II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

A new method for detection of rhythmic signals in oscillatory systems

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, Spain
[2]Biostatistics Branch, NIEHS (NIH), Research Triangle Park, NC, USA

Abstract

The study of biological rhythms is receiving a lot of attention in the literature in recent years. At the core of this research lies the methodological problem of how to detect rhythmic signals in measured data. Night and day, or dark and light patterns impact on human health in many different ways. For this reason, researchers are studying the effect of sleep on the circadian clock in human body during various stages of life. Important components of this clock are the circadian genes which have rhythmic expression overtime with phases suitably matching the night and day. Consequently, the identification of rhythmic signals is a problem of considerable interest for biologists.

In this work, we develop a novel statistical procedure to detect rhythmic signals in oscillatory systems based on Order Restricted Inference (ORI). This methodology is tested both on simulations and on real data bases. Moreover the obtained results are compared with the most widely extended rhythmicity detection algorithms in literature.

Key words: Order Restricted Inference, Conditional Tests, Circadian Rhythms, Cyclic Signal, Oscillatory Systems

1. INTRODUCTION

Human health is influenced by rhythmicity patterns, like day and night, in many ways with circadian clock regulating metabolic and physiologic processes. There is abundant literature documenting that people are having, on average, less sleep during the night and this disruption or reduction in sleep is associated with numerous health outcomes including obesity. Among teenagers this may affect the production of their growth hormones and result in abnormal growth patterns. For these reasons, researchers are in studying the effect of sleep on the circadian clock in human body during various stages of life. Important components of this oscillatory system are the circadian genes which present rhythmic expression overtime according to night and day cycles. Circadian genes consist of two well differentiated compounds, the signal term that we denote by $\boldsymbol{\mu}$, and the random error term which is derived from the experimental noise. Thus, the research in oscillatory systems focuses on the detection of which of the signals they origin are rhythmic.
The identification of rhythmic genes in the circadian clock among several thousands of genes is not a simple problem due to the large variability in the data, the high number of data (several

thousands of genes) to be processed and the fact that in many of the available datasets the number of retrievable data for each gene is not high enough to properly fit mathematical models such as Fourier's models. This difficulty is reflected in the richness of the literature on this subject as well as in the wealth of methods and algorithms devoted to this task (e.g. Straume [2004], Hughes et al. [2010], Thaben and Westermark [2014]). These methods assess the fit between the expression pattern of each transcript and a series of curves, most often sinusoidal ones, with different period lengths and phases, and they estimate the period length and phase of each circadian transcript as that of its best-matched curve.

2. CONTRIBUTIONS

Circadian gene expressions data are observed during two periods of length 24 hours ($T = 24$) and usually the time sampling frequency is fixed to be 1 hour, i.e., there are 24 time points ($n = 24$) in each period. The general mathematical formulation of rhythmicity (periodicity) in the two periods $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$ (from now on *property 1*) may be too rigid a model to be useful to detect interesting genes in practice. Real data evidence that the expression profiles of many genes are rhythmic but do not come from a sinusoidal signal, although quite often the signal has an unique maximum in each period corresponding to the moment where the gene is activated, and an unique minimum generating an up-down up pattern, (from now on *property 2*). Finally, in order to define alternative signal patterns that fit to the different gene expression data appearing in practice, we have distinguished four main classes of genes that correspond to the four signal patterns shown in Figure 2.
In this line, this work proposes a broader formulation for a rhythmic gene signal that we call *cyclic signal*, (see Definition 1), which includes as a particular case the sinusoidal signal. Figure 1 shows a typical cyclic pattern according to Definition 1.

Definition 1 *Cyclic signal.* $\boldsymbol{\mu}$ *is a cyclic signal* $\iff$ $\boldsymbol{\mu} \in \mathcal{C} = \bigcup_{L,U} C_{LU}$, *where* $L, U \in \{1, \ldots, n\}$ *and* $C_{LU} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_L \leq \mu_{L+1} \leq \cdots \leq \mu_U \geq \mu_{U+1} \geq \cdots \geq \mu_{L-1}\}$.

