Biophysical Journal Volume 101 December 2011 2563–2571

Effective Models of Periodically Driven Networks

Jason Shulman,* Lars Seemann, and Gemunu H. Gunaratne
Department of Physics, University of Houston, Houston, Texas

ABSTRACT Circadian rhythms are governed by a highly coupled, complex network of genes. Due to feedback within the network, any modification of the system's state requires coherent changes in several nodes. A model of the underlying network is necessary to compute these modifications. We use an effective modeling approach for this task. Rather than inferred biochemical interactions, our method utilizes microarray data from a group of mutants for its construction. With simulated data, we develop an effective model for a circadian network in a peripheral tissue, subject to driving by the suprachiasmatic nucleus, the mammalian pacemaker. The effective network can predict time-dependent gene expression levels in other mutants.

INTRODUCTION

Safe and effective manipulation of gene networks, including the design of successful therapeutic intervention for hereditary diseases, requires the ability to control the associated networks of genes/proteins/metabolites. The task is nontrivial. The networks contain a large number of nodes, coupled through (often unknown) nonlinear biochemical interactions (1). Adding to the difficulty of therapeutic intervention is the robustness of the solutions of most networks (2). Biological processes are necessarily insensitive to perturbations and even small changes to the network itself, e.g., mutations. Such a feature is required to ensure proper function and survival in a variable environment. This robustness, however, can complicate traditional therapies and has led to the inefficacy of medications, especially those designed to act on single molecular targets (2–4). Effects of an external modification on a gene may propagate through a large portion of the network, and feedback from downstream nodes may negate intended changes. These observations have led to the suggestion that successful interventions of complex diseases will require carefully designed multitarget therapy (2–4), a task that demands a comprehensive, global understanding of the underlying network. Even though gene expression levels in an organism can be obtained easily with current microarray technology, the number of nodes and nonlinearity of the couplings between them make it extremely difficult to infer an accurate model of the network (1,5).

These difficulties motivated us to introduce effective models (6), constructed from empirical data. The basic assumption in the formulation is the presence of a few core (or master) nodes, each of which controls actions of many other nodes on the network. These core nodes may be transcription factors (7,8) or perhaps microRNAs (9,10), as both can affect the level of a significant number of genes within a cell. The expression levels of genes outside the core group can be thought of as "slaved" to the core nodes, i.e., the expression levels of the slaved nodes are assumed to be algebraically related to those of the master nodes (11). Although the separation of master and slave nodes is artificial, the solutions of the effective model approximate those of the full network under conditions to be discussed later. Further, it allows for solutions of the full network to be manipulated by controlling the (small number of) master genes, suggesting the possibility of its use in the design of multitarget drugs. Previously, this approach was used to predict gene expression levels of mutants of an oxygen deprivation network of *Escherichia coli* (6). In this article, we extend the effective network approach to periodically driven systems (11).

Nearly all organisms possess an internal clock, a biological time-keeping device, which allows them to coordinate internal activities with their environment (12). It is thought that clocks, which are not required for viability, add to the survivability of an individual (12) and its robustness against external changes. Indeed, there is much evidence that circadian rhythms generated by the clock play an important role in the healthy function of organisms (12,13). Disruption of rhythms can cause a myriad of adverse effects, from the molecular level to that of the organism itself (14,15).

Circadian systems have a hierarchical structure (12). Environmental cues, such as the light/dark or temperature cycles, entrain an autonomous oscillator called a pacemaker, which is composed of circadian genes (12,16). The pacemaker, in turn, entrains and drives rhythms in other oscillators and genes under clock control (16). In this way, the circadian system regulates or is involved in a diverse set of functions and behaviors. The result is a vast network of genes, far larger than the core circadian system that is driven by the pacemaker. In mammals, as much as 10% of gene expression in specific tissues can exhibit circadian rhythmicity (17). At least 1% of the genes in *Drosophila* and ~2–6% in *Arabidopsis thaliana* are under circadian control (18). Among many examples are the couplings to the immune system (18–20), and cell cycle (21,22), as well as

Submitted July 18, 2011, and accepted for publication October 6, 2011.
*Correspondence: jshulman@uh.edu
Editor: Reka Albert.

0006-3495/11/12/2563/9 $2.00
doi: 10.1016/j.bpj.2011.10.008

the rhythm's role in tumorigenesis (23). Additionally, (possibly indirect) couplings have been identified between key clock and nonoscillatory genes, further expanding the network influenced by the circadian system (18,20).

The circadian pacemaker of mammals is the suprachiasmatic nucleus (SCN), a small collection of neurons located in the hypothalamus (12,16). Within each neuron, the core circadian system generates oscillations in gene expression, either autonomously or through entrainment by photic input received via the optic nerve, while the neuron itself exhibits oscillations in firing rate (24). These rhythms are responsible for coordinating the cycling of genes in the peripheral tissues, e.g., liver and kidneys (16,17). The oscillators in these tissues contain the same genes as in the SCN, but they cannot be entrained by light or, in general, maintain sustained oscillations when isolated from the SCN, i.e., oscillations in peripheral tissues are damped (17). Additionally, their oscillations are commonly phase-delayed by several hours relative to those of the identical genes within the SCN, which has been suggested to be the time required for the pacemaker signals to travel to the periphery (17).

At the most fundamental level, the circadian system itself is essentially a time-delayed, negative feedback loop (12,25). Although simple feedback loops can generate rhythms, they are sensitive to fluctuations and do not have robustness against mutations. In particular, for a single feedback loop, the elimination of just one component may be sufficient to destroy rhythmicity; this is not the case for organisms. Embellishments have been added to the single loop archetype as more genes and couplings have been discovered, resulting in multiple loop models that demonstrate increased robustness and adaptability (26). These observations motivate the synthetic circadian system we designed to test the effective model methodology. It consists of two coupled subsystems. The natural frequencies of the subsystems are chosen to be sufficiently close to each other, so that the coupling between them can provide frequency-locking (11). The analysis described below can be extended to systems that contain more than two feedback loops as well (17). We have also examined single networks, i.e., one system, with results similar to those presented here.

