Identification and Modeling of Genes with Diurnal Oscillations from Microarray Time Series Data

Wenxue Wang, *Member, IEEE,* Bijoy K. Ghosh, *Fellow, IEEE,* and Himadri B. Pakrasi, *Fellow, AAAS*

Abstract—Behavior of living organisms is strongly modulated by the day and night cycle giving rise to a cyclic pattern of activities. Such a pattern helps the organism to coordinate their activities and maintain a balance between what could be performed during the 'day' and what could be relegated to 'night'. This cyclic pattern, called the 'Circadian Rhythm', is a biological phenomenon observed in a large number of organisms. In this paper, our goal is to analyze transcriptome data from *Cyanothece* for the purpose of discovering genes whose expressions are rhythmic. We cluster these genes into groups that are close in terms of their phases and show that genes from a specific metabolic functional category are tightly clustered, indicating perhaps a 'preferred time of the day/night' when the organism performs this function. The proposed analysis is applied to two sets of micro array experiments performed under varying incident light patterns. Subsequently we propose a model with a network of three phase oscillators together with a central master clock and use it to approximate a set of 'circadian controlled genes' that can be approximated closely.

Index Terms—Gene expression, Circadian rhythm, Microarray time series, Diurnal cycle, Phase oscillation, Cyanothece, KaiC protein, Oscillator Network.

1 Introduction

Circadian rhythms are biological phenomena observed in a large number of organisms ranging from unicellular bacteria to human beings (see [4], [9] for a recent survey). With circadian rhythms, the temporal coordination of internal biological processes, both among these processes and with external environmental cycles, is crucial to the health and survival of these organisms (see [3]). The underlying biochemical mechanisms are well understood for many of the organisms in varying degree of details (see [2], [3], [12], [27], [41]). In recent years, there have been many excellent books written on this subject (see [6], [9], [35]). Roughly speaking, circadian rhythms have to do with periodic variations in the data triggered by at least one internal circadian clock that is set to the day and night oscillation of the light pattern (diurnal cycle) and other environmental cues impinging on the organism. The circadian clocks share a common basic mechanism involving oscillators that are composed of positive and negative elements, which vary in different organisms and form autoregulatory feedback loops (see [3], [34], [44]). By the feedback loops, the circadian clocks are entrained directly or indirectly to environmental time and do not change their responses even if the external cycles are momentarily altered. However the multiple oscillators in circadian clock systems are coordinated in different ways in diverse organisms. In unicellular organisms, cyanobacteria and fungi, at least one oscillator is directly linked to the environmental cues for entrainment and serves as the pacemaker synchronizing slave oscillators through direct and indirect coupling. In multicellular species, circadian complexity arises from the presence of molecular oscillators in various cell types. The circadian clock in mammals contains a complete SCN pacemaker that is formed by a collection of coupled cell-autonomous oscillators, whereas the avian circadian system is more complex and involve several coupled pacemakers that are present in the pineal gland, the retina, and the SCN. The centralized pacemakers respond to environmental inputs and coordinate the rhythms of peripheral oscillators in various tissues. In Drosophila *melanogaster*, unlike mammals and birds, the clock system lack a centralized pacemaker in the brain and consists of multiple light-responsive oscillators throughout the head and body that have pacemaker properties ([3]).

We hypothesize that there are processes in the cell that are 'in sync' with the internal clock and are therefore immune to changes in the light pattern. We shall call these processes 'circadian controlled' and it is important to isolate these processes from those which are not. In this paper, we analyze transcriptome data from *Cyanothece*, referred to as 'gene expression' data from microarray, and our objective is to identify the circadian processes and classify them in relation to their phases.

Because of its importance in understanding the dynamics of the cell cycle, especially in understanding how cells coordinate among each other, circadian oscillations have attracted the attention of modelers interested in the study of a cascade of oscillators. In this context we would like to refer to the work of Kronauer [19] and Winfree [42] who studied how

- *W. Wang is in the Institute for Collaborative Biotechnologies, University of California, Santa Barbara, CA 93106.*
E-mail: wenxue@engineering.ucsb.edu.
- *B. K. Ghosh is in the Department of Mathematics and Statistics, Texas Tech University, Lubbock, Texas, 79409.*
E-mail: bijoy.ghosh@ttu.edu.
- *H. B. Pakrasi is in the Department of Biology, Washington University in Saint Louis, MO 63130.*
E-mail: pakrasi@wustl.edu.

a pattern of light results in entraining a diurnal cycle of the same frequency. Kronauer's work was subsequently pursued by Strogatz [39] in analyzing asymptotically a large array of oscillators and figuring out when these oscillators synchronize. In contrast, we have a small array of oscillators that are not interacting among themselves strongly [15], but are under tight control of an internal clock. Global synchronization takes place as a result of each oscillator synchronizing with the clock.

In cyanobacteria, circadian rhythms have been reported in amino acid uptake [7], cell division [18], [26] and for various other gene expressions [23], [31]. Extensive studies have identified almost 100% of genes in *Synechococcus elongatus* PCC 7942 [23] and 77% in *Synechocystis* sp. PCC 6803 [1], respectively, showing circadian rhythmic activity using promoter trap analysis. These two species of cyanobacteria do not perform nitrogen fixation. In another specie of cyanobacteria, *Cyanothece* sp. ATCC 51142, which performs nitrogen fixation, it was recently reported that about 30% of genes was identified exhibiting circadian rhythm using global transcriptomic analysis of microarray gene expression under nitrogen-fixing condition in alternating light-dark cycles [38]. A circadian clock in cyanobacteria can be synthesized using the interaction between the products of three kai genes, kaiA, kaiB and kaiC [14], [16]. The autophosphorylation cycle of KaiC oscillates robustly in the cell with a 24 hr. period and is essential for the basic timing of the clock [28], [30].

The *Cyanothece*, we have referred to, is a unicellular diazotrophic cyanobacterium capable of photoautotrophic as well as heterotrophic growth [33]. In order to temporally separate the oxygen labile nitrogenase enzyme from oxygenic photosynthesis, it exhibits robust circadian rhythms in photosynthesis, nitrogen fixation, respiration and the synthesis and degradation of storage granules [37]. It is also the first unicellular nitrogen-fixing cyanobacterium to be fully sequenced and thus provides a basis upon which the regulation of such circadian controlled metabolic processes and storage capabilities may be further investigated.

In this paper, gene expression data from *Cyanothece* [38] has been analyzed for the purpose of discovering rhythmic genes and subsequently group the genes into clusters that are rhythmically close (co-rhythmic). Our main interest is to pick a cluster of genes with a 24 hr. period and map this 'gene cluster' to its corresponding metabolic processes and identify the phase relationship between several sub processes [10], [38]. Using a network of three phase oscillators, synchronized by an underlying master clock, we describe a model for the rhythmic genes. We show that many of the genes in the cluster can be approximated by the proposed model even when the frequencies of the phase oscillators vary within an interval. Throughout the paper, we compare transcriptome data from *Cyanothece* coming from two different sets of experiments [38], [40], performed under identical experimental conditions except that the choice of the incident light is different. An important highlight of this paper is the conclusion that many of the sub processes isolated from the cluster of genes with 24 hr. period do not change appreciably in phase under varying light conditions.

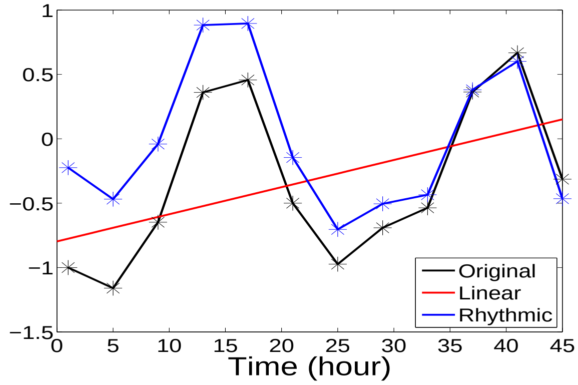

Fig. 1: The expression pattern of a rhythmic gene from the microarray data (black), its linear trend (red) and rhythmic component (blue).

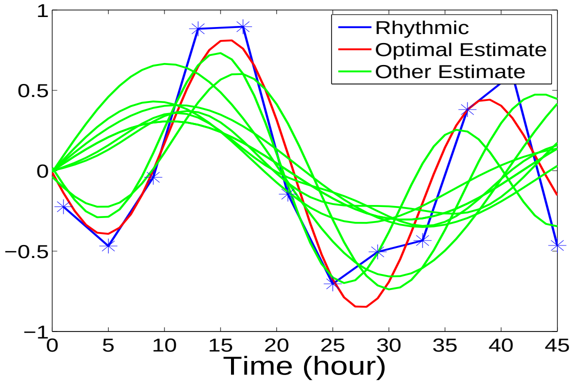

Fig. 2: The rhythmic component of a gene expression and its estimates with two sinusoidal functions over different pairs of frequencies. The red curve is the estimate with the optimum frequency pair.

2 Two Genome Wide Studies Utilizing Microarrays

For the purposes of studying circadian oscillations of co-rhythmic genes of different sub processes and their phase relations throughout a 24 hr. diurnal period, two sets of transcriptome data from *Cyanothece* sp. ATCC 51142 have been investigated [38], [40]. The unicellular diazotrophic cyanobacterium, *Cyanothece*, is capable of atmospheric nitrogen fixation and oxygenic photosynthesis, which are two important metabolic processes. The two processes are at odds with each other and usually do not exist in the same cell in many other cyanobacterial species, such as *Anabaena* and *Nostoc* [13], [43], since the nitrogen fixing enzyme, nitrogenase, is highly sensitive to oxygen produced during the photosynthesis. Nitrogen fixation is spatially separated and exists in specialized heterocysts, wherein cells differentiate with specific expressions of numerous genes [5]. The co-existence of two processes in the same cell of *Cyanothece* is achieved by a temporal regulation of intracellular environment in which nitrogen fixation occurs at night and photosynthesis during the day throughout a diurnal cycle [37]. Two sets of microarray experiments were implemented using *Cyanothece* cultures, which were grown in ASP2-N medium. In the 'minus nitrogen' medium, the cells had to perform nitrogen fixation. Both experiments were carried out with the same microarray chip platform and under the same experimental conditions, except for the pattern of the controlled variable, 'light'. One experiment was performed in alternating 12h/12h light-dark (LD) cycle with white light [38]. Samples were extracted every 4 hours for two days, starting with one hour into dark period, 'D1' (DLDL). Another experiment applied 12h/12h LD conditions first, and then switched to constant light condition (LL) [40]. Samples were taken every 4 hours for two days, under LD and LL conditions respectively, starting with two hours into light period, 'L2' (LDLL). In total, 12 samples were collected for each of the two experiments.

