doi: 10.1093/nar/gkw771

Order restricted inference for oscillatory systems for detecting rhythmic signals

Yolanda Larriba[1], Cristina Rueda[1], Miguel A. Fernández[1] and Shyamal D. Peddada[2,*]

[1]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Paseo de Belén 7, 47011 Valladolid, Spain and [2]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences (NIEHS), Alexander Dr., RTP, NC 27709, USA

Received May 12, 2016; Revised July 28, 2016; Accepted August 23, 2016

ABSTRACT

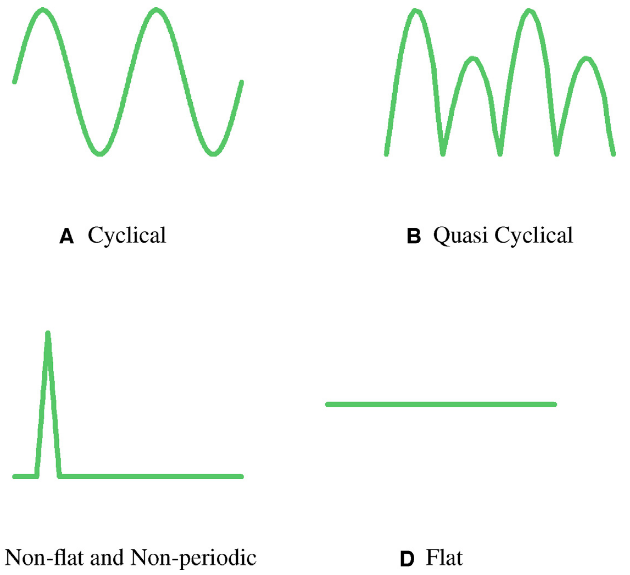
Motivation: Many biological processes, such as cell cycle, circadian clock, menstrual cycles, are governed by oscillatory systems consisting of numerous components that exhibit rhythmic patterns over time. It is not always easy to identify such rhythmic components. For example, it is a challenging problem to identify circadian genes in a given tissue using time-course gene expression data. There is a great potential for misclassifying non-rhythmic as rhythmic genes and vice versa. This has been a problem of considerable interest in recent years. In this article we develop a constrained inference based methodology called Order Restricted Inference for Oscillatory Systems (ORIOS) to detect rhythmic signals. Instead of using mathematical functions (e.g. sinusoidal) to describe shape of rhythmic signals, ORIOS uses mathematical inequalities. Consequently, it is robust and not limited by the biologist's choice of the mathematical model. We studied the performance of ORIOS using simulated as well as real data obtained from mouse liver, pituitary gland and data from NIH3T3, U2OS cell lines. Our results suggest that, for a broad collection of patterns of gene expression, ORIOS has substantially higher power to detect true rhythmic genes in comparison to some popular methods, while also declaring substantially fewer non-rhythmic genes as rhythmic. Availability and Implementation: A user friendly code implemented in R language can be downloaded from http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/peddada/index.cfm. Contact: peddada@niehs.nih.gov

INTRODUCTION

Oscillatory systems arise naturally in biological sciences such as in metabolic cycle (1), cell biology (2–7), endocrinology (8), circadian biology (9–11) and so on. Examples of oscillatory systems include, cell division cycle, circadian clock, hormonal monthly cycle in women and so on. An oscillatory system typically consists of several components that have a rhythmic pattern of expression, i.e. those non-flat patterns that repeated over a fixed period of time. For instance, genes (i.e. components) participating in cell division cycle (i.e. the oscillatory system) follow a rhythmic pattern of expression where the peak expression of a gene corresponds to its biological function (5,6). Often biologists are interested in identifying such genes to understand their functions in the cell division cycle. This paper is motivated by recent interest among pharmacologists and medical doctors to understand circadian rhythms and their role in human physiology, metabolism, and medical treatment etc. Researchers are discovering that numerous health outcomes, such as obesity, production of growth hormones (and abnormal growth patterns) among teenagers and so on. are linked to the sleep patterns and the circadian rhythms (12–14). Recently, Zhang *et al.* (15) discovered that even the efficacy of a drug is related to the time of the day a patient received the drug. To understand circadian clock and its implications on health, there is considerable interest among researchers in identifying and studying genes participating in the circadian clock. Specifically, there is interest in exploring genomic data to identify genes with rhythmic pattern over time in a given tissue (15). When dealing with an oscillatory system consisting of a large number of components, the identification of components that display rhythmic pattern over time is a challenging problem. For example, for a given tissue, the identification of rhythmic genes actively participating in a circadian cycle using a time-course gene expression data is not a simple problem due to (a) variability in time-course expression data, and (b) the absence of a natural flexible parametric model that fits well for a broad collection of rhythmic genes not all of which have a sinusoidal pattern of expression. Challenges in fitting mathematical models such as the Fourier models and other parametric models, especially when less than 50 time points in two or three cycles are available, is well acknowledged in the literature (16–27).

[*]To whom correspondence should be addressed. Tel: +1 919 541 1122; Fax: +1 919 541 4311; Email: peddada@niehs.nih.gov

Published by Oxford University Press on behalf of Nucleic Acids Research 2016.



A Cyclical

B Quasi Cyclical

C Non-flat and Non-periodic

D Flat

Figure 1. Idealized shapes of signals in two time periods of a circadian clock.