Our goal is to illustrate the power of the effective network methodology and to demonstrate that it extends beyond gene networks with steady-state solutions. The construction of an effective network requires knowledge of the set of N genes that belongs to it. Genes that cycle with a period of ~24 h are likely to belong to this set. In principle, they can be identified using several microarray experiments performed throughout the day. The development of an effective network also requires the identification of n ($<<N$) master genes of the network. This requires biological information, but transcription factors and microRNAs are possible candidates for master nodes. The experimental input necessary to generate an effective network is comprised of microarray data from wild-type (WT) organisms and single knock-out (SKO) mutants of each of the n master nodes. We require additional data to compute the effective network of the first harmonic; in the derivation below, we use WT data when the network is driven at a second frequency. This can be obtained by slightly changing the entraining environmental factors, such as the light/dark cycle.

For each organism, expression data will need to be collected 8–10 times over 24 h to extract the time dependences of gene expression levels. Thus, a total of $8 \times (n + 2)$ experiments are needed to generate an effective circadian model. This is a considerable amount of data; however, it is significantly less than that required to construct a first-principles model of the network, if such a task were possible. The effective model can be used to predict gene expression levels in other mutants. We will demonstrate this by using the model to determine expression levels of the $(N - 2)$ network genes in double-knock-out (DKO) mutants. In addition, network behavior can be calculated for individuals wherein some expression levels have been modified, e.g., heterozygous knockouts. Such a predictive ability should allow biologists to better design their experimental program and allow for effective allocation of their resources. In the future, we plan on demonstrating how one can externally modify a small number of nodes such that the entire circadian network is shifted into a prespecified state.

METHODS

Construction of the effective model

A detailed description of the development of the effective models has been presented previously (6). Here, we briefly illustrate the procedure and its adaptation to periodically driven networks, specifically, a circadian system in a peripheral (non-SCN) tissue. Let the state of the network with N genes $G = \{G_1,G_2,\ldots G_N\}$ be represented by $x(t) = \{x_1,x_2,\ldots,x_N\}$, where x_i is the expression level of the i^{th} gene. The dynamics of the network is assumed to be describable by a set of ordinary differential equations,

$$\begin{aligned} \dot{x}_1 &= F_1(x) + d_1(t), \\ \dot{x}_2 &= F_2(x) + d_2(t), \\ &\vdots \\ \dot{x}_N &= F_N(x) + d_N(t), \end{aligned} \tag{1}$$

where $F_k(x)$ are unknown functions of $x(t)$, and $d_k(t)$ is driving, which may represent the action of the SCN on clock genes in peripheral tissues. As mentioned above, these actions are time-delayed by several hours (17). The vector **d** describes the effect on the peripheral gene and not the original signal in the SCN.

A key requirement in constructing an effective model is the identification of master nodes. Together, they are needed to control the behavior of the remaining (slaved) nodes to a large degree. For example, in a gene network associated with a biological process, the transcription factors could be the master nodes.

The expression levels of master and slave nodes will be represented by X_M and x_s, where $M(i)$ and $s(j)$ denote the i^{th} master and j^{th} slave node, respectively. Consider the steady-state solutions to the expressions in Eq. 1 in the absence of driving. If the expression levels of the master nodes are fixed, the corresponding n equations are no longer valid and the

expression of the (N-n) slaved genes is determined by solving the remaining equations subject to the constraints imposed on X_M. As the master variables are changed, the solution for each slaved variable lies on an n-dimensional surface. Fig. 1 shows a two-dimensional representation of such a surface from the network used throughout this work.

The unperturbed, WT, system has a unique, stable, steady-state solution denoted by $P^{(0)} = \{P_M^{(0)}, p_s^{(0)}\}$, which corresponds to a point $\mathcal{P}_0$ that lies on this surface. Further, the steady-state equilibrium for a SKO of the K^{th} master gene in the system, $P^{(K)}$, will also be represented by a point on the surface. Knocking out the n genes of the master system individually, when combined with the WT $\mathcal{P}_0$, gives $(n + 1)$ surface points that define a unique n-dimensional plane. Although the size and nonlinearity of the network prevent access to the surfaces describing the expression of the slaved genes, experimentally obtaining the points $\mathcal{P}_K$ with $K = 0\ldots n$ allows for the determination of the n-dimensional plane. Due to the constraints, the plane lies close to the surface in the region of interest. Thus, points on the plane provide approximations to gene expression of the slaved system when the levels of the master genes are set externally. Consequently, one can approximately control the behavior of the slaved system by suitable manipulation of a small set of master genes. We next search for a system whose solutions lie on the plane and, as such, is linear. It can be written as

$$x_k - p_k^{(0)} = \sum_{I=1}^{n} B_{kI}\left(X_{M(I)} - P_{M(I)}^{(0)}\right), \tag{2}$$

for each slave index k. In Eq. 2 we have used the fact that $\mathcal{P}_0$ lies on the plane. The B_{kI} are determined by the points $\mathcal{P}_0,\ldots\mathcal{P}_n$. The crucial approximation in the construction of the effective model is that solutions of the network can be determined entirely (but approximately) by the state of the master system.