The data from each microarray were independently normalized using LOWESS normalization algorithm [32] to correct for variations in labeling intensities between two channels of the microarray: target channel and control channel. The statistical quality assessment using T-test indicated a good consistency in data. After normalization, $\log_2$(target/control) ratios were calculated at all time points to get gene expression patterns of about 4660 genes. Some expression patterns show a rhythmic change as shown in Fig. 1 (black curve).

In [38] the experimental data (DLDL) has been analyzed in some details. These experiments were conducted with the primary purpose of investigating global transcriptional changes in *Cyanothece* under nitrogen fixing conditions. The primary goal was to identify genes with diurnal rhythms in their expression patterns. This was accomplished by computing Pearson Correlation between pairs of genes and visualizing this data using Cytoscape (version 2.3.2 [36]). The 'correlation network' on the Cytoscape clearly showed four distinct subnetworks of coregulated genes corresponding to central cellular functions, including nitrogen fixation and photosynthesis and it was found that *Cyanothece* cells have increased transcriptional activity at night, implying that the demands of nitrogen fixation trigger major metabolic activities.

In a subsequent study [10], the experimental data (DLDL) from [38] was combined with the data (LDLL) from [40] in order to identify genes that continue to oscillate under both LD and LL conditions. In the same study [10], genes were identified that oscillated under LD but failed to oscillate under the LL condition. Processes that predominantly contained genes in the former group and the latter group were separately identified and these results were compared with those obtained in [38].

In this paper, our goal is to analyze the (DLDL) and the (LDLL) data sets to separately identify processes that show a diurnal rhythm. We show that many processes have genes that are tightly 'phase locked' and argue that these processes are active during a suitable 'preferred time'. This conclusion is not fundamentally different from what was noted in [38], in the sense that certain processes contain a large number of genes that peaked at a certain specific time during the 24 hr. cycle. Our analysis here combines an additional data set from [40] and utilizes quantitative estimates of the 'phase' leading to a detailed description of sub processes that are active at a specific preferred time.

Cycling Genes in Functional Categories

Functional Category	I	II	III
ACB (96)	48	34	43
BCPC (128)	53	38	52
CE (79)	39	38	38
CP (106)	37	37	42
*CIM (58)	34	27	34
DNA RMRR (84)	7	12	4
*EM (113)	53	51	51
FAPSM (51)	25	19	21
OC (380)	120	125	110
*PSR (148)	116	108	114
PPNN (46)	13	13	11
RF (194)	52	48	45
TC (38)	21	18	20
*TL (190)	98	78	96
TBP (249)	55	58	53
Unassigned (2702)	742	845	711
Total	1513	1549	1445

TABLE 1: Genes with a primary frequency of 24 hr. period and their assignment into different functional categories. The columns **I** and **II** refer to the optimization procedure described in this paper for the two experiments with light patterns DLDL and LDLL respectively. Column **III** refers to the algorithm described in the literature [38] that utilizes the peak value of the gene expressions. See Table 2 for definition of the functional categories.

3 Isolation of Rhythmic Genes from the Two Microarray Time Series Data

In this section, we analyze and isolate genes whose log ratio patterns (the 'gene expressions') have rhythmic components. Example of one such expression has been shown in Fig. 1 (black curve). We denote a gene expression as a time function $g(t)$ and expand it using Fourier components as follows:

$$\begin{aligned} g(t) \quad = \quad & a + bt + \alpha_1 \sin(\omega_1 t + \theta_1) \\ & + \alpha_2 \sin(\omega_2 t + \theta_2) + error(t). \end{aligned} \tag{1}$$

In (1), the term '$a+bt$' is the linear trend of the expression $g(t)$ (see Fig. 1 (red curve)). The rhythmic component has been approximated by a pair of frequencies ω_1 and ω_2 in order to approximate expressions that do not have a pure tone. The linear trend of gene expression pattern can be obtained using linear regression. Starting from the pairs $((t_i, g(t_i)), i = 1, ..., 12)$, the observed values of the expression, the parameters a and b are obtained as follows:

$$a = \frac{\sum g_i - b \sum t_i}{N}, \qquad b = \frac{N \sum g_i t_i - \sum g_i \sum t_i}{N \sum t_i^2 - (\sum t_i)^2}.$$

The rhythmic components in the gene expression can be obtained by removing the linear trend and expressing

$$\bar{g}(t_i) = g(t_i) - a - bt_i, \quad i = 1, \cdots, 12 \tag{2}$$

as follows:

$$\bar{g}(t_i) = \alpha_1 \sin(\omega_1 t_i + \theta_1) + \alpha_2 \sin(\omega_2 t_i + \theta_2) + error(t).$$

We shall call ω_1 the primary frequency and ω_2 the secondary frequency. The frequencies ω_1 and ω_2 are obtained using a

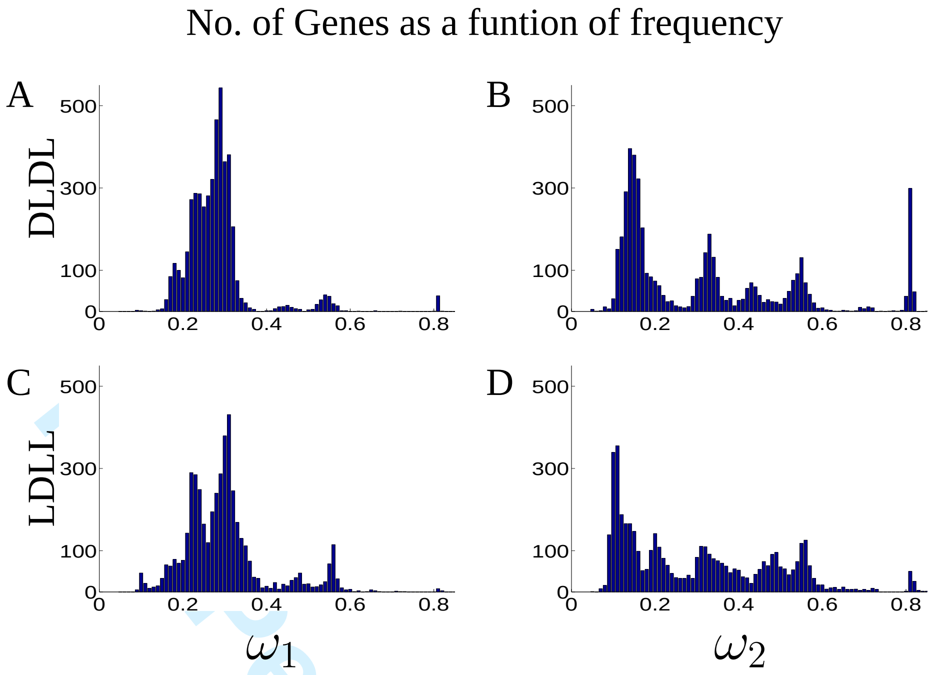

Fig. 3: Histograms showing the number of genes as a function of the primary frequency ω_1 (Panel A and Panel C) and secondary frequency ω_2 (Panel B and Panel D). The top two histograms (Panel A and Panel B) are for the light pattern DLDL and the bottom two histograms (Panel C and Panel D) are for the light pattern LDLL.

stepwise algorithm. First we obtain the best ω_1, θ_1, α_1 that minimizes the error

$$\sum_{i=1}^{N} (\bar{g}(t_i) - \alpha_1 \sin(\omega_1 t_i + \theta_1))^2. \tag{3}$$

The best ω_2, θ_2, α_2 are obtained by repeating the procedure on the residue. The optimization problem is tackled as follows.

Given a frequency, ω_1, the parameters α_1, θ_1 are chosen so that the cost function

$$E(\omega_1) = \min_{\alpha_1,\theta_1} G(\alpha_1, \theta_1)$$

is minimized where $G(\alpha_1, \theta_1)$ is the cost function described in (3). The optimal parameters α_1, θ_1 are obtained using the gradient method:

$$\begin{bmatrix} \alpha_1 \\ \theta_1 \end{bmatrix}_{k+1} = \begin{bmatrix} \alpha_1 \\ \theta_1 \end{bmatrix}_{k} - \mu \nabla G(\alpha_1, \theta_1),$$

where μ is the size of change in the vector (α_1, θ_1) and $\nabla G(\alpha_1, \theta_1)$ is the gradient of the function $G(\alpha_1, \theta_1)$. The frequency, ω_1, is varied over an interval and the estimation errors over ω_1 are computed. The frequency that makes the smallest error is chosen as the optimal frequency of the first sinusoid. As noted earlier, the optimization algorithm is repeated a second time to obtain the best ω_2, θ_2, α_2 on the residue. Fig 2 shows the best estimate (red curve) with the optimal frequencies, ω_1 and ω_2 together with many other curves that are not optimal.

The two step algorithm is repeated for all gene expressions from each of the two microarray experiments. In Figs. 3 the histograms show the number of genes as a function of the optimal frequencies ω_1 and ω_2 under the light condition DLDL

The Abbreviation of Functional Categories and SubCategories

Abbreviation of Functional Categories	
Abbreviation	Functional Category Name
ACB	Amino acid biosynthesis
BCPC	Biosynthesis of cofactors, prosthetic groups, carriers
CE	Cell envelop
CP	Cellular processes
CIM	Central intermediary metabolism
DNA RMRR	DNA replication, modification, recombination, repair
EM	Energy metabolism
FAPSM	Fatty acid, phospholipid, sterol metabolism
OC	Other categories
PSR	Photosynthesis and respiration
PPNN	Purines, pyrimidines, nucleosides, nucleotides
RF	Regulatory functions
TC	Transcription
TL	Translation
TBP	Transport and binding proteins

Abbreviation of Functional SubCategories	
Abbreviation	Functional SubCategory Name
NF	Nitrogen fixation
PG	Polysaccharides and glycoproteins
PPP	Pentose phosphate pathway
GL	Glycolysis
PB	Phycobilisome
PS II	Photosystem II
CF	CO_2 fixation
PS I	Photosystem I
ATP	ATP synthase
RP	Ribosomal proteins
DPPG	Degradation of proteins, peptides, and glycopeptides

TABLE 2: The functional categories and subcategories in *Cyanothece* and their abbreviation.