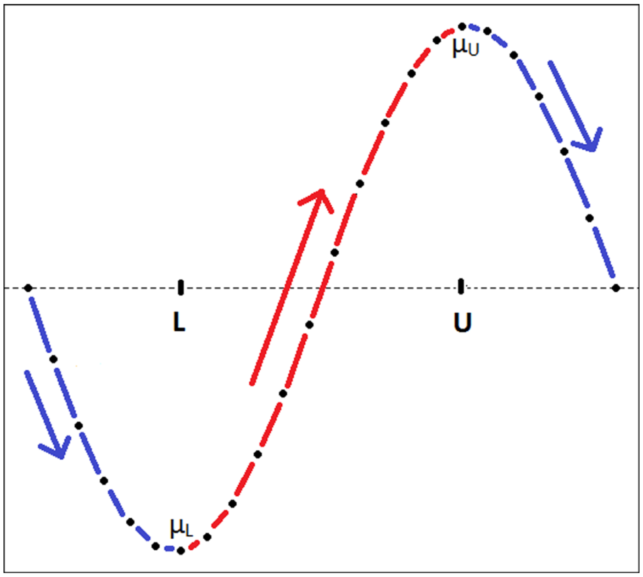


Figure 2. Cyclical signal $\boldsymbol{\mu}$ satisfying a sinusoidal shaped pattern.

Although throughout this paper we focus on circadian clock gene expression data, the methodology discussed in this paper is potentially applicable to other oscillatory systems as well. As an alternative to some of the existing methodologies, in this paper, we develop a method based on order restricted inference (28,29) to classify genes according to various time-course profiles, such as, cyclical, quasi cyclical (these two profiles will be considered as rhythmic), non-flat and non-periodic, or flat (which will be considered as non-rhythmic). For a sample of patterns, see Figure 1. The proposed methodology is non-parametric in the sense that it does not assume a parametric form for the time-course profile, but instead relies on the mathematical inequalities among mean expressions at various time points.

Performance of the proposed methodology, relative to two existing methods JTK_Cycle (JTK) algorithm (22) and RAIN (26), is evaluated using extensive simulation experiments as well as a publicly available circadian clock data at the NCBI GEO database (30).

MATERIALS AND METHODS

Notation and definitions

Without loss of generality, we assume that period of a circadian clock is 24 h long and, as is usually the case, two periods of data are available (31). The methodology described here can be trivially extended to the case when there are more than two periods of data.

For each gene, within each period, we assume that data are available at n different time points. Let $\boldsymbol{Y}_j = (Y_{1j}, \ldots, Y_{nj})'$ denote the vector of gene expressions of a gene at the n time points in the jth period $j = 1, 2$ and let $\boldsymbol{Y} = (\boldsymbol{Y}_1', \boldsymbol{Y}_2')'$. We further assume that the sampling variance is constant at all time points and that the gene expression at each time point follows a normal distribution and the expressions are uncorrelated at all time points. In other words, the observed data are modeled by a signal plus error model $\boldsymbol{Y}_j = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_j$, where $\boldsymbol{\epsilon}_j \sim N_n(0, \sigma^2 I)$ independent and $j = 1, 2$.

A gene is said to be periodic if and only if it satisfies Definition 1, i.e., for any given time point, its mean expression does not change from period to period.

DEFINITION 1. *Periodic signal*

$\boldsymbol{\mu}$ *is said to be periodic* $\iff \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$.

Periodic signal includes a wide range of shapes (some examples are provided in Supplementary Figure S1 in the Supporting Materials). Among these, a common shape of interest to a biologist is the cyclical signal with a unique peak (U) and a unique trough (L) within a period (panel (A) in Figure 1). The time point corresponding to the peak expression represents the time to peak expression of a gene which potentially corresponds to its function in the circadian clock. Several authors, such as (19,32), or (22), modeled $\boldsymbol{\mu}$ using a parametric function of time such as the sinusoidal function. These methods use least squares principle to fit the observed expression data to a sinusoidal curve and identify the best fitting sinusoidal curve. However, from our experience with the cell-cycle and circadian clock data, the sinusoidal function is too rigid and that the real circadian clock data, although cyclical, need not have a perfect sinusoidal signal depicted in Figure 2. For examples of figures that are cyclical but not perfectly sinusoidal see Figure S2 in the Supporting Materials. For this reason, we expand the class of sinusoidal shaped functions to a class of non-parametric cyclical shaped signals where the shape is entirely described by the mathematical inequalities among the components of $\boldsymbol{\mu}$ as follows:

DEFINITION 2. *Cyclical signal*

$\boldsymbol{\mu}$ *is said to be cyclical* $\iff \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$ *and* $\boldsymbol{\mu} \in \mathcal{C} = \bigcup_{L,U} C_{LU}$, *where* $L < U \in \{1, \ldots, n\}$ *and* $C_{LU} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_1 \geq \mu_2 \geq \ldots \geq \mu_L \leq \mu_{L+1} \leq \ldots \mu_{U-1} \leq \mu_U \geq \mu_{U+1} \geq \ldots \geq \mu_n\}$.

In other words, a signal is said to be cyclical if (a) it has the same pattern $\boldsymbol{\mu}$ in both periods (i.e. $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$), and

(b) in each period the signal monotonically decreases up to time point L and then increases up to a time point U (with $L < U$) before decreasing again. This definition is more general than the classical sinusoidal shape. Note that, for convenience and without loss of generality, in the above definition we have the trough followed by the peak (i.e. $L < U$). Our definition of cyclical signal allows for the opposite pattern where in each period the signal monotonically increases up to time point U and then decreases up to a time point L (with $U < L$) before increasing again. A first description of cyclical patterns using inequalities appears in (33), and was implemented in (34) for short time-course experiments (see (34,35)). In this paper, we develop a different methodology that is also able to cope with long time-course data as those appearing in circadian clock studies.

We have observed from cell cycle and circadian clock data that there are rhythmic genes that do not have a cyclical signal. For example, they have multiple local peaks and/or troughs within each period and such patterns are repeated in both periods. We refer to such genes as quasi cyclical (panel (B) in Figure 1).

It is also common to find genes with non-flat and non-periodic signals such as those depicted in panel (C) in Figure 1. For example, a gene may have a distinct cyclical pattern in one period but may have a flat pattern of expression in the other period. Patterns not represented by the above three are regarded as flat patterns (panel (D) in Figure 1) and that the observed data are merely noise around a flat line.