We have yet to consider the relationship between the master variables themselves. Indeed, due to coupling between them, perturbation of one or a few of the master nodes will modify the others. To continue with the linear approximation, we assume that the dynamics of the master system only depends on the master variables, and the relationships are

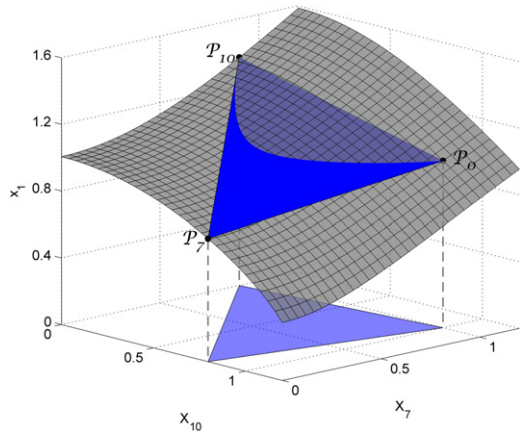

FIGURE 1 Expression levels of the slave gene G_1 (*gray surface*) as master genes G_7 and G_{10} are varied in the ranges shown. Lying on the surface are the WT equilibrium value $\mathcal{P}_0$ and those of the two SKO values, $\mathcal{P}_7$ and $\mathcal{P}_{10}$. The remaining master genes have been fixed to their equilibrium values to generate a three-dimensional plot for illustration. In general, the $n + 1$ points define an n-dimensional plane, which provides an approximation to gene expression levels of the slaved genes. The set of master gene values giving rise to the slave gene approximations is shown in the projection on the X_7-X_{10} plane.

$$\dot{X}_K = \sum_{I=1}^{n} A_{KI}\left(X_{M(I)} - P_{M(I)}^{(0)}\right). \tag{3}$$

The A_{KI} values are also obtained from the (experimental) gene expression data. They represent the effective interactions of genes in the network. Equations 2 and 3 define an effective empirical subnetwork (EES). Upon modification of some values of the master system, e.g., by knockout, the expression levels of the remaining master genes are approximated by Eq. 3. This information is then used in Eq. 2 to predict expression levels of the slaved nodes. The error generated by using the EES is often less than the experimental noise in the microarray data (experimental error and biological variability) used to generate the EES (27). Thus, the EES approximations are typically not the principal source of error in such a method. Use of the effective network results in a significant simplification. However, it can only be used to compute changes when the master variables are modified.

In this article, we adapt the EES methodology to analyze periodically driven networks. Here, the expression levels oscillate about the equilibrium values $P^{(0)}$ for the WT and $P^{(K)}$ for the SKO mutants. The oscillations of the slaved genes lie on the n-dimensional surface (*gray surface* in Fig. 1). Corresponding oscillations of the EES occur on the plane. If the equilibrium of another mutant, e.g., a DKO, is reasonably close to a $\mathcal{P}_K$ or if the surface is sufficiently flat, it is likely that the slope of the plane will be close to that of the surface. Then, the predicted oscillations will resemble that of the actual system, again providing an approximation to the time-dependent gene expression levels.

Here, we will concern ourselves with predicting behavior of the 0^{th} and first harmonics of gene expression levels (the EES formalism is adaptable to higher harmonics as well). The required time-dependent WT and SKO experimental input is Fourier-transformed to obtain the harmonics. For time-dependent gene expression, the EES construction for the 0^{th} harmonic is simply that of the steady-state case (Eqs. 2 and 3). For the first harmonic, the system of differential equations describing the dynamics of the master nodes is

$$\dot{X}_{1:K}(t) = \sum_{I=1}^{n} A_{KI}^{(1)} X_{1:M(I)}(t) + D_K(t), \tag{4}$$

or in Fourier space,

$$i\Omega_d X_{1:K} = \sum_{I=1}^{n} A_{KI}^{(1)} X_{1:M(I)} + D_K, \tag{5}$$

where $D_k(t)$ represents sinusoidal driving (from the SCN) at a frequency Ω_d. The value $X_{1:K}$ is the first harmonic of the expression level, and is complex with an amplitude and phase. The first harmonic expression of the slaved nodes, in analogy with Eq. 2, is a linear combination of the master variables, i.e.,

$$x_{1:k} = \sum_{I=1}^{n} B_{kI}^{(1)} X_{1:M(I)} + D_k. \tag{6}$$

Equations 2, 3, 5, and 6 are used to calculate the matrices **A** and **B** for the 0^{th} and first harmonics and the first harmonic driving vector **D**, which completely define the effective model. The model so constructed can be used to predict expression levels of mutant species.

The EES approach is unique in that its only input is expression data from microarray experiments. It does not require or assume connectivity or functional relationships between genes, nor does it need coupling parameter information, either experimental or estimated by spanning the parameter space. Despite its linearity, the EES is able to accurately describe gene expression of nonlinear networks due to the constrained geometry associated with its construction.

A synthetic circadian system

The expression level input required to generate the EES is rather specific. As such, experimental data with which we can test this approach is not available. In lieu of such data, we developed a class of synthetic circadian networks to generate oscillatory expression levels. We will demonstrate the viability of the effective network approach using data generated from a synthetic circadian system with $N = 20$ genes. However, the method has been tested extensively, for many networks of various sizes and with different forms of coupling.

The synthetic network was designed to represent a circadian system in a peripheral tissue, containing the core circadian genes and those under circadian control, which is subject to driving from the SCN. As is common for mathematical descriptions of biochemical reactions, the functions $F_k(x)$ in the expressions in Eq. 1 are chosen to be linear combinations of sigmoidal Hill functions,

$$H(x;c) = \frac{x^h}{x^h + c^h}.$$

That is,

$$F_k(x) = \sum_{i=1}^{N} a_{ki}\left[H(x_i;c_{ki}) - H\left(P_i^{(0)};c_{ki}\right)\right], \qquad (7)$$

with $P_i^{(0)}$ representing the equilibrium expression level of the i^{th} gene and h chosen to be 2. Note that $F_k(x) = 0$ if, for each i, $x_i = p_i^{(0)}$. The coefficient a_{ki} describes the nature and strength of the effect gene i has on gene k. It can be activating, inhibitory, or absent, corresponding to a_{ki} being positive, negative, or zero, respectively. The driving terms $d_k(t)$ in the expressions in Eq. 1 are sinusoidal, with frequency dictated by environmental signals or the endogenous free running period of the pacemaker. The values of a few driving terms are chosen to be zero, rendering the corresponding genes free from direct pacemaker control; their behavior is solely influenced by the actions within the network.