(Fig. 3A and Fig. 3B) and LDLL (Fig. 3C and Fig. 3D). Most of the expressions have a primary frequency ω_1 around $\frac{2\pi}{24} = 0.26$ radians/hr. (i.e. a 24 hr. period) and a small portion of gene expressions have ω_1 around $\frac{2\pi}{12} = 0.52$ radians/hr. (i.e. a 12 hr. period) as shown in Fig. 3A and Fig. 3C. This fact has already been observed and reported in [10]. The secondary frequency component ω_2 appears to be distributed over the entire range and occurs with an overall small magnitude except perhaps when ω_2 is close to $\frac{2\pi}{24} = 0.26$ radians/hr. as well (see Fig. 3B and Fig. 3D).

In summary, with the exception of few rhythmic genes that have a 12hr. period, most of the rhythmic genes have both their primary and secondary frequencies in the vicinity of 24 hr. period with perhaps different phases θ_1 and θ_2. The 12 hour genes are also observed and reported by [10], wherein the temporal responses of two such genes, cce_1889 and cce_3226 have been plotted. In this paper we do not proceed to isolate 12 hour genes but isolate genes with a significant 24 hr. cycle based on the following threshold:

$$|\omega_1 - 0.26| < 0.1 \text{ and } |\alpha_1| > 0.2. \tag{4}$$

Using the criterion (4), we isolate approximately 1513 genes for the light condition DLDL and 1549 genes for the light condition LDLL to have a significant rhythm with 24 hr. period. These two numbers should be compared with the number 1445, which is the number of genes isolated to have the same rhythm in [38]. In essence, about one third of a total of 4600 genes have a significant rhythm with 24 hr. period. The genes isolated are parsed according to the 'Biological Processes' they belong to, and the result is shown in Table 1. We show a total of 15 functional categories and the corresponding number of participating genes with a significant rhythm of 24 hr. period. In the next section, we analyze four of these processes in detail (indicated by a * in Table 1) and compute the phases of the associated participating genes.

4 Phase Responses of Functionally Connected Genes

In the last section 3, we have isolated rhythmic genes with a primary period of 24 hr. ($\omega_1 = 0.26$), from a total of 15 functional groups (see Table 1). In this section, we concentrate on their 'phases'. To save space, we would concentrate only on 4 of the 15 functional categories (indicated in Table 1 by a *) and closely look at their functional subcategories. Our first goal is to associate a phase to every gene with a 24 hr. period that has already been isolated using the threshold (4). If $g(t)$ is the expression of one such gene, we define

$$\bar{g}(t) = g(t) - (a + bt)$$

as in (2) and represent

$$\bar{g}(t) = \alpha_1 \sin(0.26t + \theta) + error(t).$$

The parameter θ is the phase associated with $g(t)$. Alternatively, we can write

$$\bar{g}(t) = (p_1 \quad p_2) \begin{pmatrix} \sin(\omega t) \\ \cos(\omega t) \end{pmatrix} + error(t)$$

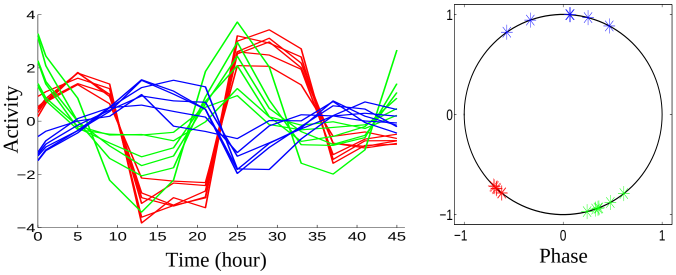

Fig. 4: The expressions of a few candidate genes (Left) from three different functional subcategories (indicated in different colors) and their phase vectors mapped on the unit circle (Right). Each point on the circle represents a gene with its corresponding phase based on the whole temporal expression pattern.

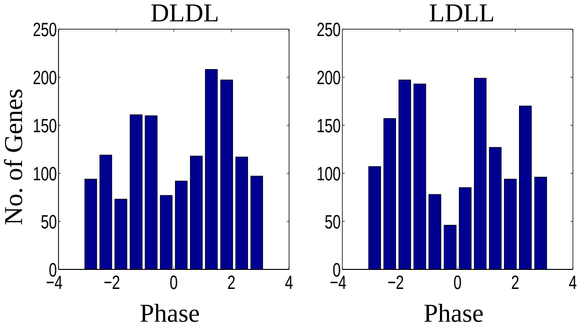

Fig. 5: A histogram shows the number of genes as a function of the phase angle $\theta \in [-\pi, \pi]$ on the unit circle. The left histogram is for the light pattern DLDL and the right histogram is for the light pattern LDLL.

where $\omega = 0.26$. The angle that the vector $(p_1 \quad p_2)$ makes with respect to the positive x-axis is precisely the phase angle θ and we would call the normalized vector

$$p = (p_1 \quad p_2)/(p_1^2 + p_2^2)$$

to be the unit phase vector associated with $g(t)$. Note that the phase vector is a point on the unit circle and we have mapped the expression function $g(t)$ onto this point. In Fig. 4 we show expressions of a few candidate genes (Left) and their phase vectors mapped on the unit circle (Right). Each point on the circle represents a gene with its corresponding phase based on the whole temporal expression pattern.

For each of the 1513 genes (see column **I** in Table 1), isolated from the microarray experiment DLDL, the corresponding phase vectors are computed and mapped onto the unit circle. In Fig. 5 (Left), a histogram is plotted for the number of genes as a function of the phase angle $\theta \in [-\pi, \pi]$. Likewise, the corresponding phase vectors of the 1549 genes (see column **II** in Table 1), isolated from the microarray experiment LDLL, are computed and mapped on the phase circle. In Fig. 5 (Right), the histogram plot is repeated for the light pattern LDLL. In Fig. 6, the activities of the 1513 genes from the experiment DLDL have been plotted as a function of time. The activities have been color coded, and Red indicates High and Blue indicates Low. All panels in the figure carry the

Phase Distribution Density of Cycling Genes under light conditions DLDL/LDLL

Functional SubCategory	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	Total No.
CIM: Nitrogen fixation(16)	0/0	13/0	0/0	1/0	0/0	0/0	0/0	0/12	0/1	0/1	0/0	0/0	14/14
CIM: Polysaccharides and glycoproteins(23)	1/1	0/0	0/1	2/5	1/0	0/0	0/0	1/0	2/0	5/2	3/1	1/0	16/10
EM: Amino acids and amines (14)	1/0	0/0	0/2	2/0	0/0	2/0	0/0	0/2	2/1	0/1	0/1	2/0	9/7
EM: Pentose phosphate pathway(10)	0/0	0/1	0/0	5/1	1/0	0/0	0/0	0/0	1/0	0/2	0/3	1/1	8/8
EM: Sugars(24)	2/0	0/2	0/0	1/3	0/0	0/0	0/1	2/1	1/1	1/2	0/1	2/0	9/11
EM: TCA cycle(10)	0/0	1/0	1/0	3/0	1/0	0/0	0/0	0/0	0/2	0/1	0/3	0/0	6/6
EM: Glycolysis(22)	0/0	1/0	2/0	4/1	1/0	1/0	0/0	0/1	0/0	1/0	0/5	0/1	10/8
EM: Pyruvate and acetyl-CoA metabolism(12)	0/1	0/0	0/0	1/1	1/0	1/1	0/0	0/0	1/2	0/0	2/2	0/0	6/7
PSR: Phycobilisome(17)	0/1	0/5	0/3	1/0	0/0	0/1	0/0	0/1	0/0	4/0	8/0	1/0	14/11
PSR: Photosystem II(27)	0/1	0/3	0/9	1/8	0/2	0/0	1/0	0/1	7/1	12/0	3/1	0/0	24/26
PSR: Soluble electron carriers(17)	0/0	0/2	1/1	5/3	1/0	0/0	0/0	0/1	3/1	2/1	1/3	0/1	13/13
PSR: NADH dehydrogenase(24)	0/5	0/3	1/0	0/1	2/1	4/0	2/0	3/0	1/2	4/0	0/2	0/3	17/17
PSR: CO_2 fixation(16)	0/0	0/1	2/1	0/5	1/2	0/0	0/2	0/0	0/1	2/1	6/1	3/0	14/14
PSR: Photosystem I(16)	0/5	0/1	0/0	0/0	0/0	0/0	0/1	4/0	5/0	2/0	1/0	2/2	14/9
PSR: ATP synthase(15)	0/0	0/0	0/0	0/9	2/1	0/0	0/0	0/0	0/0	0/0	9/0	1/0	12/10
TL: Ribosomal proteins(52)	17/0	21/1	1/4	1/10	0/9	0/2	0/1	0/0	0/0	0/0	1/0	2/0	43/27
TL: Degradation of proteins, peptides, and glycopeptides(40)	1/3	0/7	1/2	2/0	2/0	1/0	0/1	3/3	7/0	1/0	0/2	0/1	18/19
TL: Aminoacyl tRNA synthetases and tRNA modification(48)	0/2	2/2	1/1	2/2	1/0	1/0	0/0	1/3	2/0	1/1	0/3	1/0	12/14
TL: Nucleoproteins(15)	0/1	3/0	0/2	0/2	0/0	1/1	0/0	0/0	0/0	2/0	1/1	2/0	9/7
TL: Protein modification and translation factors(35)	1/1	0/1	1/3	1/3	1/0	0/0	0/0	1/1	3/2	1/0	1/0	6/0	16/11

TABLE 3: Phase distribution density of significant cycling genes of 24 hr. period under light conditions DLDL/LDLL. I_k-s, $k = 1, 2, \cdots, 12$, denote phase intervals of equal width, from $[-\pi, \pi]$ along the counter clockwise direction. For a specific metabolic process, the two numbers a/b refer to the number of genes that have a phase response in the corresponding phase interval under DLDL/LDLL respectively. The abbreviations of functional categories refer to Table 2.