Consequently, our proposed algorithm ORIOS, described in the following subsection, classifies each gene into one of the four shapes described in Figure 1. More precisely, ORIOS classifies cyclical and quasi-cyclical as rhythmic genes, while flat and non-flat and non-periodic are classified as non-rhythmic genes. Some examples of genes corresponding to these four shapes from real data are displayed in Supplementary Figure S3 in the Supporting Materials.

Order restricted inference for oscillatory systems (ORIOS)

The flowchart of ORIOS is described in Figure 3. It consists of two major steps, a filtering step and a classification step. For each gene, we first estimate landmarks L and U as follows:

$$\hat{L} = \arg\min_i \overline{Y}_{i.} \qquad \hat{U} = \arg\max_i \overline{Y}_{i.}$$

In the first step, called filtering step, we distinguish between non-rhythmic genes and potentially rhythmic genes. A gene that is declared potentially rhythmic in the filtering step is classified either as cyclical, quasi cyclical or flat in the classification step. A non-rhythmic gene is classified as either flat or non-flat and non-periodic. Statistical tests performed in ORIOS are conditional tests developed in order restricted inference (cf. (7,36)).

Throughout this paper we assume that each gene is normally distributed as $\boldsymbol{Y}_j \sim N_n(\boldsymbol{\mu}_j, \sigma^2 I)$ for $j = 1, 2$.

Filtering Step. Corresponding to each gene, in this step we test the following hypotheses using the methodology described in Subsection 1.1 in the Supporting Materials, where c_1 and c_2 are some arbitrary unknown constants and $\mathbf{1} = (1, \ldots, 1)'$:

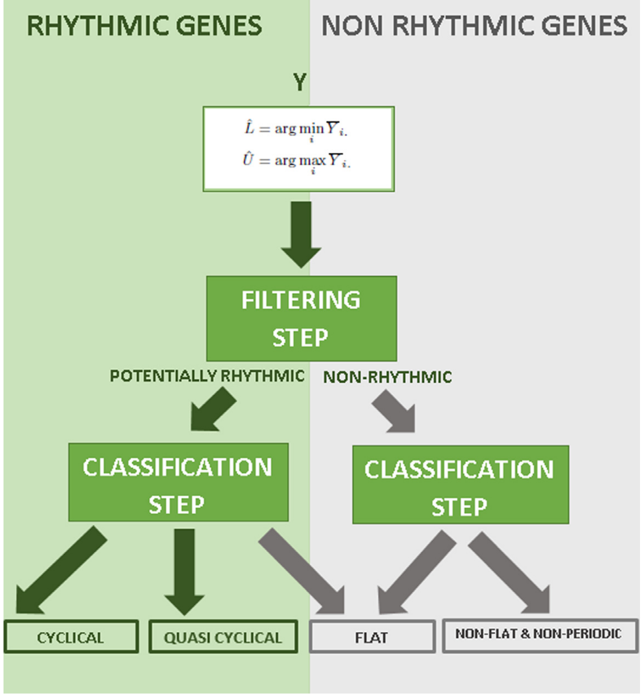


Figure 3. Flowchart of the ORIOS algorithm.

$$\begin{array}{ll} H_{10} : \boldsymbol{\mu}_1 = c_1\mathbf{1} & H_{20} : \boldsymbol{\mu}_2 = c_2\mathbf{1} \\ H_{11} : \boldsymbol{\mu}_1 \in C_{LU} & H_{21} : \boldsymbol{\mu}_2 \in C_{LU} \end{array}$$

Let p_{0j}, denote the p-value obtained for testing H_{j0} against the alternative H_{j1}, $j = 1, 2$. Let $p_0 = \max_{j=1,2}(p_{0j})$. Since there are a larger number of genes therefore a large number of tests are being performed. For this reason, we adjust the resulting p-values p_0 using the Benjamini–Hochberg (BH) procedure (37), to control the false discovery rate (FDR). Genes with BH adjusted p-values less than α are declared to be potentially rhythmic, otherwise they are declared to be non-rhythmic (see Figure 3).

Classification step. For genes that are declared to be non-rhythmic in the filtering step, we classify them into either flat signal (panel (D) in Figure 1) or non-flat and non-periodic signal (panel (C) in Figure 1). It is achieved by testing the null hypothesis $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu} = c\mathbf{1}$, where c some arbitrary unknown constant, against the alternative hypothesis that $\boldsymbol{\mu} \in C_{LU}$. More precisely, we test:

$$\begin{array}{l} H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu} = c\mathbf{1} \\ H_1 : \boldsymbol{\mu} \in C_{LU} \end{array} \qquad (1)$$

If the null is rejected, after adjusting the p-values by the BH procedure for multiple testing, then we conclude that the gene is non-flat and non-periodic . Otherwise we declare the gene to be flat.

For genes that are declared to be potentially rhythmic in the filtering step, we also distinguish among cyclical, quasi cyclical or flat observed expressions. To distinguish between

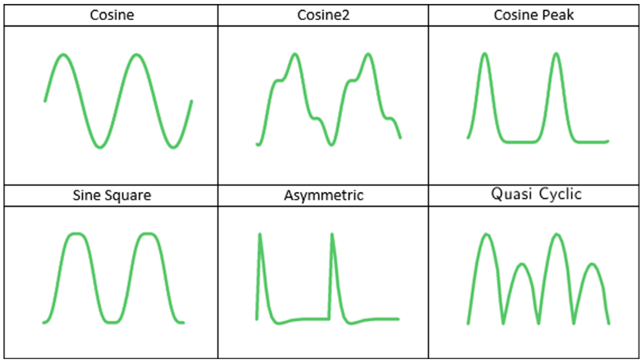


Figure 4. Rhythmic signal shapes in two periods for simulating rhythmic genes.