In keeping with models of circadian systems, the synthetic network contains a number of coupled subsystems (two), each of which is chosen to have a natural period close to 24 h. Oscillations in the peripheral tissues are not self-sustained. That is, they dampen when uncoupled to the pacemaker (17). Thus, our synthetic peripheral network must be damped as well. To achieve this, a linearization of Eq. 7 was used to compute the a_{ki} for the subsystems. We begin the construction by setting the eigenvalues of each subsystem. We require a pair of complex conjugate eigenvalues whose real parts, ε_r, are slightly negative and whose imaginary parts correspond to a period close to 24 h. The remaining eigenvalues are chosen to be smaller than $-\varepsilon_r$. Thus, each subsystem is a damped driven oscillator, and a sufficiently large driving at a period close to 24 h will drive them to a periodic solution.

We now perform a random orthonormal transformation on the diagonal matrix containing the eigenvalues of each subsystem. After the transformation, we have two nondiagonal matrices whose eigenvalues were predetermined. Each of the two subsystems contains 10 genes, numbered 1–10 and 11–20 for subsystems 1 and 2, respectively. Next, the genes between the two subsystems are sparsely coupled such that the coupling matrix has the form

$$a = \begin{pmatrix} \begin{pmatrix} a_{1,1} & \cdots & a_{1,10} \\ \vdots & \ddots & \vdots \\ a_{10,1} & \cdots & a_{10,10} \end{pmatrix} & \begin{pmatrix} a_{1,11} & \cdots & a_{1,20} \\ 0 & \ddots & 0 \\ a_{10,1} & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} a_{11,1} & \cdots & a_{11,10} \\ 0 & \ddots & 0 \\ 0 & 0 & a_{20,10} \end{pmatrix} & \begin{pmatrix} a_{11,11} & \cdots & a_{11,20} \\ \vdots & \ddots & \vdots \\ a_{20,11} & \cdots & a_{20,20} \end{pmatrix} \end{pmatrix}. \qquad (8)$$

The diagonal blocks represent the intrasubsystem couplings whereas the off-diagonal blocks describe the couplings between the two systems. One gene from each subsystem (genes 1 and 11) is subject to the influence of all genes in the opposite subsystem. The range of the coupling parameters is ± the average of the intrasubsystem couplings. Further, each of the other genes impacts exactly one gene in the opposite subsystem, with a range of values for the magnitude reduced by a factor of 5. The weak and sparse coupling between the subsystems does not significantly change the eigenvalues of Eq. 8 from those of the uncoupled systems. A graph of the network is displayed in Fig. 2. We emphasize that many coupling schemes have been analyzed, in addition to varying system size, with results similar to those presented here.

Synthetic WT data are obtained by solving the expressions in Eq. 1 for the time-dependent gene expression levels $x(t)$. For SKO mutants of the i^{th} gene, the i^{th} equation in the expressions in Eq. 1 is no longer valid as its level is externally set to zero. Data for the mutant are calculated by solving the remaining $(N - 1)$ equations with $x_i = 0$. The mutant input required for the model is not limited to the homozygous knockouts described here; heterozygous knockouts and mutants in which a gene is fixed to a nonzero value can also be incorporated into the methodology. Fig. 3 shows the time dependence of expression for a selected set of genes (G_1, G_{16}, G_{17}) after entrainment to a 24-h environmental cycle. Upon knockout of a network gene, the expression of the remaining nodes is modified and manifested by changes in amplitude and phase of the principle and higher harmonics. The extent of the modification is determined by the coupling strength between the nodes. As an example, knocking out gene 3 has a significant effect on the expression of gene 1 (*dashed line* in Fig. 3). In contrast, it has only a minor influence on G_{16} and G_{17}, members of subgroup 2. In general, modification of a gene results in significant changes to other members of the subnetwork and negligible-to-moderate adjustments of expression levels of genes in the complementary subnetwork.

This synthetic nonlinear circadian network will be used to generate the WT and SKO data, necessary to construct effective models for the 0^{th} and first harmonics. Additionally, the network will provide DKO data that will be used to test the predictive abilities of the effective model. We emphasize that the synthetic system is simply a platform for testing the ability of the EES to predict periodic gene expression. Its role is to produce data, because the necessary experimental data are currently unavailable. The EES is constructed from the data, but is otherwise entirely independent of the system used to generate it.

RESULTS

We randomly selected three genes from each subsystem as master nodes. These are genes [G_3, G_7, G_{10}, G_{11}, G_{17}, G_{18}].

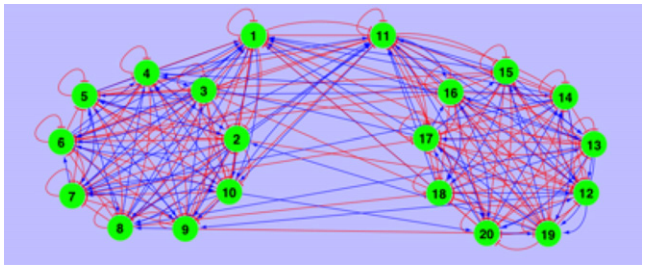

FIGURE 2 Graph of the synthetic gene network. The network is composed coupled subgroups, genes (1–10) and (11–20). Full coupling exists within the subgroups, and each gene influences two genes in the opposite subgroup. (*Arrows* and *bars*) Activating and inhibitory interactions, respectively. Each subgroup is constructed with a natural period close to 24 h.