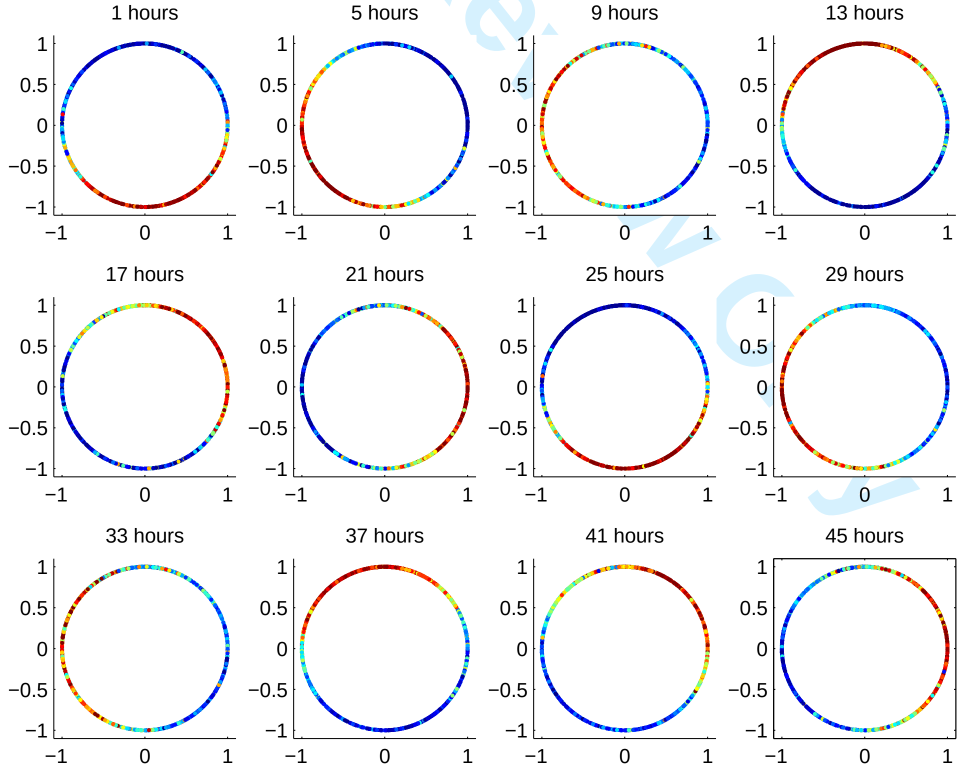

Fig. 6: The expressions of cyclic genes from the microarray experiment DLDL are plotted over time. The genes are mapped on the circle using their corresponding phase responses, which does not change with time. The gene expressions over time are encoded in color – Red stands for a high level of activity and Blue stands for a low level of activity. The genes that have close phases, peak at the same time. The expressions peak in turn as time goes on, along the 'phase circle' in a clockwise direction.

same phase information of the genes since the phase vectors are computed using the whole temporal patterns, and each panel carries the corresponding activity values of those genes at the certain time (1 hours, 5 hours, ..., 45 hours) using the color coding. Each gene has been mapped to the same phase position on all circles with different colors that indicate the various expression values over time. From this figure we learn that, by and large, a group of genes are highly active at any given time (indicated by a block of red dots on the circle) and that this group is clustered in a certain region of the circle. We would now like to argue that genes associated with a functional subcategory are often clustered on the phase circle and we would now like to display these clusters. In Fig. 4, the candidate genes are selected from three different functional subcategories (indicated in different colors) and it is clearly visible that they are clustered on the unit phase circle. Let $I_k, k = 1, 2, \cdots, 12$, denote 12 equispaced phase intervals from $[-\pi, \pi]$ along the counter clockwise direction. In Table 3 we consider 4 of the 15 functional categories described in Table 1 and their subcategories. For each interval I_k, we count the number of genes in a subcategory whose expressions have phases in the interval I_k. We repeat this for $k = 1, 2, \cdots, 12$ and for each of the two experiments with light patterns DLDL and LDLL. In Table 3, we have displayed 20 of the total of 28 possible functional subcategories associated with the 4 functional categories considered. The 8 subcategories have been eliminated because they have a small number of genes with a 24 hr. period. For the light pattern DLDL, a total of 284 genes with 24 hr. period belong to the 20 functional subcategories displayed in Table 3. For the other experiment LDLL, this number is given by 249 genes. Thus only a small percentage ($\approx 12\%$) of genes drop out as a result of a changed light condition. A closer examination of Table 3 reveals that for 11 of the 20 subcategories, the phase variables of the associated genes are localized to within a subset of the unit circle. This has been displayed in Fig. 7 for both the experiments with light patterns DLDL and LDLL. This indicates that for 11 of the 20 sub processes, the participating genes are active only at a specific time of the diurnal cycle. The pattern observed in Fig. 6 is primarily due to the genes of these 11 sub processes expressing themselves in a certain sequence for the data set DLDL. Interestingly, these 11 sub processes correspond to respectively 66% and 62% of the total number of participating genes in Table 3. In Table 4 we show the phase span of the participating genes for each of the 11 sub processes and compare the two experiments, DLDL and LDLL. In this table we observe that 7 of the 11 processes, viz CIM:NF, CIM:PG, EM:PPP, EM:GL, PSR:PS II, PSR:CF and TL:DPPG, the phase spans of the participating genes overlap. In 3 others, viz PSR:PB, PSR:PS I and PSR:ATP, the phase spans do not overlap but are almost adjacent to each other (i.e. separated by 1 hour). In only one process, TL:RP, the phase spans are not adjacent (i.e. apart by 3 hours).

In this section, we show that 4 processes have about 10 circadian controlled sub processes having a preferred time of activity which remains unaltered by changes in the light condition. Each of the other sub processes either do not have enough participating genes or do not have a preferred time of activity that remains conserved under changing light condition.

Functional Sub Category	Phase Distribution in Terms of Time	
	DLDL	**LDLL**
CIM: NF	D5 – D7	D6 – D8
CIM: PG	D11 – L5	L2 –L4
EM: PPP	D1 – D3	L12 – D4
EM: GL	D1 – D5	L12 – D2
PSR: PB	D11 – L3	L4 – L8
PSR: PS II	L1 – L5	L2 – L8
PSR: CF	D11 – L3	D12 – L4
PSR: PS I	L3–L7	L8–L12
PSR: ATP	D11–L1	L2–L4
TL: RP	D5 –D9	D12–L6
TL: DPPG	L3 –L7	L4–L10

TABLE 4: Phase Distribution in terms of the time of the day, for light patterns DLDL and LDLL. The abbreviations of functional categories and subcategories refer to Table 2.

5 A Prototype Model of the Circadian Genes with Phase Oscillators

So far in this paper we have observed that restricted to genes with expressions with a 24 hr. period, there are many sub processes with a specific time of peak activity during the diurnal cycle. Even under changing light condition, the phases of these sub processes do not change appreciably. This indicates that perhaps these sub processes are coupled to an internal clock, ie they are circadian controlled. The precise biochemical coupling between an internal clock and genes participating in a circadian controlled process is poorly understood. What is well understood is that in Cyanobacteria, a circadian clock is sustained by three interacting proteins, KaiA, KaiB and KaiC (See [14], [16], [24]), with KaiC showing autophosphorylation in the presence of ATP. KaiC is an enzyme with autokinase and autophosphatase activities. KaiA enhances KaiC function while KaiB diminishes the effect of KaiA on KaiC ([17]). Given the dual function of KaiC and cooperation between KaiA and KaiB, it has been suggested that autonomous oscillation of KaiC phosphorylation might be achieved [29]. Several mathematical models have been proposed recently to account for the molecular mechanisms involved in the circadian phosphorylation process of Kai proteins by assuming different reaction schemes (see [8], [11], [22], [24], [25]).

In one approach, a dynamical model was built ([24]) and it was concluded that the KaiA and KaiB-assisted autocatalytic phosphorylation and dephosphorylation of KaiC are the source for circadian rhythmicity. The autocatalytic activity induces a positive feedback on the KaiC phosphorylation process and results in synchronization at the highest phosphorylation level of the hexamers. The oscillatory activities of unphosphorylated KaiC and phosphorylated KaiC (KaiC*) simulated using the model form a limit cycle (shown in Fig. 8). The topology of this approach has a deficit in robustness against common changes in protein levels, which has been overcome by another approach in which a negative feedback loop was induced by sequestration of KaiA proteins([8]). In this section, we

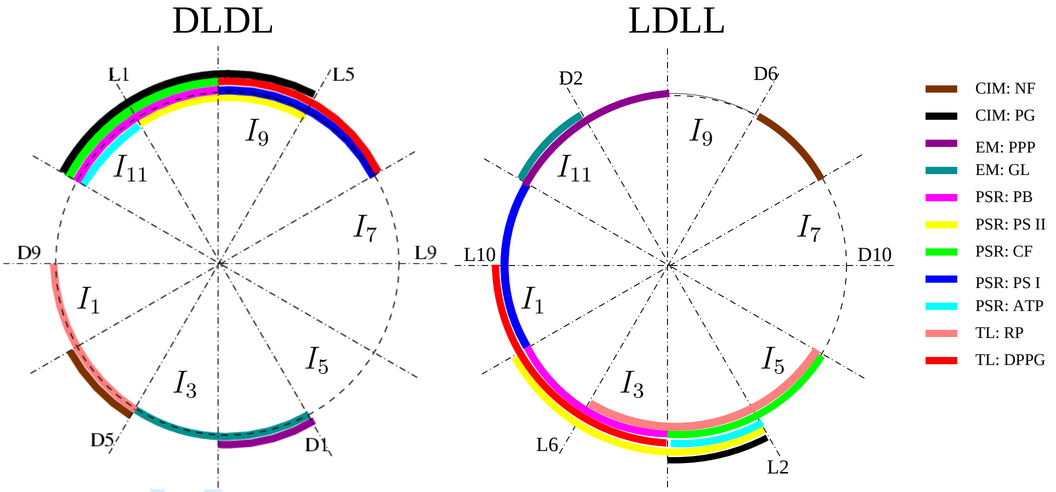

Fig. 7: Phase spans of genes that belong to 11 functional subcategories under light patterns, DLDL (Left) and LDLL (Right) respectively. The figure shows that these subcategories have genes with phases localized within a small interval of the unit circle indicating that they have a preferred time of activity. In the left figure, D1, D5 and D9 refer to the 1^{st}, 5^{th} and 9^{th} hours of the dark cycle, L1, L5 and L9 refer to the 1^{st}, 5^{th} and 9^{th} hours of the light cycle. The Di and Li, $i = 2, 6, 10$ in the right figure also mean the same except that for the second 24 hr., D2, D6 and D10 have to be replaced by L14, L18 and L22. The abbreviations of functional categories and subcategories refer to Table 2.

synthesize a phase oscillator from the dynamic model of the circadian clock ([24]) using its limit cycle behavior (Fig. 8) despite the deficit in robustness, since our focus in this study is to construct a phase model of oscillator network. We call the derived oscillator the 'master clock'. We propose an oscillator network with the master clock and three other peripheral oscillators (shown in Fig. 10) and study to what extent the expressions of genes in the 11 sub processes are approximated by the network proposed. Finally we also study to what extent the oscillator network is robust with respect to changes in the frequencies of the peripheral oscillators. We conclude that many circadian controlled genes are robustly approximated by the proposed network.