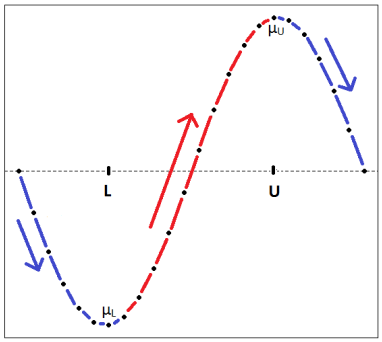
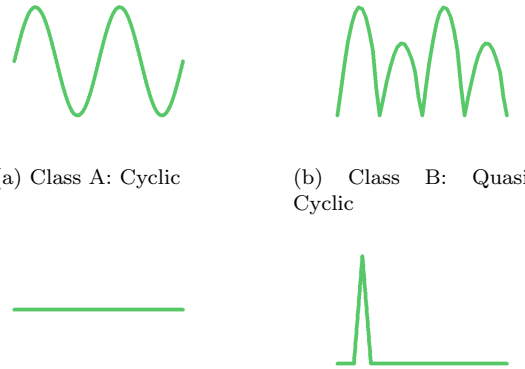


Figure 1: Profile of a Cyclic Signal $\boldsymbol{\mu}$

Two elements are remarkable in the definition of a cyclic signal. On one side, the two landmarks L and U, that determine the minimum and maximum in a period. On the other side, the restrictions among the components of $\boldsymbol{\mu}$, determining the up-down-up pattern. This order pattern points to the use of ORI methodology (see Robertson et al. [1988]). In addition, the properties defined above let us classify genes into one of the main classes of genes according to their signal patterns, see Figure 2. To be more precise, most typical periodic genes verify *properties 1* and *2* and belong to Class A. Genes in class B verify *property 1* but not *property 2*, i.e., they present for each period more than one local maximum or minimum. Class C includes the genes whose expression data variability, in both periods, is totally explained by the random error term. Finally, genes in Class D fail *property 1* but verify *property 2* at least in one period. From the biological interest point of view, classes A and B include periodic genes and C and D involve the non periodic ones.



(a) Class A: Cyclic

(b) Class B: Quasi Cyclic

(c) Class C: Flat

(d) Class D: Non homogeneous

Figure 2: Class gene patterns in the two periods

This work also presents the algorithm we propose to identify periodic genes, and classify them into the classes appearing in Figure 2. It consists of three basic steps: estimate the landmarks L and U; conduct a filtering step and then a classification step. The algorithm relies heavily on ORI methodology and involves a series of sequentially likelihood ratio conditional tests (see Menéndez et al. [1991], Fernández et al. [2012]). In addition, a false discovery rate (FDR) procedure is also considered since multiple testing problems are conducted.

3. RESULTS

To evaluate the performance of ORI to detect periodicity, different simulation studies have been conducted, including the simulation of an *artificial dataset* which imitates what occurs in circadian real data bases. This *artificial dataset* contains 40000 simulated genes. The proportion of periodic and non periodic genes are 30% and 70%, respectively. The first group of patterns in Table 1 simulates periodic genes and the second one does non periodic genes. We also compare the ORI results with those from the most widely used rhythmicity detection algorithms in literature, JTK_Cycle (JTK) and RAIN (see Hughes et al. [2010] and Thaben and Westermark [2014], respectively), which are non parametric algorithms to detect sinusoidal signals.

	ORI		JTK		RAIN	
Shape pattern	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 Outlier	0.000		0.000		0.010	
Two Outliers	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**

Table 1: False positive and negative rates for each algorithm for the *artificial dataset*

From Table 1 we conclude that JTK presents the highest false negative rates for asymmetric and quasy cyclic shape patterns and that RAIN increases JTK false positive rates for non periodic patterns. In contrast with this, ORI outperforms the simulation results obtained by JTK and RAIN.