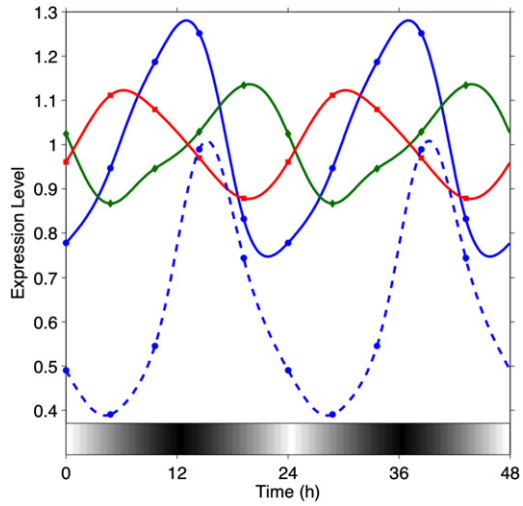

FIGURE 3 Partial collection of output from the synthetic network under 24 h entrainment. WT gene expression levels for G_1 (●),G_{16} (■), and G_{17} (◆) are shown. The impact of a knockout of gene 3 on gene 1 is included (*dashed line*). Only the real parts of the expression levels are shown.

All others are relegated to the slave system. Construction of the EES requires (time-dependent) gene expression levels from the six mutants with the master nodes individually knocked out, plus the WT organism at two driving frequencies. This is obtained by solving the expressions in Eq. 1 under the appropriate conditions. The 0^{th} and first harmonics are extracted from the data, and the model matrices (**A** and **B**) for each harmonic as well as the driving vector **D** for the first harmonic are calculated.

The expression of the master nodes can now be used to approximate the behavior of the rest of the network. Consider the example presented in Fig. 1. The gray surface describes the expression level of G_1 as a function of X_7 and X_{10}, the expression levels of G_7 and G_{10}, respectively. The EES predictions for G_1 lie on the plane defined by steady-state points (data) [$\mathcal{P}_0$, $\mathcal{P}_7$, $\mathcal{P}_{10}$]. The difference between the surface and plane is the error of the prediction. Given the master gene values denoted by the projection of the EES plane on to the X_7-X_{10} plane in Fig. 1, the quality of the predictions for G_1 has been evaluated. The average difference between an actual and EES solution is 3.1%, with a maximum of ~10%.

To demonstrate the predictive ability of the EES on the entire network, we will choose two master nodes and externally set their expression levels to zero, i.e., a DKO mutant will be constructed. The full network will accommodate this manipulation by an adjustment in gene expression, as determined by the couplings (Eq. 7 for the synthetic network). For the effective model, the result of fixing these values on the other master and slaved nodes is determined by Eqs. 2, 5 and 3, 6, respectively. These are the predictions of the model. They will be compared to the actual expression levels, the solutions to the expressions in Eq. 1, subject to the DKO constraints.

As an example, consider the effect of knocking out genes 3 and 11. There are 18 expression levels of the mutant to be estimated. The average error between the predicted and actual 0^{th} harmonic values is 0.93%, whereas that of the amplitude and phase of the first harmonic is 1.35% and 1.40°, respectively. These results are typical for this system. One can gain a qualitative appreciation for the accuracy of the predictions by examining the time dependence of the gene expression. This is shown for a representative set of genes in Fig. 4. The actual expression levels from the full network are plotted as solid lines, which clearly demonstrate the presence of higher harmonics. The corresponding predictions of the EES are used to construct an approximate time dependence, $x_k(t) = x_{0:k} + x_{1:k}e^{i\Omega_d t}$, where $x_{i,k}$ is the i^{th} harmonic prediction for G_k. This is represented by dashed lines in the figure. The predicted expression levels (by construction) contain only the primary harmonic, yet the magnitude and phase are aligned well with the actual values from the synthetic system.

To fully analyze the predictive ability of the effective model, gene expression from all possible DKO mutants was predicted and evaluated in a manner similar to that described for the DKO[3, 11] mutant. For a six-member master system such as the one provided above, there are $\binom{6}{2} = 15$ such DKO mutants. Each of these exhibited oscillatory, bounded expression levels, which was considered the criterion for viability of the synthetic mutant. With 18 functioning genes in each mutant, there are $15 \times 18 = 270$ gene expression levels to be predicted. For the 0^{th} harmonic, the average error is 0.63%. There are 220 predictions (81%) that are within 1% of their exact values

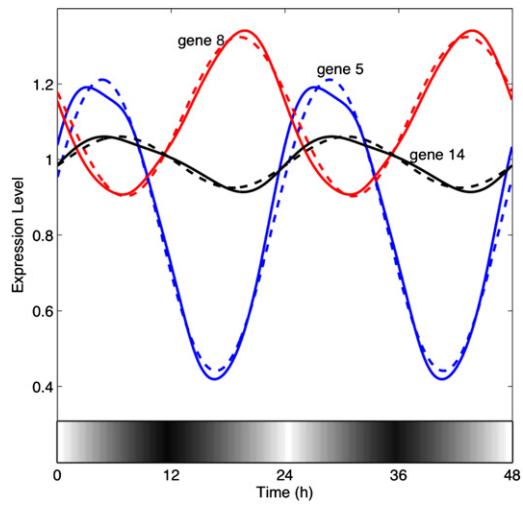

FIGURE 4 Expression levels (real part) for genes in a DKO[3, 11] mutant. (*Solid lines*) Solutions of the double knockout. (*Dashed lines*) Corresponding predictions from the EES. Despite the approximate nature of the effective model, the time dependence of the predictions matches that of the actual values quite well. Note that, by construction, the effective model only contains the 0^{th} and first harmonics.

whereas 47 (17%) fall between 1 and 5%. Only three genes have 0^{th} harmonic errors >5%. These results are summarized in the histogram of Fig. 5 and Table 1.

The coefficient of each first harmonic is complex, and illustrated using the amplitude and phase. The accuracies of these components were investigated individually (see Table 1). The average amplitude error was found to be 1.68% whereas that of the phase was 1.31°. Fig. 6 displays the results of the first harmonic calculations in polar form with the radial axis representing the ratio of the amplitudes ($|X_{1:p}|/|X_{1:a}|$or $|x_{1:p}|/|x_{1:a}|$) and the angular axis corresponding to the phase difference between the predicted (p) and actual (a) expression levels. Thus, the point (1,0) corresponds to an exact prediction. The ratios for the DKOs closely scatter about this location. The size and proximity of the data points render it difficult to accurately judge the quality of the predictions, so the inset of Fig. 6 displays an expanded view that includes predictions that are accurate to within 5° and 5% in amplitude. This accounts for 91% of the predictions.