5.1 A phase model of the circadian clock

A dynamic model of the master clock proposed in [24] is given by

$$\begin{aligned}
\dot{x}_1 &= k_5x_8 - k_1x_1x_3 - k_3x_6x_1x_3\\
\dot{x}_2 &= k_6x_7 - k_4x_6x_2\\
\dot{x}_3 &= k_7x_4 - k_1x_1x_3 - k_3x_6x_1x_3\\
\dot{x}_4 &= k_6x_7 - k_7x_4\\
\dot{x}_5 &= k_1x_1x_3 - k_2x_5\\
\dot{x}_6 &= k_2x_5 + k_3x_6x_1x_3 - k_4x_6x_2\\
\dot{x}_7 &= k_5x_8 - k_6x_7\\
\dot{x}_8 &= k_4x_6x_2 - k_5x_8
\end{aligned}$$

where $x_1, \ldots, x_8$ represent KaiA, KaiB, KaiC, KaiC*, KaiAC, KaiAC*, KaiBC* and KaiABC* respectively. The state variable x_3 (unphosphorylated KaiC protein) and x_4 (phosphorylated KaiC protein KaiC*) are of great interest because they are oscillatory and converge to a limit cycle on the phase plane. Fig. 8 shows the oscillatory activities of KaiC and KaiC*. In Fig. 8 (Right) the associated limit cycle has been sketched after translating the plot in such a way that the cycle encircles the origin. There are of course several ways of doing this and

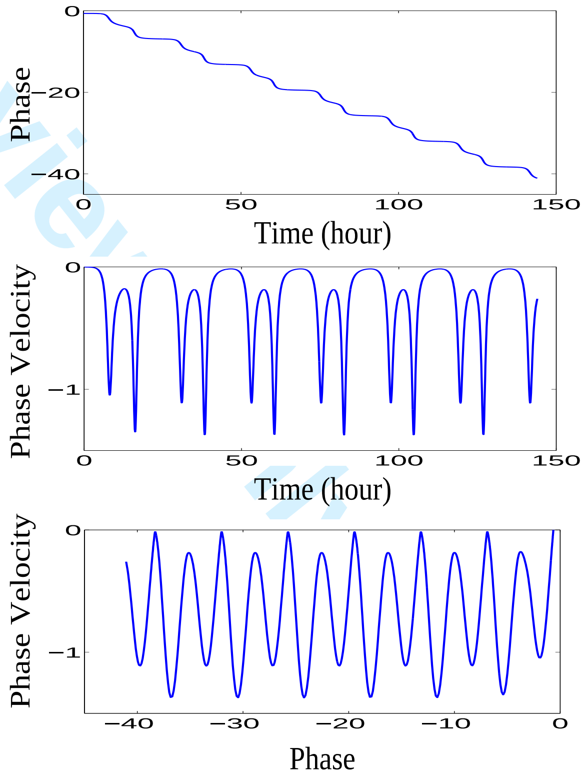

Fig. 9: The phase pattern of the master clock (Top) and its velocity pattern (Middle) as a function of time. The phase velocity has been plotted as a function of phase (Bottom).

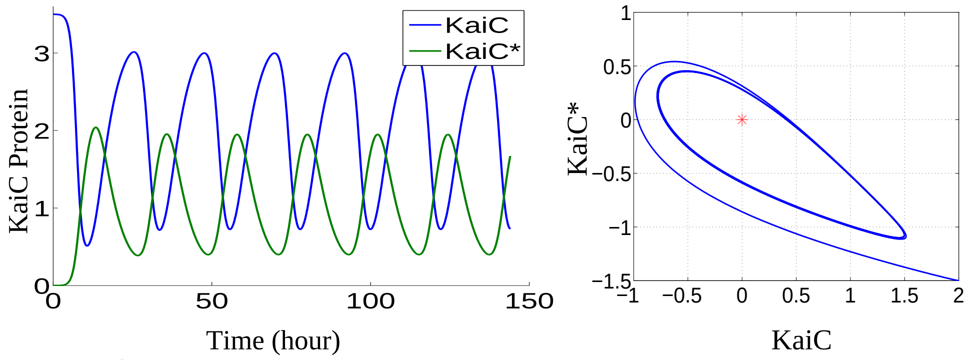

Fig. 8: The oscillatory activities of unphosphorylated KaiC and phosphorylated KaiC (KaiC*) in time (Left) and the limit cycle behavior around the origin on the phase plane by taking a proper linear transformation (Right).

one canonical way to translate the limit cycle is to ensure that the maximum and minimum values of KaiC and KaiC* are equal in magnitude. In Fig. 8 however, the limit cycle has been translated so that for KaiC the ratio of minimum to maximum values is at $1 : 2$, and for KaiC* this ratio is at $2 : 1$. The specific choices of these numbers are arbitrary.

We can now attach a phase variable ϕ to every point on the limit cycle and plot ϕ and $\dot{\phi}$ as a function of time and $\dot{\phi}$ as a function of ϕ (see Fig. 9). Our eventual goal is to write $\dot{\phi}$ as a Fourier series given by

$$\dot{\phi} = \omega + \sum_{m=1}^{N} (a_m \sin(m\phi) + b_m \cos(m\phi)) \tag{5}$$

where a_m and b_m are real numbers. The equation (5) describes the phase dynamics of the circadian clock derived from the limit cycle oscillation of KaiC/KaiC* shown in Fig. 8. It was observed that for $N = 30$, the solution of the ordinary differential equation (5) matches with the phase function sketched in Fig. 9 quite closely for a suitable choice of ω and $a_i, b_i, i = 1, \cdots, 30$.

5.2 An oscillator network model of the circadian sub processes

The oscillator network model, sketched in Fig. 10, is described by the following sets of equations:

$$\begin{array}{rcl}
\dot{\phi}_0 & = & \omega_0 + \sum_{m=1}^{N} (a_m \sin(m\phi_0) + b_m \cos(m\phi_0)) \\
\dot{\phi}_1 & = & \omega_1 + 0.05c \sin(-\phi_0 - \phi_1) \\
 & & +0.05 \sin(\phi_2 - \phi_1 + \frac{2\pi}{3}) \\
 & & +0.05 \sin(\phi_3 - \phi_1 + \frac{4\pi}{3}) \\
\dot{\phi}_2 & = & \omega_1 + 0.05c \sin(-\phi_0 - \phi_2 - \frac{2\pi}{3}) \\
 & & +0.05 \sin(\phi_1 - \phi_2 - \frac{2\pi}{3}) \\
 & & +0.05 \sin(\phi_3 - \phi_2 + \frac{2\pi}{3}) \\
\dot{\phi}_3 & = & \omega_1 + 0.05c \sin(-\phi_0 - \phi_3 - \frac{4\pi}{3}) \\
 & & +0.05 \sin(\phi_1 - \phi_3 - \frac{4\pi}{3}) \\
 & & +0.05 \sin(\phi_2 - \phi_3 - \frac{2\pi}{3})
\end{array} \tag{6}$$

In (6), ϕ_0 is the phase variable of the master clock and the parameters are set to what has been described in (5). Thus ϕ_0 follows the phase of the KaiC/KaiC* oscillation as sketched in Fig. 9. The variables ϕ_1, ϕ_2 and ϕ_3 are phase variables of three peripheral oscillators that are modeled after what was originally proposed by Kuramoto [20], [21]. The parameter ω_1 is set to a nominal value of 0.26 (i.e. 24 hr. period). The parameter c describes the strength factor of the connection between the master clock and the three peripheral oscillators and is an important tuning parameter that we adjust optimally.

If we assume that $g_j(t)$ is the expression of the j^{th} gene, we augment the dynamics (6) with

$$g_j(t) = \sum_{i=1}^{3} \beta_{ji} \sin \phi_i(t), \tag{7}$$

where we define $\beta_j = (\beta_{j1}, \beta_{j2}, \beta_{j3})$ and call the vector β_j the 'output vector' for the j^{th} gene. When the parameter c is large, the master clock dominates and all the peripheral oscillators follow the clock with an appropriate phase difference. When the parameter c is small, each of the peripheral oscillators have a pure tone with frequency ω_1. The parameter ω_0 is assumed to be close to the 24 hr. period whereas ω_1 could drift away in a small neighborhood of the 24 hr. period. The connection strength factor c is adjusted so the peripheral oscillators have a period close to 24 hr. even at the cost of a 'non pure tone'. In Fig. 11 we show that when ω_0 and ω_1 are close, there is no particular advantage in choosing a large value of c. In fact, as shown by the Red Curve, the error goes up because the shape of the phase oscillation deviates from the pure tone. On the other hand, when ω_1 does not match the precise frequency of 0.26, as shown by the Blue Curve, there is an optimum connection strength factor when the error is minimal. The two curves in Fig. 11 has been plotted for a single gene. The design question we discuss in the next section is

"How to optimally choose the parameters β_j and c of the network model (6), (7) for the entire collection of genes with expression frequency close to 0.26 isolated using threshold criterion (4)?"