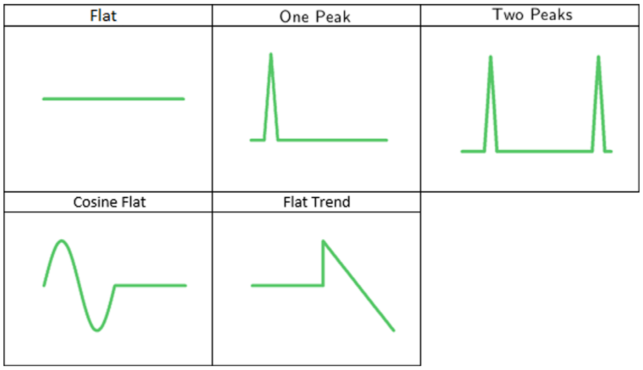


Figure 5. Non-rhythmic signal shapes in two periods for simulating non-rhythmic genes.

them, we first test the following hypotheses:

$$H_1 : \boldsymbol{\mu} \in C_{LU} \qquad (2)$$
$$H_2 : \boldsymbol{\mu} \in \mathbb{R}^n$$

If the null hypothesis H_1 is rejected then we conclude that the gene is periodic but not cyclical or flat, and hence classify it as quasi cyclical gene. If the null hypothesis H_1 is not rejected then we conduct the test described in (1). If the null H_0 in (1) is rejected (after adjusting for multiple testing using the BH procedure) then the gene is declared to be cyclical otherwise it is declared to be flat. Table 1 summarizes the gene classification according to the results obtained from the above testing problems. The theoretical details of this step are described in second subsection of the Supporting Materials.

RESULTS

Performance of ORIOS in a *Simulated dataset*

In this section we compare the performance of ORIOS with JTK_Cycle algorithm (22) and RAIN (26) using several simulated datasets described below. Specifically, we are interested in comparing the three methods in terms of the proportion of false positive and false negative identification of a given pattern as in related works in literature (38,39).

Study design. To compare the three methods, we simulated a total of 40 000 'genes'. Corresponding to each gene, within each period j, $j = 1, 2$, we simulated expression data $\boldsymbol{Y}_j$ for the 24 time points using simulated dataset $N_{24}(\boldsymbol{\mu}_j, \sigma^2 I)$ where σ^2 is fixed to be 1. The values of $\boldsymbol{\mu}_j$ were chosen so as to represent different shapes of signals. Motivated by circadian gene databases (40), we generated 30% rhythmic and 70% non-rhythmic genes, i.e. 12 000 rhythmic genes and 28 000 non-rhythmic genes.

The 12 000 patterns of rhythmic genes consisted of six shapes (2000 each) depicted in Figure 4, namely, *Cosine, Cosine Two, Cosine Peak, Sine Square, Asymmetric and Quasi Cyclic*. Of the 28 000 non-rhythmic genes, 26 000 were chosen flat pattern, and the remaining 2000 were chosen equally among the remaining 4 non-rhythmic patterns in Figure 5, namely, *One Peak, Two Peaks, Cosine Flat and Flat Trend*.

In this simulation study the time points representing the two periods were taken to be $\{0, 1, 2, \ldots, 47\}$, the phase shifts were chosen from an uniform distribution in [0, 47] and the median level amplitude is fixed according to (39). The α level for all our tests was taken to be 0.01. When BH procedure is used it represents the nominal FDR level.

Results of the simulation study. Results of our simulation study comparing the performance of ORIOS, JTK (version 3) and RAIN (according to the default parameters described in (26)) are summarized in Table 2. In each case, we computed the proportion of times an algorithm missed to identify a particular rhythmic pattern (i.e. false negative, FN) and the proportion of times it falsely declares a non-rhythmic pattern to be rhythmic (i.e. false positive, FP).

The simulation study illustrates that ORIOS has the smallest FP and FN rates. The FP rate is close to the nominal level of 0.01 and the FN rate is estimated to be 0. These rates are remarkably low compared to the two competitors, JTK and RAIN, which have an overall FP rates of 12% and 30%, respectively and FN rates of 33% and 22%, respectively.

Detection of rhythmic signals in published circadian gene datasets

We re-analyzed publicly available time-course gene expression data of (22) which are online available at NCBI GEO, (http://www.ncbi.nlm.nih.gov/geo/). The mouse liver and pituitary gland as well as the NIH3T3 cell lines data consisted of 45 101 probe sets each, whereas the U2OS human cell lines data consisted of 32 321 probe sets. The data were normalized using RMA grouping transcripts ENSEMBL gene annotations (41). Each data had 48 time points representing two periods of data. All analyses were performed with FDR α = 0.01. We compared ORIOS with JTK (version 3) and RAIN.