We have thus far only considered differences between the solutions to the synthetic system and predictions of the effective model. However, in any real-world application of the EES, the input expression levels will contain experimental error. The effect of such error must be small if the method is to be useful. Consider, as an example, the DKO [3,11] introduced previously. To model experimental error, we multiply the input data for the EES (i.e., solutions from the synthetic model) by a random value from a normal distribution with a mean of 1 and standard deviation ε. Levels of the DKO[3,11] mutant were calculated for 10,000 such experiments. For realistic values of ε (27), the difference between the predicted and exact expression levels is principally governed by the experimental error, with a smaller contribution due to use of the EES. Of course, if the measurement errors are small, the EES is the primary source of error in the analysis, although, in this case, it also is quite small (<2%). The phase predictions of the first harmonic are somewhat insensitive to the experimental error. The average ranged from 1.4 to 2.59° as the experimental error varied from 0 to 10%. Thus, we note that experimental error in microarray data is more significant than error due to the approximations used in the EES.

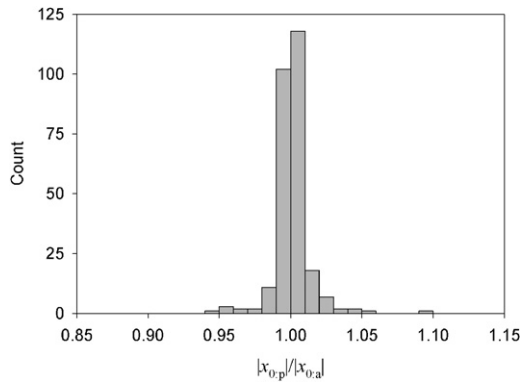

FIGURE 5 Histogram for the ratio of the predicted (p) and actual (a) 0^{th} harmonics of DKO mutants. All predictions are within 10% of the actual values.

TABLE 1 Number and percentage of genes (out of 270) predicted within an error range

	0^{th} harmonic	First harmonic	
Error range (percent or degree)	Amp	Amp	Phase
(0,1)	220 (81%)	153 (57%)	162 (60%)
(1,5)	47 (17%)	99 (37%)	95 (35%)
$(5,\infty)$	3	18	13

Error has units of percent for amplitudes and degrees for phase.

It is clear that the effective model can accurately predict the time-dependent expression levels of the circadian network. It should be noted that the utility of the EES is not limited to predictions in DKOs. Any manipulations of the master system (but only of the master system) can be approximated by the effective model. This includes up- or downregulation of the master genes as well as triple knock-outs, etc. Such considerations, combined with the quality of the results, justify the use of the EES formulation to study of nonlinear, periodically driven gene networks.

DISCUSSION AND CONCLUSIONS

Large nonlinear models have been proposed for circadian systems (28,29). However, a great deal of effort goes into

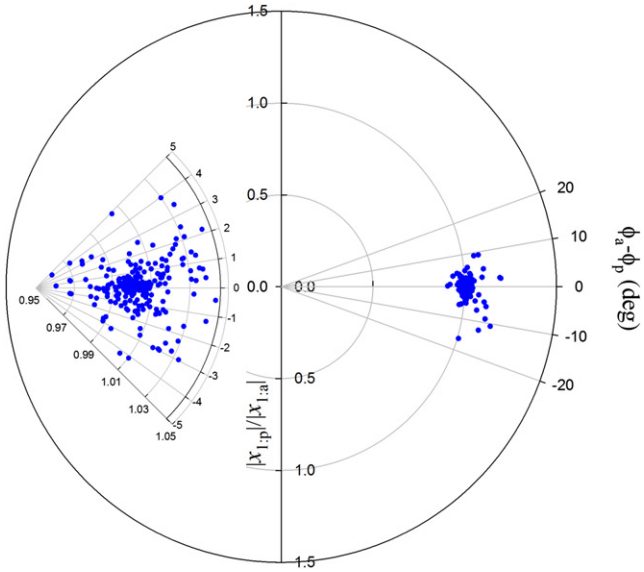

FIGURE 6 Plot indicating quality of first harmonic predictions in DKO mutants. The ratio of the predicted/actual magnitudes is plotted in the radial direction. The angular direction indicates phase difference. The data cluster around the point (1,0), indicating the predictions accurately describe the first harmonics generated by the network. (*Inset*) Expanded view of main figure. Includes predictions within 5% amplitude and 5° of phase.

determining the governing equations, interactions, and control parameters. Often, such information (for example, reaction rates, diffusion constants) is unknown or known only approximately, so such models can only be used to infer qualitative features of the underlying system. We have introduced an alternative approach to modeling that relies not on biochemical information but microarray data. For example, the complex role that proteins play within the network, which is not generally understood, is not needed for our effective model approach. Further, in the synthetic system, we used Hill functions to couple genes, yet the construction of the EES does not rely on this fact. However, it is important to appreciate that the effective network can only make a limited set of predictions about the underlying network.

One might suspect that increasing the size of the master system, i.e., inputting more data into the model, would increase its accuracy. Consider the extreme case of including all network genes in the master system. When this is done for the synthetic network, there are $\binom{20}{2} = 190$ possible DKO combinations. Four of these mutants were found to be not viable. There are 18 active genes in each of the 186 DKO mutants, resulting in 3348 gene expression levels to be calculated. The average error for the predictions is 0.66% and 1.98% for the amplitudes of the 0^{th} and first harmonics, respectively. The average phase error for the first harmonic is 1.32°. This is very similar to what was found for the six-member master system above, despite the inclusion of significantly more data. This example illustrates that enhancing the set of master nodes does not increase the accuracy of the predictions. However, more such predictions can be made.

