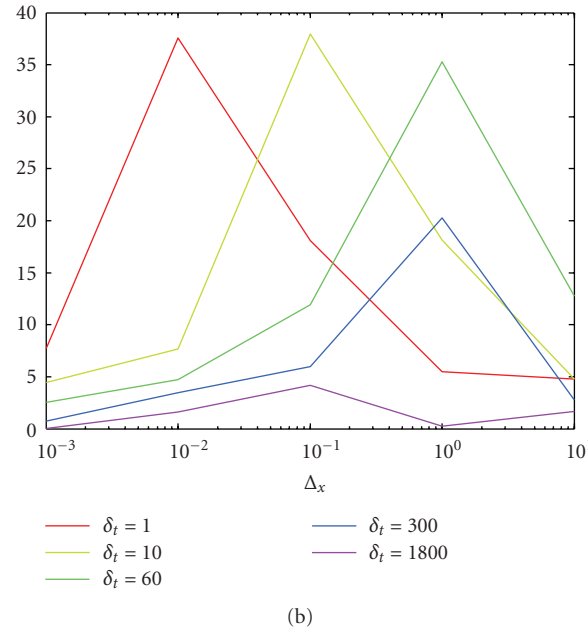
4. CONCLUSION

This study has proposed a metric to quantitatively compare two networks and has demonstrated the utility of the metric via a simulation study involving different quantizations of the reference network. A key property of the proposed metric is that it allows comparison of networks of different natures. It also takes into consideration differences in the steady-state behavior and is invariant under time shifting and scaling. The metric can be used for various purposes besides quantization issues. Possibilities include the generation of a projected network from a reference network by removing proteins from the equations and connectivity reduction by removing edges in the connectivity matrix.

The metric facilitates systematic study of the ability of discrete dynamical models, such as Boolean networks, to approximately represent more complex models, such as differential-equation models. This can be particularly important in the framework of network inference, where the parameters for projected models can be inferred from the reference model, either analytically or via synthetic data generated via simulation of the reference model. Then, given the reference and projected models, the metric can be used to determine the level of abstraction that provides the best inference; given the amount of observations available, this approach corresponds to classification-rule constraint for classifier inference in pattern recognition.



(a)



(b)

FIGURE 10: Results for the second simulation: the vertical axis shows the distance $\widetilde{d}_{L_1}(f_{(\Delta_x,\delta_t)}, f_{(\Delta_x=0,\delta_t)})$ as function of quantization levels for both the values (labeled "Δ_x") and the time (labeled "δ_t"). Part (a) shows the distance as a function of Δ_x for several values of δ_t. Part (b) shows the distance as a function of δ_t for several values of Δ_x.

NOMENCLATURE

Trajectory:	A function $f(t)$
Distance Function:	The proposed distance between networks

NOTATIONS

t:	Time
ψ:	Network
x_0:	Starting Point
$f(t), g(t), h(t)$:	Trajectories
f_{ss}, g_{ss}:	Steady-State trajectories
$f_{\psi,x_o}(t)$:	Trajectory
f_{tran}:	Transient part of the trajectory
f_{ss}:	Steady-state part of the trajectory
$F(x), G(x), H(x)$:	Cumulative distribution functions
$d_{tr}(\cdot,\cdot)$:	Distance between two trajectories
$d_{ss}(\cdot,\cdot)$:	Distance between two periodic or constant trajectories
$\lambda(A)$:	Lebesgue measure of set A
$\mathbf{f}(t)$:	Multivariate trajectory

ACKNOWLEDGMENTS

We would like to thank the National Science Foundation (CCF-0514644) and the National Cancer Institute (R01 CA-104620) for sponsoring in part this research.

REFERENCES

[1] H. De Jong, "Modeling and simulation of genetic regulatory systems: a literature review," *Journal of Computational Biology*, vol. 9, no. 1, pp. 67–103, 2002.

[2] R. Srivastava, L. You, J. Summers, and J. Yin, "Stochastic vs. deterministic modeling of intracellular viral kinetics," *Journal of Theoretical Biology*, vol. 218, no. 3, pp. 309–321, 2002.

[3] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[4] S. Kim, H. Li, E. R. Dougherty, et al., "Can Markov chain models mimic biological regulation?" *Journal of Biological Systems*, vol. 10, no. 4, pp. 337–357, 2002.

[5] R. Albert and H. G. Othmer, "The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*," *Journal of Theoretical Biology*, vol. 223, no. 1, pp. 1–18, 2003.

[6] S. Aburatani, K. Tashiro, C. J. Savoie, et al., "Discovery of novel transcription control relationships with gene regulatory networks generated from multiple-disruption full genome expression libraries," *DNA Research*, vol. 10, no. 1, pp. 1–8, 2003.

[7] J. Goutsias and S. Kim, "A nonlinear discrete dynamical model for transcriptional regulation: construction and properties," *Biophysical Journal*, vol. 86, no. 4, pp. 1922–1945, 2004.

[8] H. Li and M. Zhan, "Systematic intervention of transcription for identifying network response to disease and cellular phenotypes," *Bioinformatics*, vol. 22, no. 1, pp. 96–102, 2006.

[9] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks," *Machine Learning*, vol. 52, no. 1-2, pp. 169–191, 2003.

[10] A. Choudhary, A. Datta, M. L. Bittner, and E. R. Dougherty, "Control in a family of boolean networks," in *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS '06)*, College Station, Tex, USA, May 2006.

[11] L. Devroye, L. Györffi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, NY, USA, 1996.

[12] I. Ivanov and E. R. Dougherty, "Modeling genetic regulatory networks: continuous or discrete?" *Journal of Biological Systems*, vol. 14, no. 2, pp. 219–229, 2006.

[13] I. Ivanov and E. R. Dougherty, "Reduction mappings between probabilistic boolean networks," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 1, pp. 125–131, 2004.

[14] S. Ott, S. Imoto, and S. Miyano, "Finding optimal models for small gene networks," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '04)*, pp. 557–567, Big Island, Hawaii, USA, January 2004.

[15] L. F. Wessels, E. P. van Someren, and M. J. Reinders, "A comparison of genetic network models," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '01)*, pp. 508–519, Lihue, Hawaii, USA, January 2001.

[16] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, no. 5584, pp. 1183–1186, 2002.

[17] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, New York, NY, USA, 1993.

[18] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, Garland Science, New York, NY, USA, 4th edition, 2002.

[19] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, 1969.

[20] P. A. Lynn, *An Introduction to the Analysis and Processing of Signals*, John Wiley & Sons, New York, NY, USA, 1973.

[21] A. Arkin, J. Ross, and H. H. McAdams, "Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells," *Genetics*, vol. 149, no. 4, pp. 1633–1648, 1998.

[22] V. Iyer and K. Struhl, "Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 11, pp. 5208–5212, 1996.

[23] J. R. Lorsch and D. Herschlag, "Kinetic dissection of fundamental processes of eukaryotic translation initiation in vitro," *EMBO Journal*, vol. 18, no. 23, pp. 6705–6717, 1999.