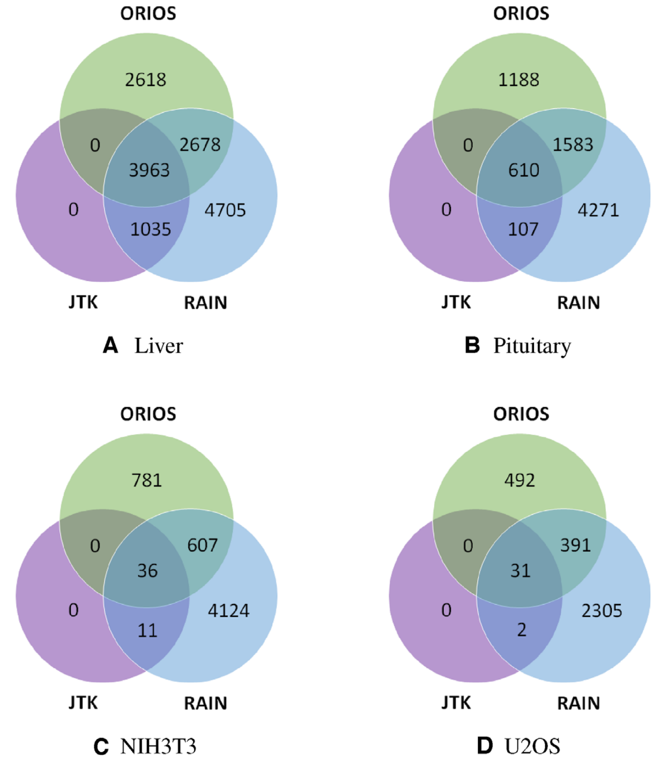
Number of genes identified as rhythmic by the three methods, for the two tissues and cell lines, is summarized in Tables 3 and 4 and in the Venn diagrams in Figure 6. In each dataset RAIN identified the most number of genes to be rhythmic, whereas JTK identified the fewest. Recall from the study reported in Table 2 that RAIN tends to have a very high false positive rate, while JTK tends to have a very high false negative rate. On the other hand, ORIOS has negligible false negative rate while controlling the false positive

Table 1. Gene classification according to ORIOS algorithm

Filtering Stage	Classification Stage		Result
	H_1 vs $H_2 - H_1$	H_0 vs $H_1 - H_0$	
Reject?	Reject?	Reject?	
Yes	Yes	-	Quasi cyclical
Yes	No	Yes	Cyclical
Yes	No	No	Flat
No	-	Yes	Non-flat and non-periodic
No	-	No	Flat

Table 2. False positive and negative rates and mean error for the different signals in the simulated datasets for each classification algorithm considered (α = 0.01)

	ORIOS		JTK		RAIN	
	FP+	FN-	FP+	FN-	FP+	FN-
Cosine		0.000		0.000		0.000
Cosine Two		0.000		0.000		0.000
Cosine Peak		0.000		0.003		0.000
Sine Square		0.000		0.000		0.000
Asymmetric		0.001		0.973		0.652
Quasi Cyclic		0.001		1.000		0.687
Flat	0.052		0.000		0.018	
One Peak	0.000		0.000		0.010	
Two Peaks	0.000		0.000		0.014	
Cosine Flat	0.012		0.504		0.900	
Flat Trend	0.008		0.102		0.572	
MEAN ERROR	**0.014**	**0.000**	**0.121**	**0.329**	**0.303**	**0.223**



A Liver

B Pituitary

C NIH3T3

D U2OS

Figure 6. Number of rhythmic genes overlapping between ORIOS (green), JTK (purple) and RAIN (blue) for the four datasets considered (α = 0.01).

rate within 1%. Therefore in view of the simulation study results, it is plausible that many of the genes identified by RAIN are false positives and JTK may have failed to identify truly rhythmic genes (false negatives). ORIOS, on the other hand, may have correctly identified many rhythmic as well as non-rhythmic genes in the four datasets.

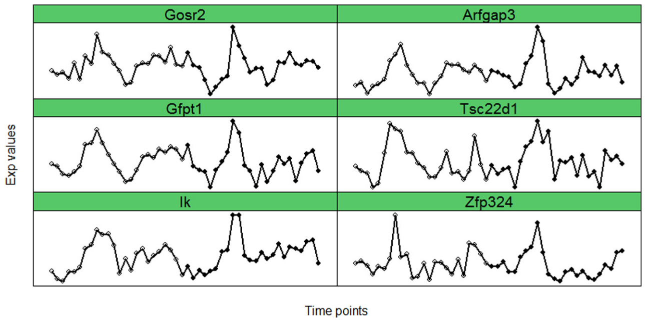


Figure 7. Some rhythmic circadian genes in mouse liver according to ORIOS, which are detected as non-rhythmic by JTK and RAIN (α = 0.01).

Unlike JTK and RAIN, ORIOS not only identifies rhythmic and non-rhythmic genes but it further classifies them as cyclical, quasi cyclical, flat or non-flat and non-periodic. For each tissue/cell line, in Table 5 we summarize the classifications obtained by ORIOS. Some examples of patterns detected by ORIOS but not detected by JTK or RAIN are provided in Figure 7. Note that genes such as Zfp324 and Gosr2 display rhythmic but not perfectly sinusoidal patterns. They display *asymmetric* and *quasi cyclic* rhythms, respectively. Similar patterns are observed for genes in the pituitary gland and the two cell lines (see genes Trim8, Mvh14 and IK in Supplementary Figures S4, S5 and S6 in the Supporting Materials).

Table 3. Number of genes identified as rhythmic by ORIOS, JTK and RAIN for mouse liver, pituitary gland, NIH3T3 and U2OS cell lines (α = 0.01)

	ORIOS	JTK	RAIN
Liver	9259	4998	12381
Pituitary	3381	717	6571
NIH3T3	1424	47	4778
U2OS	914	33	2729

Table 4. Rhythmic and non-rhythmic joint gene detection for ORIOS vs JTK and ORIOS vs RAIN in the four datasets considered (α = 0.01)

		JTK		RAIN	
	ORIOS	Rhythmic	Non-rhythmic	Rhythmic	Non-rhythmic
Liver	Rhythmic	3963	5296	6641	2618
	Non-rhythmic	1035	34 807	5740	30 120
Pituitary	Rhythmic	610	2771	2193	1188
	Non-rhythmic	107	41 613	4378	37 342
NIH3T3	Rhythmic	36	1388	643	781
	Non-rhythmic	11	43 666	4135	39 542
U2OS	Rhythmic	31	883	422	492
	Non-rhythmic	2	31 405	2307	29 100

Table 5. Number of genes classified according to different shape categories by ORIOS

	Rhythmic Signals		Non-rhythmic Signals	
	Cyclical	Quasi Cyclical	Flat	Non-flat and Non-periodic
Liver	9167	92	35 788	54
Pituitary	3363	18	41 720	0
NIH3T3	1411	13	43 677	0
U2OS	906	8	31 407	0

Consistent with published literature (22), ORIOS identified considerably more rhythmic genes in liver than pituitary gland (9,22,42) and far more than in synchronized cell lines.