This puzzling property is due to the nature of the EES construction. Given two master systems, possibly of different sizes, master and slave node predictions common to both systems will be identical. For example, the DKO[3,11] mutant described above can be calculated with both the 6- and 20-member master systems. The two effective models predict the 0^{th} harmonic expression of, for instance, G_1 to be 0.85. Similar equivalence is found for the 0^{th} and first harmonics of the remaining genes in the mutant. This effect will be illustrated graphically on a three-node network with steady-state expression levels (the mathematical treatment for master systems of different sizes is provided in the Supporting Material). Consider a three-node network in which G_1 is the sole member of the master system. In this case, the EES requires the WT $P^{(0)}$ and the SKO of G_1, $P^{(1)}$. As in Fig. 1, the slave system expression can be written in terms of the master expression level (Fig. 7 *a*). Here, however, the plane has been replaced by a line formed by the two points $\mathcal{P}_0$ and $\mathcal{P}_1$. As before, by fixing the expression level of G_1 (to a value Q_1), the predicted values of the slave system are determined (Q_2 and Q_3). If, however, genes

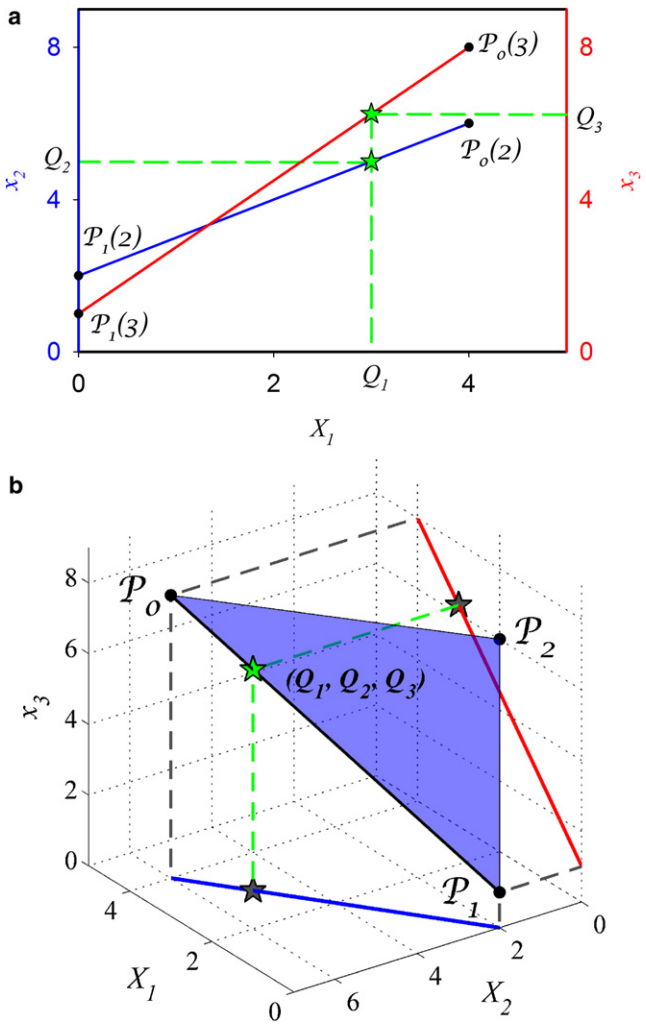

FIGURE 7 Slave gene predictions for a three-node network when the master system contains one gene (*a*) and two genes (*b*). By choosing the expression level of G_1 to be equal to Q_1, the remaining genes, master or slave, will be determined. In each case, the final state of the network is predicted to be $(Q_1, Q_2, Q_3) = (3, 5, 6.25)$. In general, the predictions are independent of size or membership of the master system. Of course, a larger set of predictions can be made with a larger set of master nodes.

G_1 and G_2 are chosen to form the master system, the required points $\mathcal{P}_0$, $\mathcal{P}_1$, and $\mathcal{P}_2$ define a plane that approximates the slave gene G_3 (Fig. 7 *b*). To compare the predictions from the two master systems, one must limit consideration to those common to both systems. Thus, the plane of the larger system must be restricted to the line defined by $\mathcal{P}_0$ and $\mathcal{P}_1$. Again, if one chooses $X_1 = Q_1$, the values of the other master variable, X_2, and the slave variable will be determined (see Eqs. 2 and 3 for the 0^{th} and Eqs. 5 and 6 for the first harmonics). This can be seen by considering the projections of the line on to the X_1-X_2 plane and the X_1-x_3 plane, respectively. Fixing X_1 determines a point on each of these projections, resulting in $X_2 = Q_2$ and $x_3 = Q_3$. Thus, by predetermining X_1,

both systems predict the same values for the remaining members of the network.

Such a feature has some important implications. First, note that the solution surfaces of the full network are not dependent on the choice of the master system. We have just shown that predictions are also independent of size and membership of the master system. Thus, the differences between the solution surfaces and the predictions, i.e., the prediction error, are also independent of the master system. Therefore, the accuracy of the predictions is not sensitive to the choice of master system. Without detailed knowledge of the full network, determination of a suitable master system may be difficult. The equivalence described above ensures that the model accuracy does not suffer if the size of the master system is underestimated. As an example, consider a gene that strongly regulates many others and, therefore, is a natural candidate for the master system. If a mutant is not viable, one cannot include the gene and loses the ability to utilize it to influence the other members of the network. However, based on the discussion above, one can be sure that the accuracies of the predictions generated by the chosen master system will be the same as those if the gene in question had been included.

It should be noted that, although the predictions common to both systems are identical, the larger system allows for more control over the network. For the hypothetical three-node network above, with the larger system, one can use both G_1 and G_2 to manipulate the expression of G_3, with the possible values indicated by the plane. For the smaller system, only those values along the lines in Fig. 7 *a* are available through control of G_1 alone.