We have also analysed four real datasets including (each of them) 45101 circadian genes from mouse liver, pituitary and NIH3T3 cell lines; and 32321 circadian genes from U2OS human cell lines. In all four real datasets, RAIN is the approach which identifies more periodic genes and JTK is which does less, while ORI is between them, reiterating the results obtained in simulations. In addition, we found particularly interesting the case of NIH3T3 and U2OS cell lines, where as it is known the influence of noise is higher; but even so ORI still advantages over JTK.
Finally, from biological interest point of view, the relevant contribution is that ORI detects new periodic circadian genes. Figure 3 illustrates six of those new periodic circadian genes from the mouse liver tissue. In contrast with ORI, JTK and RAIN detect all of them as non periodic genes, despite that a simple inspection of them confirms a strong circadian rhythmiticity.

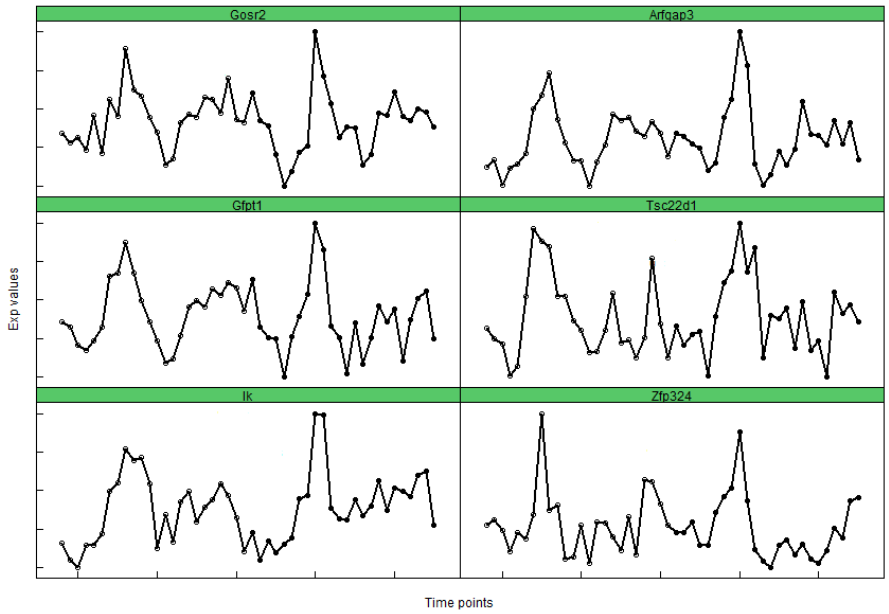


Figure 3: Six new periodic circadian genes detected by ORI from the mouse liver tissue

References

M.A. Fernández, C. Rueda, and S.D. Peddada. Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Research*, 40(7):2823–2832, 2012.

M.E. Hughes, J.B. Hogenesch, and K. Kornacker. JTK CYCLE: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of Biological Rhythms*, 25(5):372–380, 2010.

J.A. Menéndez, C. Rueda, and B. Salvador. Conditional test for testing a face of the tree order cone. *Communications in statistics. Simulation and computation*, 20(2-3), 1991.

T. Robertson, F.T. Wright, and R.L. Dykstra. *Order Restricted Statistical Inference*. John Wiley & Sons, 1988.

M. Straume. DNA microarray time series analysis: Automated statistical assessment of circadian rhythms in gene expression patterning. In Ludwig Brand and Michael L. Johnson, editors, *Numerical Computer Methods, Part D*, volume 383 of *Methods in Enzymology*, pages 149–166. Academic Press, 2004.

P.F. Thaben and P.O. Westermark. Detecting rhythms in time series with rain. *Journal of Biological Rhythms*, 29(6):391–400, 2014.