DISCUSSION

Modeling gene expression patterns in time-course experiments using a parametric function is a challenging problem as not all genes may obey the same functional form. Even if they did, the determination of a flexible functional form is a challenging task.

Unlike methods based on a parametric function, such as the sinusoidal, ORIOS is free of any modeling assumptions. Consequently, it is fairly flexible to detect a wide range of rhythmic temporal patterns of gene expression such as those depicted in Figure 1. Such non-cyclical but periodic gene expression patterns are rather common in circadian clock or cell cycle gene expression studies (see Supplementary Figures S1 and S2 in the Supporting Materials). Although ORIOGEN (34) is also a methodology based on order restricted inference, by design it will not be powerful for detecting patterns in a long series experiments for oscillatory systems such as the cell-cycle and the circadian clock. This is because ORIOGEN formulates the pattern recognition problem as a union-intersection test. As the number of time points increases, as it would in the case of circadian clock and cell-cycle experiments, the number of alternative hypotheses tested in the union-intersection test in ORIOGEN increases. This results in a substantial loss of power. In fact, (35) discussed this issue and recommended against using ORIOGEN for long series time course experiments. On the other hand ORIOS is an efficient procedure that circumvents the union-intersection test conducted in ORIOGEN and takes a more direct approach to the problem.

As seen from our simulation studies, the detection of additional non-cyclical but periodic gene expression patterns presents a distinct advantage to ORIOS. ORIOS tends to have negligible false negative rate (i.e. failing to detect a rhythmic pattern when there is one) while controlling the false positive rate (i.e. falsely declaring a pattern to be rhythmic when it is not). Moreover, as seen in Supplementary Table S1 in the Supplementary material, the computation time for ORIOS is similar to that of JTK and lower than that of RAIN. We conducted additional simulation studies (results reported in the Supplementary text) to investigate the performance of ORIOS, JTK and RAIN for sparse time course data where fewer time points are available within each period. We simulated experiments where samples are obtained every 2 h over 2 days (denoted as 2/2 design) and 4 h over 2 days (denoted as 4/2 design). Results are summarized in the Supplementary text (Supplementary Tables S2 and S3). The results corresponding to the 2/2 design (Supplementary Table S2) are very similar to those of the 1 hour/2 days design. ORIOS generally performs well in comparison to JTK as well as RAIN. In the case of 4/2 design, there are only 6 time points within each period, which are not large enough to describe *Cosine Two* and *Quasi Cyclic* patterns and hence these 2 patterns were not included for the 4/2 design. From the results reported in Supplementary Table S3, we notice that for the 4/2 design ORIOS performs well in terms of the overall false positive and false negative rates. However, in terms of individual patterns, RAIN seems to have smaller false negative rates than ORIOS for the *Cosine* and the *Sine Square* patterns. On the other hand, RAIN has inflated false negative rates for other patterns and has a large false positive rate compared to ORIOS. JTK tends to have large false negative rates in comparison to both ORIOS and RAIN. It is important to note that when the frequency of sampling is hourly then we have a total of 48 time points over 2 days. Due to this large sample size, our simulation studies suggest that, even at a level of significance as small as 0.01, ORIOS had sufficient power to recognize various patterns quiet accurately. However, as the sampling frequency decreases we expect the power to decrease. For this reason, for 2/2 and 4/2 designs we used the usual nominal level of 0.05 in our simulations reported in the Supplementary text. In fact, in practice we recommend the users to the usual nominal level of 0.05 for low sampling frequency but use smaller level of significance, such as 0.01, when the sampling frequency is high.

As seen in the case of real data, in comparison to JTK and RAIN method, ORIOS successfully identifies rhythmic genes such as 1444048_at or Piga (see Supplementary Figures S7a and S7b in the Supporting Materials) when JTK and RAIN declare them as non-rhythmic. Not only does ORIOS declare these genes as rhythmic, it further classifies them as cyclical and quasi cyclical rhythms, respectively (see panels (a) and (b) in Figure 1, respectively). Conversely, ORIOS also declares genes such as PSMF1 (see Figure S7c in the Supporting Materials) as non-rhythmic when JTK and RAIN declare it as rhythmic (i.e. false positive detection). More precisely, ORIOS classifies PSMF1 as a non-flat and non-periodic gene (see panel (c) in Figure 1).

Most circadian clock and cell-cycle gene expression studies available in the literature, that we are familiar with, consist of data corresponding to at most two periods. For examples, please refer to the large circadian clock database CircaDB (40) (http://circadb.org) or the famous cell-cycle data base Cyclebase (43) (http://www.cyclebase.org) among others. Each of these websites contains numerous data sets. In each case, the number of periods is at most two. Typically, the long series time course experiments are intrinsically expensive and hence it is not common to study more than two periods. Not only that, as noted in (44), due to cost considerations, researchers consider single replicates at each time point. For these reasons, we developed our methodology for data involving two periods.

Although we have illustrated our methodology and algorithm for circadian clock data, this procedure can also be used for other oscillatory systems such as, for example, the cell division cycle. Often time course course experiments, such as in the cell cycle experiments (e.g. (45)) may contain more than two periods. In some cases there may be data available on partial third period. Although the methodology described in this paper assumes there are two periods, it can easily be extended to cases when there are more than two periods of data (even if it is partial third period). Secondly, in some instances the exact length of the period may be unknown a priori. In such cases, ORIOS can be modified to first estimate the period of the cycle. Once that is done, the ORIOS algorithm proposed in this paper can be implemented.