6 Tuning the Parameters in the Network Model

The network of oscillators that we have proposed in Fig. 10 has the property that if the connection strength parameter c is set to zero, i.e. if the master clock has no influence on the oscillators, then the oscillators maintain a frequency of ω_1

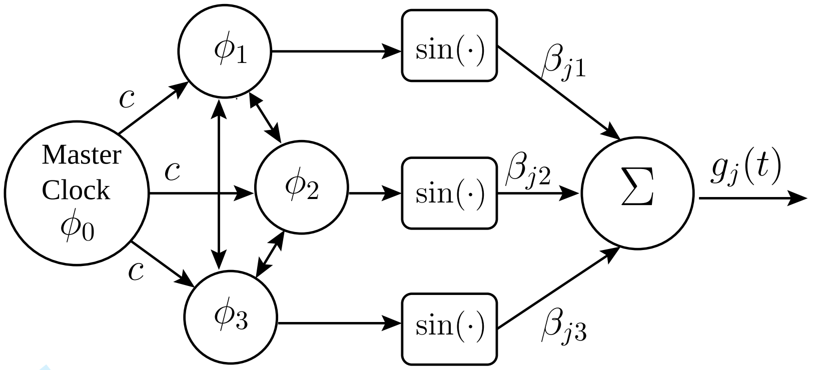

Fig. 10: The oscillatory network including one master clock (ϕ_0) and three Kuramoto type peripheral oscillators ϕ_1, ϕ_2, ϕ_3. The phase variables of the peripheral clocks are used to reproduce expression patterns of circadian controlled genes.

with an appropriate phase difference that has been set to 120 degrees. Nominally, ω_1 is set to the value 0.26 which is what one requires to maintain a period of 24 hours. In this case the model is able to precisely reconstruct every gene with a frequency given by 0.26 and with an arbitrary phase. As is evident from Fig. 3, the gene expressions for DLDL and LDLL are not exactly of frequency 0.26, and even though we filter those genes that satisfy the condition (4), the genes we pick do not all have a period of 24 hours. Hence the proposed model is able to approximate only a subset of the circadian controlled genes that we have isolated.

Additionally we assume that the parameter ω_1 deviates from the nominal value of 0.26, i.e. deviates from the 24 hr. cycle. In this paper we have assumed that ω_1 is perturbed in the set

$$\{0.26 \pm 0.02k\},\ k = 1, 2, 3, 4 \text{ and } 5, \qquad (8)$$

which corresponds to deviation of period in the interval $[17.5, 39.3]$. This could be as a result of a random drift in the parameter ω_1 or perhaps as a result of a drift in the diurnal cycle, changed light pattern for example. It is under this situation that the internal clock, which is assumed to robustly maintain a period of 24 hours, comes to the rescue. Using a single gene as an example, we show in Fig. 11 that, in this situation, a suitable value of the connection strength parameter c actually pulls the frequencies of the three oscillators close to the nominal value of 0.26. The price paid is that the phase variables do not oscillate with a pure tone – because the master clock does not oscillate with a pure tone. In summary, when there is an uncertainty of the two kinds described above, it is not possible to precisely reproduce the gene expressions with the model. The connection strength c and the output vector β_j need to be chosen for the j^{th} gene appropriately to minimize the error. Note that the parameter c is independent of j and also does not depend on the uncertain values of ω_1, the frequency of the peripheral oscillator. We now describe how these parameters are chosen.

We start with a priori assumption that c is known as an element of an interval. We choose seven different values of c (from the set $\{20, 25, 30, 35, 40, 45, 50\}$) and for each value of c, we compute the optimum value of β_j for each gene expression assuming that ω_0 and ω_1 are set to the nominal value of 0.26. All the vectors β_j, viewed as row vectors, are stacked together in a column to define a matrix which would be called the output matrix M_c. We now keep the optimal value of the output matrix fixed and vary c in the above interval while computing the error percentage averaged over all the uncertain values of ω_1 taking values in the set (8). The total of average error percentage is plotted (see Fig. 12) as a function of the connection strength c for each possible choice of the optimal output matrices. All the seven curves are shown in Fig. 12, wherein the optimal output matrix M^* corresponds to the 'solid green curve' (where $\mathrm{M}^* = \mathrm{M}_{25}$) and the value of the connection strength, indicated by a green star, is chosen as 35.

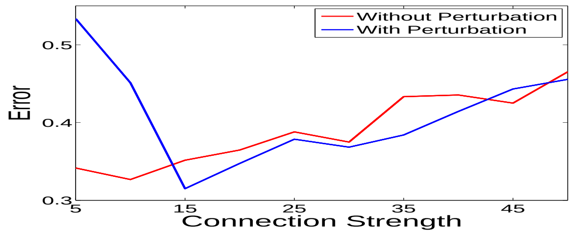

Fig. 11: The estimation error percentage as a function of the connection strength of the network computed for a single gene. The red curve is when the parameters ω_0 and ω_1 are close to the 24 hr. period. Increasing c has no particular advantage. The blue curve is when the parameter ω_1 has been perturbed to a 39.3hr period. The Figure shows that increasing c to a value close to 15 pulls the period close to 24 hours, giving rise to a small error.

We remark that the output matrix is optimized only for the nominal value of ω_1 $(= 0.26)$. The connection strength parameter 'c' minimizes the average error over all the uncertain values of ω_1. A somewhat more difficult optimization problem, where 'output matrix' and 'connection strength' parameters are simultaneously chosen to optimize the average error has not been discussed in this paper. We have also not considered perturbations in ω_0 and assumed that the master clock maintains its rhythm quite accurately.

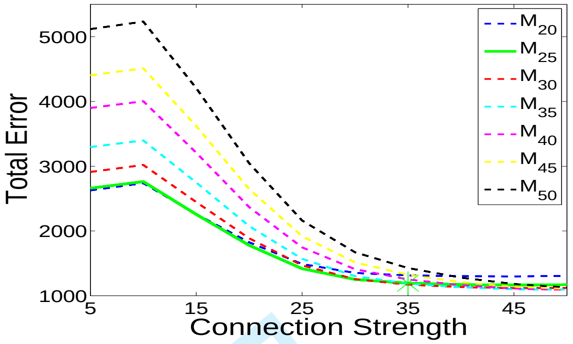

Fig. 12: The total of average error percentages of 1513 circadian controlled genes from the experiment DLDL over frequency perturbation under each connection strength factor for every candidate output matrix (7 curves). The solid green curve is associated with the optimal output matrix M_{25} and the green star points the optimal connection strength factor $c = 35$ of the network.

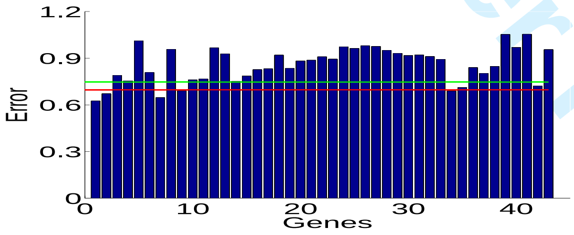

Fig. 13: The estimation error percentages of gene expressions in the sub process *Translation: Ribosomal Protein* for the experiment DLDL with the model of oscillatory network. The red line is the mean error percentage μ averaged over the 11 sub processes and the green line is the threshold value $\mu + 0.4\sigma$.

Number of genes well estimated

Functional	DLDL			LDLL		
SubCategory	Total NO.	g*	$\hat{g}$	Total NO.	g*	$\hat{g}$
CIM: NF	14	13	13	14	14	14
CIM: PG	16	14	13	10	9	10
EM: PPP	8	7	7	8	8	7
EM: GL	10	7	7	8	6	8
PSR: PB	14	12	13	11	4	2
PSR: PS II	24	18	16	26	22	23
PSR: CF	14	12	11	14	11	11
PSR: PS I	14	2	8	9	6	9
PSR: ATP	12	12	12	10	9	10
TL: RP	43	12	7	27	15	22
TL: DPPG	18	12	7	19	12	16
Total	187	121	114	156	116	132

TABLE 5: Number of genes in each of 11 sub processes that are well estimated with respect to g^* and $\hat{g}$ under both light conditions, DLDL and LDLL.

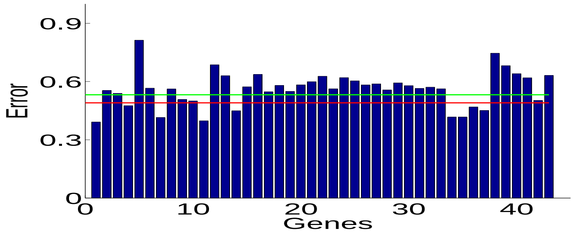

Fig. 14: The estimation error percentages of gene expressions in the sub process *Translation: Ribosomal Protein* for the experiment DLDL using the functions $g_j^*(t)$. The red line is the mean error percentage μ^* averaged over the 11 sub processes and the green line is the threshold value $\mu^* + 0.4\sigma^*$.

7 Analysis of the Model in Capturing the Expressions of Circadian Genes

The parameters of the network model have been chosen in such a way that the output vector β_j is optimum provided that the frequencies of the oscillators are close to the 24 hr. period. The connection strength factor c is optimized so that the model reproduces the expressions even when the oscillator frequencies fluctuate. The model was tested for each of the 187 (DLDL) and 156 (LDLL) genes from the 11 sub processes under two light conditions described in Table 4. We recall from Section 4 that 10 of these 11 sub processes (i.e. with the exception of TL: RP) do not change its phase activity even when the light condition is changed.