The adaptation of the EES methodology to periodically driven systems extends the applicability of the technique beyond the steady-state gene networks to include, in particular, circadian networks of peripheral tissues. Through its many couplings to other subnetworks, such as those governing immune response and tumorigenesis, the circadian machinery has a vast influence on an organism. The ability to control the key clock genes may perhaps open novel avenues for potential treatment of complex diseases. The EES formulation has potential to provide such control while simultaneously requiring relatively few resources. This is evidenced by the quality of the DKO predictions. The majority of the amplitude predictions, for both the 0^{th} and first harmonics, were within 1% of the actual values, whereas the majority of the phase predictions were within 1°. This is true for both the 270 predictions using the six-node master system as well as the 3348 for the 20-member master system. Further, the 0^{th} and first harmonics combine to generate the predicted time-dependent gene expression levels, which closely match those of the underlying network, e.g., Fig. 4.

We hope our work motivates an experimental study on the construction of effective networks and a validation of the methodology outlined here.

SUPPORTING MATERIAL

Additional information with supporting equations is available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(11)01197-0.

This work was supported in part by the Texas Center for Superconductivity at the University of Houston. The authors rely on Cytoscape software to generate the network graph. J.S. acknowledges the Mathematical Biosciences Institute for workshop support and discussions. L.S. is supported by National Science Foundation award No. 0840889.

REFERENCES

1. Gardner, T. S., D. di Bernardo, …, J. J. Collins. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science.* 301:102–105.
2. Kitano, H. 2007. A robustness-based approach to systems-oriented drug design. *Nat. Rev. Drug Discov.* 6:202–210.
3. Csermely, P., V. Agoston, and S. Pongor. 2005. The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.* 26:178–182.
4. Zimmermann, G. R., J. Lehár, and C. T. Keith. 2007. Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discov. Today.* 12:34–42.
5. Shandilya, S. G., and M. Timme. 2011. Inferring network topology from complex dynamics. *New J. Phys.* 13:013004.
6. Gunaratne, G. H., P. H. Gunaratne, …, A. Török. 2010. Using effective subnetworks to predict selected properties of gene networks. *PLoS ONE.* 5:e13080.
7. Babu, M. M., N. M. Luscombe, …, S. A. Teichmann. 2004. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* 14:283–291.
8. Gill, G. 2001. Regulation of the initiation of eukaryotic transcription. *Essays Biochem.* 37:33–43.
9. Ambros, V. 2004. The functions of animal microRNAs. *Nature.* 431:350–355.
10. Bartel, D. P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell.* 116:281–297.
11. Cross, M. C., and P. C. Hohenberg. 1993. Pattern formation outside of equilibrium. *Rev. Mod. Phys.* 65:851–1112.
12. Dunlap, J. C., J. J. Loros, and P. J. DeCoursey. 2004. Chronobiology: Biological Timekeeping. Sinauer, Sunderland, MA.
13. Sephton, S., and D. Spiegel. 2003. Circadian disruption in cancer: a neuroendocrine-immune pathway from stress to disease? *Brain Behav. Immun.* 17:321–328.
14. Kuhn, G. 2001. Circadian rhythm, shift work, and emergency medicine. *Ann. Emerg. Med.* 37:88–98.
15. Meyer-Bernstein, E. L., and L. P. Morin. 1996. Differential serotonergic innervation of the suprachiasmatic nucleus and the intergeniculate leaflet and its role in circadian rhythm modulation. *J. Neurosci.* 16:2097–2111.
16. Hastings, M. H., A. B. Reddy, and E. S. Maywood. 2003. A clockwork web: circadian timing in brain and periphery, in health and disease. *Nat. Rev. Neurosci.* 4:649–661.
17. Roenneberg, T., and M. Merrow. 2003. The network of time: understanding the molecular circadian system. *Curr. Biol.* 13:R198–R207.
18. McDonald, M. J., and M. Rosbash. 2001. Microarray analysis and organization of circadian gene expression in *Drosophila*. *Cell.* 107:567–578.
19. Wang, W., J. Y. Barnaby, …, X. Dong. 2011. Timing of plant immune responses by a central circadian regulator. *Nature.* 470:110–114.
20. Panda, S., M. P. Antoch, …, J. B. Hogenesch. 2002. Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell.* 109:307–320.

21. Matsuo, T., S. Yamaguchi, …, H. Okamura. 2003. Control mechanism of the circadian clock for timing of cell division in vivo. *Science.* 302:255–259.

22. Zámborszky, J., C. I. Hong, and A. Csikász Nagy. 2007. Computational analysis of mammalian cell division gated by a circadian clock: quantized cell cycles and cell size control. *J. Biol. Rhythms.* 22: 542–553.

23. Fu, L., and C. C. Lee. 2003. The circadian clock: pacemaker and tumor suppressor. *Nat. Rev. Cancer.* 3:350–361.

24. Welsh, D. K., D. E. Logothetis, …, S. M. Reppert. 1995. Individual neurons dissociated from rat suprachiasmatic nucleus express independently phased circadian firing rhythms. *Neuron.* 14:697–706.

25. Hastings, M. H. 2000. Circadian clockwork: two loops are better than one. *Nat. Rev. Neurosci.* 1:143–146.

26. Saithong, T., K. J. Painter, and A. J. Millar. 2010. The contributions of interlocking loops and extensive nonlinearity to the properties of circadian clock models. *PLoS ONE.* 5:e13867.

27. Shi, L., L. H. Reid, …, W. Slikker, Jr, MAQC Consortium. 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* 24:1151–1161.

28. Goldbeter, A. 1995. A model for circadian oscillations in the *Drosophila* period protein (PER). *Proc. Biol. Sci.* 261:319–324.

29. Leloup, J. C., and A. Goldbeter. 2003. Toward a detailed computational model for the mammalian circadian clock. *Proc. Natl. Acad. Sci. USA.* 100:7051–7056.