The expression data on each gene in an oscillatory system is an average over thousands of cells. Although all cells may be synchronized at the beginning of the cell cycle experiments, over time the cells cease to be synchronized. When that happens the time course gene expression pattern of a cell cycle gene would display attenuation of expression, a phenomenon that is common to cell cycle experiments (46). Although the present methodology does not specifically model such cell-cycle data, it can be extended to cope with this issue. Moreover, although beyond the scope of this paper, the ORIOS methodology can also be extended to handle heteroscedasticity (i.e. non-constant variance across time), non-normality and dependent time course experiments (e.g. repeated measurements) using resampling procedures such as the bootstrap.

In conclusion, we have introduced a simple methodology that does not make any modeling assumptions, for identifying temporal patterns in gene expression studies. The proposed methodology is flexible and robust to the shape of the gene expression curve and an easy to use R-code is available from the following website to implement ORIOS (http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/peddada/index.cfm).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Drs Delong Liu (NIH) and S. Barragán (Universidad de Valladolid) for their valuable comments on an earlier draft which have led to an improved presentation of the manuscript. The authors also thank the editor and two anonymous reviewers for their comments that improved the presentation of the manuscript.

FUNDING

Spanish Ministerio de Ciencia e Innovación [MTM2015-71217-R to C.R. and M.A.F.]; Spanish Ministerio de Educación, Cultura y Deporte [FPU14/04534 to Y.L.]; Intramural Research Program of the National Institute of Environmental Health Sciences (NIEHS) [Z01 ES101744-04 to S.D.P.]. Funding for open access charge: NIEHS.

Conflict of interest statement. None declared.

REFERENCES

1. Slavov,N., Airoldi,E.M., van Oudenaarden,A. and Botstein,D. (2012) A conserved cell growth cycle can account for the environmental stress responses of divergent eukaryotes. *Mol. Biol. Cell*, **23**, 1986–1997.
2. Oliva,A., Rosebrock,A., Ferrezuelo,F., Pyne,S., Chen,H., Skiena,S., Futcher,B. and Leatherwood,J. (2005) The cell-cycle-regulated genes of Schizosaccharomyces pombe. *PLoS Biol.*, **3**, 1239–1260.
3. Rustici,G., Mata,J., Kivinen,K., Lio,P., Penkett,C.J., Burns,G., Hayles,J., Brazma,A., Nurse,P. and Bahler,J. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.*, **36**, 809–817.
4. Peng,X., Karuturi,R.K.M., Miller,L.D., Lin,K., Jia,Y., Kondu,P., Wang,L., Wong,L., Liu,E.T., Balasubramanian,M.K. and Liu,J. (2005) Identification of cell cycle-regulated genes in fission yeast. *Mol. Biol. Cell*, **16**, 1026–1042.
5. Jensen,J.L., Jensen,T.S., Lichtenberg,U., Brunak,S. and Bork,P. (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature*, **443**, 594–597.