For every gene expression $g_j(t)$ we obtain $\hat{g}_j(t)$ the corresponding reconstruction of the expression using the model (6), (7), assuming optimal parameter values. We compute the error percentage $\|g_j(t) - \hat{g}_j(t)\| / \|g_j(t)\|$ and the corresponding mean μ and standard deviation σ over all the genes selected. We hypothesize that if the error is within the threshold $\mu + 0.4\sigma$, then the model faithfully reproduces the corresponding gene expression. Fig. 13 shows the estimation error percentage of 43 gene expression in the sub process *Translation: Ribosomal Protein* (TL:RP) from the experiment DLDL (we specifically chose TL:RP for illustration because the model could capture only 7 of the 43 genes in this sub process and we wanted to know why). The red line is the mean error percentage μ averaged over all selected gene expressions and the green line is the threshold value $\mu + 0.4\sigma$. In Table 5, under the column $\hat{g}$ we show the number of genes that are within the above threshold for each of the 11 sub processes and for each of the two experiments DLDL and LDLL. With the exception of PSR:PSI, TL:RP, TL:DPPG for DLDL and PSR:PB for LDLL, for each of the other sub processes most of the genes are well approximated by the model. We conclude that a simple ‘three oscillator model synchronized by a single master clock’ was sufficient to model fairly large number of sub processes that we believe are ‘circadian controlled.’

We were interested in analyzing what was the cause for the model to fail for some of the genes, for example TL:RP in the experiment DLDL. For every gene expression $g_j(t)$, we computed $g_j^*(t)$, the best reconstruction using functions precisely of frequency 0.26 and arbitrary phase and amplitude. Although the gene expressions $g_j(t)$ had their primary frequencies close to 0.26, for some sub processes they were not close enough. We computed the error percentage $\|g_j(t) - g_j^*(t)\|/\|g_j(t)\|$, the corresponding mean μ^* and standard deviation σ^* over all the genes in the 11 sub process. The estimation error percentage of gene expressions using g_j^* in the sub process TL:RP under DLDL have been plotted in Fig. 14 which clearly shows that most of the genes here were not closely reproduced by the model because their frequencies were too far from the 24 hr. period. The red line is the mean error percentage μ^* and the green line is the threshold value $\mu^* + 0.4\sigma^*$. In Table 5, under the column g^* we show the number of genes that are within a threshold of $\mu^* + 0.4\sigma^*$. We find that in all the cases where a sub process was poorly reproduced, the error $\|g_j(t) - g_j^*(t)\|/\|g_j(t)\|$ was not within threshold. We conclude that the performance of the model was closely tied to the frequency of the genes to be approximated. Hence a tighter threshold (4) would have improved the performance of the model.

We remark that choosing a threshold of the form ‘mean + .4 std. dev.’ is a bit unconventional, instead of simply choosing the threshold to be a small real number. Our motivation to choose this form of threshold is to ensure that the model accepts an average circadian gene expression. Only when the error percentage exceeds the average by more than ‘.4 std. dev.’, the model rejects the corresponding gene.

8 Discussion

In this section we would like to compare some of the main results of this paper with those reported in [10], [38]. These two papers refer to the problem of isolating genes from *Cyanothece* transcriptome data, with a 24 hr. period and classify them into metabolic processes and sub processes. In [38], the peak value of the gene expression was looked into as the time point of maximal activity and genes were clustered based on these time points. The main result was to show that many of the gene clusters came from specific processes indicating that the organism maintains a particular time of activity to perform the corresponding metabolic task. On the other hand, in [10], gene expressions were analyzed based on a suitably defined ‘Fourier Score’ and using these scores, diurnal genes were identified using a ‘p-value’ or alternatively a ‘False Discovery rate’. In this way, the corresponding frequency components were isolated for each and every gene. A gene was declared to be rhythmic if it has a strong single frequency in the neighborhood of 24 hr. period. Two sets of experimental data reported in [38], [40] were compared to show how these rhythmic genes changed their oscillatory patterns when the incident light was altered from an alternating DLDL cycle to LDLL cycle, where ‘L’ stands for Light and ‘D’ stands for Dark. An important result in [10] is to show that certain genes do not change its frequency even when the light pattern is altered and these genes were characterized as ‘circadian controlled genes’. On the other hand there were rhythmic genes in the DLDL data that failed to oscillate under LDLL, and these genes were characterized as ‘light controlled genes’. References to processes that contain a significant number of circadian controlled or light controlled genes were also made in [10], and it was noted that these processes were temporally separated.

In this paper, the rhythmic genes were isolated using a linear combination of a pair of sinusoidal functions with different phases and frequencies and optimum values of these parameters were computed. The main reason why only two frequencies were used is that the sparsity of the data set did not allow us to robustly estimate multiple frequencies, similar to that of a Fourier series representation of a rhythm. We did not start with the assumption that the rhythmic genes have only pure tones. However our analysis in this paper shows that this is indeed the case. In fact, most of the rhythmic genes have a period in the vicinity of 24 hours and only a smaller number of genes have a period close to 12 hours. The 24 hour rhythmic genes were isolated using a threshold around 24 hr. period for the dominant frequency. An important result of this paper is to map every rhythmic genes onto a point on the ‘phase circle’ and cluster genes on this circle using the phase variable. Various metabolic sub processes were looked into with the goal of finding rhythmic genes that maintain a tight cluster. It was found that many sub processes maintain a preferred time of activity, as was indicated in [10], [38], that can be distinguished from the phase plot. It was also discovered that the ‘phase span’ of some of these sub processes remain invariant under a changed light pattern, as evidenced by comparing data from the two experiments. This fact was indicated in [10] but a detailed classification of all the sub processes of four important processes, is one of the main contributions of this paper. As a final contribution, we have a phase model of the circadian genes, not previously reported in the literature.

The rhythmic gene isolation techniques proposed in [10], [38] and this paper are now compared. In [38], the genes are clustered using Pearson Correlation. The basic idea is that if a specific gene expression peaks at a certain time, all other genes that are close would peak around the same time. The technique proposed in [10] is to compare a rhythmic gene with another hypothetical gene obtained by randomizing the data points. This way, one eliminates the possibility of a gene to be categorized as rhythmic, when in fact it appears to be so by chance. The approach in this paper is to first remove the linear trend and subsequently analyze the expression time series in order to find frequency components. No study has been made to combine the three methods detailed here, although it is likely that this can indeed be possible.

To end this section, we would like to comment on the use of a single master clock oscillator driving three peripheral oscillators with a phase difference of 120 degrees between them, as shown in Fig. 10. The basic underlying question is ‘How many master clocks does the observed data set support?’ In *Cyanothece*, we presently have evidence of only one master clock and our model supports this point. We do not claim that the proposed architecture is ‘Biologically Supported’, but would like to emphasize that ‘The proposed phase model derived from a dynamic model of a circadian clock supports the rhythmic data observed in the two experiments, DLDL and LDLL.’

9 Conclusion

The phenomenon of circadian rhythm, observed in *Cyanothece*, has been studied recently in [10], [12], [37], [38] and many others. In this paper, the phase responses of circadian genes associated with particular metabolic processes have been investigated. A prototype model of the oscillatory network has been proposed to estimate the expression patterns of genes with a significant oscillatory component.

With two sets of transcriptome data, collected from *Cyanothece* under two different light conditions, identification of genes with significant diurnal oscillations has been carried out using a single sinusoidal function with a primary frequency and arbitrary phase. Under two different light conditions, the distribution density of this frequency shows that the genes can be isolated into two groups: a larger group of genes that have a dominant frequency of 24 hours and a smaller group of genes that have a dominant frequency of 12 hours. We focused our attention on genes that have a period of 24 hours and computed their corresponding phases. About $\frac{1}{3}$ of the genes in the *Cyanothece* gene pool were selected and the phase responses of these genes were investigated. An important observation that we made as a result, is that the phases of these genes are clustered. In many cases, gene clusters with sufficiently close phase spans come from a specific metabolic sub process. This indicates the existence of a preferred time when these sub processes are 'most active'. Moreover for many sub processes, the preferred time does not depend on the nature of the pattern of the incident light, indicating that the genes in there are circadian controlled, i.e. controlled by an internal clock.

Using a simple model with three peripheral oscillators that are synchronized by an internal master clock, the circadian controlled genes are modeled and their expression patterns are reproduced. The clock model has been derived from the limit cycle oscillation of KaiC/KaiC*. An optimal connection parameter and output matrix for the model were tuned by minimizing the total average estimation error over a set of frequencies of the peripheral oscillators. For most of the circadian controlled genes, the model performs adequately even when the frequency parameter is perturbed. For those gene expressions on which the model fails, we note that the model accuracy depends quite strongly on the frequency of the expression to be approximated.

10 Acknowledgement

Partial supports for this research from NSF FIBR grant 0425749 and from NSF BIC Grant 0523983 are gratefully acknowledged. The authors would like to thank J. Stockel and J. Toepel for sharing their experimental data. The authors would also like to thank T. R. Elvitigala and E. Welsh for discussions and suggestions during this study.

References

[1] S. Aoki, T. Kondo, and M. Ishiura. A promoter-trap vector for clock-controlled genes in the cyanobacterium synechocystis sp. pcc 6803. *J Microbiol Methods*.

[2] K. Arita, H. Hashimoto, K. Igari, M. Akaboshi, S. Kutsuna, M. Sato, and T. Shimizu. Structural and biochemical characterization of a cyanobacterium circadian clock-modifier protein. *Journal of Biological Chemistry*, 282(2):1128–1135, 2007.

[3] D. Bell-Pedersen, V. M. Cassone, D. J. Earnest, S. S. Golden, P. E. Hardin, T. L. Thomas, and M. J. Zoran. Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nature Rev. Genet.*, 6:544–556, 2005.

[4] E. Bunning. *The Physiological Clock*. Springer-Verlag, Berlin, 1973.

[5] E. L. Campbell, M. L. Summers, H. Christman, M. E. Martin, and J. C. Meeks. Global gene expression patterns of *nostoc punctiforme* in steady-state dinitrogen-grown heterocyst-containing cultures and at single time points during the differentiation of akinetes and hormogonia. *J. Bacteriol.*, 189:5247–5256, 2007.

[6] D. J. Chadwick and J. A. Goode. *Molecular Clocks and Light Signalling*. John Wiley & Sons, Ltd, Chichester, UK, 2003.

[7] T. H. Chen, T. L. Chen, L. M. Hung, and T. C. Huang. Circadian rhythm in amino acid uptake by synechococcus rf-1. *Plant Physiol.*, 97:55–59, 1991.

[8] S. Clodong, U. Dühring, L. Kronk, A. Wilde, I. Axmann, H. Herzel, and M. Kollmann. Functioning and robustness of a bacterial circadian clock. *Molecular Systems Biology*, 3:90, 2007.