6. Fernández,M.A., Rueda,C. and Peddada,S.D. (2012) Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Res.*, **40**, 300–330.
7. Rueda,C., Fernández,M.A. and Peddada,S.D. (2009) Estimation of parameters subject to order restrictions on a circle with application to estimation of phase angles of cell-cycle genes. *J. Am. Stat. Assoc.*, **104**, 338–347.
8. Xiao,E., Xia-Zhang,L., Barth,A., Zhu,J. and Ferin,M. (1998) Stress and menstrual cycle: relevance of cycle quality in the short- and long-term response to a 5-day endotoxin challenge during the follicular phase in the Rhesus monkey. *J. Clin. Endocrinol.*, **88**, 2454–2460.
9. Hughes,M.E., DiTacchio,L., Hayes,K.R., Vollmers,C., Pulivarthy,S., Baggs,J.E., Manda,S. and Hogenesch,J.B. (2009) Harmonics of circadian gene transcription in mammals. *PLoS Genet.*, **5**, e1000442.
10. Cermakian,N., Lamont,E.W., Bourdeau,P. and Boivin,D.B. (2011) Circadian clock gene expression in brain regions of Alzheimer's disease patients and control subjects. *J. Biol. Rhythm.*, **26**, 160–170.
11. Kondratova,A.A. and Kondratov,R.V. (2012) The circadian clock and pathology of the ageing brain. *Nat. Rev. Neurosci.*, **13**, 325–335.
12. Altevogt,B.M. and Colten,H.R. (2006) *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. National Academies Press, Washington DC.
13. Chaput,J.P., Després,J.P., Bouchard,C. and Tremblay,A. (2008) The association between sleep duration and weight gain in adults: a 6-year prospective study from the Quebec Family Study. *Sleep*, **31**, 517–523.
14. Lyytikáinen,P., Rahkonen,O., Lahelma,E. and Lallukka,T. (2011) Association of sleep duration with weight and weight gain: a prospective follow-up study. *J. Sleep Res.*, **20**, 298–302.
15. Zhang,R., Lahens,N.F., Ballance,H.I., Hughes,M.E. and Hogenesch,J.B. (2014) A circadian gene expression atlas in mammals: Implications for biology and medicine. *PNAS*, **111**, 16219–16224.
16. Wichert,S., Fokianos,K. and Strimmer,K. (2004) Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, **20**, 5–20.
17. Lomb,N. (1976) Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci.*, **39**, 447–462.
18. Scargle,J. (1982) Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J.*, **263**, 835–853.
19. Straume,M. (2004) DNA Microarray Time Series Analysis: Automated Statistical Assessment of Circadian Rhythms in Gene Expression Patterning. In: Ludwig,B and Michael,LJ (eds). *Numerical Computer Methods, Part D*. Academic Press, Toronto, pp. 149–168.
20. de Lichtenberg,U., Jensen,L.J., Fausbll,A., Jensen,T.S., Bork,P. and Brunak,S. (2005) Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, **21**, 1164–1171.
21. Ahnert,S.E., Willbrand,K., Brown,F.C.S. and Fink,T.M.A. (2006) Unbiased pattern detection in microarray data series. *Bioinformatics*, **22**, 1471–1476.
22. Hughes,M.E., Hogenesch,J.B. and Kornacker,K. (2010) JTK CYCLE: An Efficient Nonparametric Algorithm for Detecting Rhythmic Components in Genome-Scale Data Sets. *J. Biol. Rhythm.*, **25**, 372–380.
23. Cohen-Steiner,D., Edelsbrunner,H., Harer,J. and Mileyko,Y. (2010) Lipschitz functions have L p-stable persistence. *Found. Comput. Math.*, **10**, 127–139.
24. Yang,R. and Su,Z. (2010) Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics*, **26**, 168–174.
25. Yang,R., Zhang,C. and Su,Z. (2011) LSPR: an integrated periodicity detection algorithm for unevenly sampled temporal microarray data. *Bioinformatics*, **27**, 1023–1025.
26. Thaben,P.F. and Westermark,P.O. (2014) Detecting Rhythms in Time Series with RAIN. *J. Biol. Rhythm.*, **29**, 391–400.
27. Leng,N., Chu,L.F., Barry,C., Li,Y., Choi,J., Li,X., Jiang,P., Stewart,R.M., Thomson,J.A. and Endzіorski,C.K. (2015) Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods*, **12**, 947–950.
28. Robertson,T., Wright,F.T. and Dykstra,R.L. (1988) *Order Restricted Statistical Inference*. John Wiley & Sons, Chichester.
29. Silvapulle,M.J. and Sen,P.K. (2004) *Constrained Statistical Inference: Inequality, Order and Shape Restrictions*. John Wiley & Sons, Hoboken.
30. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, 991–995.
31. Panda,S., Antoch,M.P., Miller,B.H., Su,A.I., Schook,A.B., Straume,M., Schultz,P.G., Kay,S.A., Takahashi,J.S. and Hogenesch,J.B. (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, **109**, 307–320.
32. Mockler,T.C., Michael,T.P., Priest,H.D., Shen,R., Sullivan,C.M., Givan,S.A., McEntee,C., Kay,S.A. and Chory,J. (2007) The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. In: *Cold Spring Harb. Sym*. Cold Spring Harbor Laboratory Press, NY, pp. 353–363.
33. Peddada,S.D., Lobenhofer,E.K., Li,L., Afshari,C.A., Weinberg,C.R. and Umbach,D.M. (2003) Gene selection and clustering for time-course and doseresponse microarray experiments using order-restricted inference. *Bioinformatics*, **19**, 834–841.
34. Peddada,S., Harris,S., Zajd,J. and Harvey,E. (2005) ORIOGEN: order restricted inference for ordered gene expression data. *Bioinformatics*, **21**, 3933–3934.
35. Peddada,S.D., Umbach,D.M. and Harris,S. (2012) Statistical analysis of gene expression studies with ordered experimental conditions. In: Chakraborty,R, Rao,CR and Sen,P (eds). *Handbook of Statistics: Bioinformatics in Human Health and Heredity*. Elsevier, NY, pp. 39–66.
36. Rueda,C., Ugarte,M.D. and Militino,A.F. (2015) Checking unimodality and locating the break-point: An application to breast cancer mortality trends. *Stoch. Env. Res. Risk A*, **30**, 1277–1288.
37. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Roy. Stat. Soc. B Met.*, **57**, 289–300.
38. Deckard,A., Anafi,R.C., Hogenesch,J.B., Haase,S.B. and Harer,J. (2013) Design and analysis of large-scale biological rhythm studies: a comparison of algorithms for detecting periodic signals in biological data. *Bioinformatics*, **29**, 3174–3180.
39. Wu,G., Zhu,J., Yu,J., Zhou,L., Huang,J.Z. and Zhang,Z. (2014) Evaluation of five methods for genome-wide circadian gene identification. *J. Biol. Rhythm.*, **29**, 231–242.
40. Pizarro,A., Hayer,K., Lahens,N.F. and Hogenesch,J.B. (2013) CircaDB: a database of mammalian circadian gene expression profiles. *Nucleic Acids Res.*, **41**, 1009–1013.
41. Dai,M., Wang,P., Boyd,A.D., Kostov,G., Athey,B., Jones,E.G., Bunney,W.E., Myers,R.M., Speed,T.P., Akil,H. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.
42. Hughes,M., Deharo,L., Pulivarthy,S.R., Gu,J., Hayes,K., Panda,S. and Hogenesch,J.B. (2007) High-resolution time-course analysis of gene expression from pituitary. In: *Cold Spring Harb. Sym*. Cold Spring Harbor Laboratory Press, NY, pp. 381–386.
43. Gauthier,N., Larsen,M.E., Wernersson,R., de Lichtenberg,U., Jensen,L.J., Brunak,S. and Jensen,T.S. (2008) Cyclebase.org—a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Res.*, **36**, 854–859.
44. Walter,W., Striberny,B., Gaquerel,E., Baldwin,I.T., Kim,S.-G. and Heiland,I. (2014) Improving the accuracy of expression data analysis in time course experiments using resampling. *Bioinformatics*, **15**, 1–9.
45. Oliva,A., Rosebrock,A., Ferrezuelo,F., Pyne,S., Chen,H., Skiena,S., Futcher,B. and Leatherwood,J. (2005) The cell cycle-regulated genes of Schizosaccharomyces pombe. *PLoS Biol.*, **3**, 1239–1260.
46. Liu,D., Umbach,D.M., Peddada,S.D., Li,L., Crockett,P.W. and Weinberg,C.R. (2004) A random-periods model for expression of cell-cycle genes. *PNAS*, **101**, 7240–7245.