[9] J. C. Dunlap, J. J. Loros, and P. J. DeCoursey. *Chronobiology, Biological Timekeeping*. Sinauer Associates Inc., Sunderland, MA, 2004.

[10] T. R. Elvitigala, J. Stöckel, B. K. Ghosh, and H. B. Pakrasi. Effect of continuous light on diurnal rhythms in cyanothece atcc 51142. *BMC Genomics*, submitted.

[11] E. Emberly and N. S. Wingreen. Hourglass model for a protein-based circadian oscillator. *Phys Rev Lett.*, 96:038303, 2006.

[12] S. S. Golden, M. Ishiura, C. H. Johnson, and T. Kondo. Cyanobacterial circadian rhythms. *Annu Rev Plant Physiol Plant Mol Biol*, 48:327–354, 1997.

[13] R. Haselkorn. Heterocysts. *Annu. Rev. Plant Physiol.*, 29:319–344, 1978.

[14] M. Ishiura, S. Kutsuna, S. Aoki, H. Iwasaki, C. R. Andersson, A. Tanabe, S. S. Golden, C. H. Johnson, and T. Kondo. Expression of a gene cluster kaiabc as a circadian feedback process in cyanobacteria. *Science*, 281:1519–1523, 1998.

[15] C. H. Johnson. Circadian rhythms: As time glows by in bacteria. *Nature*, 430:23–24, 2004.

[16] H. Kageyama, T. Kondo, and H. Iwasaki. Circadian formation of clock protein complexes by kaia, kaib, kaic, and sasa in cyanobacteria. *J. Biol. Chem.*, 278:2388–2395, 2003.

[17] Y. Kitayama, H. Iwasaki, T. Nishiwaki, and T. Kondo. Kaib functions as an attenuator of kaic phosphorylation in the cyanobacterial circadian clock system. *The EMBO Journal*, 22:2127–2134, 2003.

[18] T. Kondo, T. Mori, N. V. Lebedeva, S. Aoki, M. Ishiura, and S. S. Golden. Circadian rhythms in rapidly dividing cyanobacteria. *Science*, 275:224–227, 1996.

[19] R. E. Kronauer, C. A. Czeisler, S. F. Pilato, M. C. Moore-Ede, and E. D. Weitzman. Mathematical model of the human circadian system with two interacting oscillators. *Am J Physiol Regul Integr Comp Physiol*, 242:3–17, 1982.

[20] Y. Kuramoto. *In international symposium on mathematical problems in theoretical physics, Lecture Notes in Physics*. Springer-Verlag, New York, 1975.

[21] Y. Kuramoto. *Chemical Oscillations, Waves, and Turbulence*. Springer-Verlag, New York, 1984.

[22] G. Kurosawa, K. Aihara, and Y. Iwasa. A model for the circadian rhythm of cyanobacteria that maintains oscillation without gene expression. *Biophys J.*, 91:2015–2023, 2006.

[23] Y. Liu, N. F. Tsinoremas, C. H. Johnson, N. V. Lebedeva, S. S. Golden, M. Ishiura, and T. Kondo. Circadian orchestration of gene expression in cyanobacteria. *Genes Dev.*, 9:1469–1478, 1995.

[24] A. Mehra, C. I. Hong, M. Shi, J. J. Loros, J. C. Dunlap, and P. Ruoff. Circadian rhythmicity by autocatalysis. *PLos Computational Biology*, 2(7):816–823, 2006.

[25] F. Miyoshi, Y. Nakayama, K. Kaizu, H. Iwasaki, and M. Tomita. A mathematical model for the kai-protein-based chemical oscillator and clock gene expression rhythms in cyanobacteria. *J Biol Rhythms.*, 22:69–80, 2007.

[26] T. Mori, B. Binder, and C. H. Johnson. Circadian gating of cell division in cyanobacteria growing with average doubling times of less than 24 hours. *Proc. Natl. Acad. Sci., USA*, 93:10183–10188, 1996.

[27] F. Naef. Circadian clocks go *in vitro*: purely post-translational oscillators in cyanobacteria. *Molecular Systems Biology*, 1:2005.0019, 2005.

[28] Y. Nakahira, M. Katayama, H. Miyashita, S. Kutsuna, H. Iwasaki, T. Oyama, and T. Kondo. Global gene repression by kaic as a master process of prokaryotic circadian system. *Proc Natl Acad Sci*, 101:881–885, 2004.

[29] M. Nakajima, K. Imai, H. Ito, T. Nishiwaki, Y. Murayama, H. wasaki, T. Oyama, and T. Kondo. Reconstitution of circadian oscillation of cyanobacterial kaic phosphorylation in vitro. *Science*, 308:414–415, 2005.

[30] T. Nishiwaki, Y. Satomi, M. Nakajima, C. Lee, R. Kiyohara, H. Kageyama, Y. Kitayama, M. Temamoto, A. Yamaguchi, A. Hijikata, M. Go, H. Iwasaki, T. Takao, and T. Kondo. Role of kaic phosphorylation in the circadian clock system of *synechococcus elongatus* pcc 7942. *Proc Natl Acad Sci*, 101:13927–13932, 2004.

[31] K. Onai, M. Morishita, S. Itoh, K. Okamoto, and M. Ishiura. Circadian rhythms in the thermophilic cyanobacterium thermosynechococcus elongatus: Compensation of period length over a wide temperature range. *J. Bacteriol.*, 186(15):4972–4977, 2004.

[32] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32:496–501, 2002.

[33] R. J. Reddy, J. B. Haskell, D. M. Sherman, and L. A. Sherman. Unicellular, aerobic nitrogen-fixing cyanobacteria of the genus cyanothece. *J. Bacteriol.*, 175:1284–1292, 1993.

[34] S. M. Reppert and D. R. Weaver. Coordination of circadian timing in mammals. *Nature*, 418:935–941, 2002.

[35] A. Sehgal. *Molecular Biology of Circadian Rhythms*. John Wiley & Sons, Inc., Hoboken, NJ, 2004.

[36] P. Shannon and *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13:2498–2504, 2003.

[37] L. A. Sherman, P. Meunier, and M. S. Colon-Lopez. Diurnal rhythms in metabolism: A day in the life of a unicellular, diazotrophic cyanobacterium. *Photosynth. Res.*, 58:25–42, 1998.

[38] J. Stöckel, E. A. Welsh, M. Liberton, R. Kunnavakkam, R. Aurora, and H. B. Pakrasi. Global transcriptomic analysis of *cyanothece* 51142 reveals robust diurnal oscillation of central metabolic processes. *Proceedings of the National Academy of Science*, 105:6156–6161, 2008.

[39] S. H. Strogatz and I. Stewart. Coupled oscillators and biological synchronization. *Scientific American*, 269:102–109, 1993.

[40] J. Toepel, E. Welsh, T. C. Summerfield, H. B. Pakrasi, and L. A. Sherman. Differential transcriptional analysis of the cyanobacterium *cyanothece* sp. strain atcc 51142 during light-dark and continuous-light growth. *Journal of Bacteriology*, 190:3904–3913, 2008.

[41] J. J. Tyson, C. I. Hong, C. D. Thron, and B. Novak. A simple model of circadian rhythms based on dimerization and proteolysis of per and tim. *Biophys J.*, 77:2411–2417, 1999.

[42] A. T. Winfree. *The geometry of biological time*. Springer, New York, 1980.

[43] C. P. Wolk. Heterocyst formation. *Annu. Rev. Genet.*, 30:59–78, 1996.

[44] M. W. Young and S. A. Kay. Time zones: a comparative genetics of circadian clocks. *Nature Rev. Genet.*, 2:702–715, 2001.

Wenxue Wang (Member, IEEE) received the B.Sc. degree in Automatical Control from Beijing Institute of Technology, M.Sc degree in Control Theory from the Institute of Systems Science, Chinese Academy of Sciences, Beijing. China, and M.Sc. and D.Sc. degrees in Systems Science and Mathematics from Washington University, Saint Louis, MO, in 1996, 1999, 2002, and 2006, respectively.

He is currently a Postdoctoral Research Fellow with the Institute for Collaborative Biotechnologies at the University of California, Santa Barbara, CA. His research interests are in the areas of Computational Neuroscience, Neural Networks, Systems Biology, Signal Analysis and Estimation, and Application of Systems Science to Biological Systems.

Bijoy K. Ghosh (Fellow, IEEE) received the B.Tech. degree from Birla Institute of Technology and Science, Pilani, India, in 1977 and the M.Tech. degree from the Indian Institute of Technology, Kanpur, India, in 1979, both in in Electrical and Electronics engineering, and the Ph.D. degree in engineering from the Decision and Control Group of the Division of Applied Sciences, Harvard University, Cambridge, MA, in 1983. From 1983 to 2006, he was a faculty member in the Department of Electrical and Systems Engineering, Washington University, St. Louis, MO, as a Professor, and directed the center for BioCybernetics and Intelligent Systems. Presently he is a Dick and Martha Brooks Regent Professor in the Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX. His current research interests are in Machine Vision, Computational Neuroscience and Bioinformatics.

Himadri B. Pakrasi (Fellow, AAAS) is the George William and Irene Koechig Freiberg Professor of Biology in Arts and Sciences and professor of Energy in the School of Engineering and Applied Science, Washington University in Saint Louis, MO. He was born in Calcutta, India, and received undergraduate and graduate training in physics at the Presidency College and University of Calcutta.

He came to the U.S. to study biology and received a doctorate at the University of Missouri-Columbia in 1984. He has been on the faculty of Washington University since 1987. Pakrasi is a biochemist recognized for his work on photosynthesis and, in particular, on membrane protein complexes in cyanobacteria and plant chloroplasts. He has a keen interest in bridging the differences between the biological and physical sciences, and leads large-scale multi-institutional systems biology projects.

Pakrasi currently serves as the Director of the International Center for Advanced Renewable Energy and Sustainability at Washington University. Pakrasi has been an Alexander von Humboldt Fellow at Munich University, Germany; a Distinguished Fellow at the Biosciences Institute, Nagoya University, Japan; and a Lady David Visiting Professor at the Hebrew University, Jerusalem, Israel.

Pakrasi serves as the Washington University ambassador from the McDonnell International Scholars Academy to Jawaharlal Nehru University, India.