A normalization-robust bootstrap-based rhythmicity measure to detect rhythmic genes in oscillatory systems

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

[1] Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Paseo de Belén 7, 47011 Valladolid, Spain

[2] Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences (NIEHS),Alexander Dr., RTP, NC 27709, USA

E-mail for correspondence: `yolanda.larriba@uva.es`

Abstract: Microarray gene expression data are extremely noisy. Normalization is widely regarded as an essential step before data analysis to remove systematic variations while maintaining biological signals of interest. However, the choice of normalization may substantially impact the detection of rhythmic genes in oscillatory systems. We introduce a rhythmicity measure and a bootstrap methodology to detect rhythmic genes robust with respect to the normalization choice.

Keywords: Microarray; Normalization; Rhythmic genes.

1 State of the Art

Microarrays are a powerful technology widely used research tool in gene expression analysis of biological systems, such as circadian clock or cell cycle. One of the major difficulties dealing with high-throughput microarray gene-expression experiments is the noisy nature of the data that is intrinsic to each array. Thus, previously to any to microarray analysis data are pre-processed, not only to reduce those non-biological sources of variations but also to take data from probe level to single expression values for every gene. A variety of pre-processing methods are available in literature, such as MBEI (*Li et al., 2001*), MAS 5.0 (*Hubbell et al., 2002*) or RMA (*Irizarry et al., 2003*). They usually involve three distinct steps, namely, Background correction, Normalization, and Summarization. Normalization is an important component of pre-processing microarray gene-expression data, since it removes (or reduces) non-biological variations among arrays. Many microarray normalization methods are available in literature, among those, the widely extended *Quantile, (Cyclic) Loess, Contrast, Constant, Invariant Set, Qspline and Variance Stabilization Normalization (VSN),* see *Huber et al., 2002* and *Gautier et al., 2004* for details. All of these methods are implemented in the R-package *Affy* (*Gautier et al., 2004*) together with background correction and summarization steps according to the lines proposed for RMA (*Irizarry et al., 2003*), so that a matrix of gene expressions is finally obtained as output of the pre-processing. However, there is no universally accepted normalization strategy, and each normalization strategy is based

on certain model and assumptions. Consequently, the downstream analyses are expected to depend upon the normalization method used.
According to the line of our research, the subsequent analyses are focused on the identification of rhythmic or periodic components over time derived from biological processes such as cell-cycle (*Oliva et al., 2005*) or circadian clock (*Larriba et al., 2016*) that are governed by oscillatory systems. There are several algorithms available in the literature to determine whether a gene is rhythmic or not. Some recent examples include JTK_Cycle (*Hughes et al., 2010*), RAIN (*Thaben et al., 2014*) and ORIOS (*Larriba et al., 2016*). The performance of such algorithms potentially depends upon, among other factors, the normalization methods used. However, there has not been a systematic evaluation of the impact of normalization methods on identifying rhythmic genes in studies involving oscillatory systems in literature. Yet, researchers are interested in identifying rhythmic genes.

2 Original Contributions

Original contributions of this work consist of introducing a bootstrap methodology that allows us to define a robust rhythmicity measure highly correlated across various normalization methods. A by-product of bootstrap procedure is that it can be used for simulating potentially realistic time-course circadian gene-expression data.
For each gene $g = 1, \dots, G$, we define the standard measure of gene rhythmicity $M^g(n, a) = 1 - \text{p-value}^g(n, a)$, where $\text{p-value}^g(n, a)$ denotes the Benjamini-Hochberg adjusted p-value of gene g, according to normalization n and algorithm a. Note that $0 \leq M^g(n, a) \leq 1$ for $g = 1, \dots, G$ so that, closer 0 indicates potentially non-rhythmic gene and closer 1 indicates potentially rhythmic gene.

2.1 Bootstrap methodology

Let $\mathbf{R}$ denote the tri-dimensional array of raw intensities obtained from a reference high-throughput circadian microarray experiment. Data in $\mathbf{R}$ are expressed at probe level, where R^g_{pt} states the raw intensity value for gene g on probe p at time point (*array*) t, where $g = 1, \dots, G$, $p = 1, \dots, P$ and $t = 1, \dots, T$. Let $\mathbf{X}$ be the tri-dimensional array derived from $\mathbf{R}$ after background correction and $\mathbf{X}^{(b)*}, b = 1, \dots, B$, the b^{th} simulated microarray datasets generated according to parametric bootstrap that is based on a linear model from corrrected intensities $\mathbf{X}$, see *Irizarry et al., 2003*. Normalization and summarization steps are then conducted on $\mathbf{X}^{(b)*}, b = 1, \dots, B$, to finally obtain a matrix of gene-expression values ready to measure gene-rhythmicity.

2.2 Robust measure of gene-rhythmicity

Given a normalization method n, a rhythmicity algorithm a, and a random realization of data, define the rhythmicity statistic $\mathbf{M}(n, a) = (M^1_{(n,a)}, \dots, M^G_{(n,a)})'$. Let $\boldsymbol{\theta}(n, a) = \mathbb{E}(\mathbf{M}(n, a))$ be the parameter of interest and $\hat{\boldsymbol{\theta}}(n, a) = \mathbf{M}(n, a)$ be its estimator. For $b = 1, \dots, B$, $\hat{\boldsymbol{\theta}}^{(b)*}(n, a)$ denotes the bootstrap estimate of $\boldsymbol{\theta}(n, a)$, for the b^{th} bootstrap

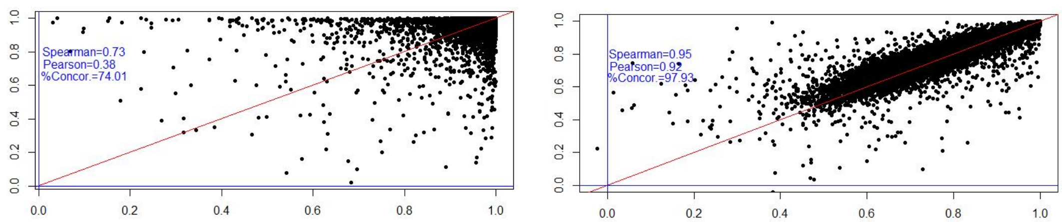

FIGURE 1. Pairwie scatter plots of $M^g(Qspline, ORIOS)$ vs $M^g(Loess, ORIOS)$ (left) and $M^g_{Robust}(Qspline, ORIOS)$ vs $M^g_{Robust}(Loess, ORIOS)$ (right) for a set of 15369 genes from mouse liver. Red line is the 45^{o} diagonal and the blue lines are the Cartesian axes.

sample. Then, $\mathbf{M}_{\text{Robust}}(n,a) = \widehat{\mathbb{E}}(\hat{\boldsymbol{\theta}}(n,a)) - \widehat{RMSE}(\hat{\boldsymbol{\theta}}(n,a))$ defines a **robust measure of gene rhythmicity**, where $\widehat{\mathbb{E}}(\hat{\boldsymbol{\theta}}(n,a)) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}^{(b)*}(n,a))$ and $\widehat{RMSE}(\hat{\boldsymbol{\theta}}(n,a)) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}^{(b)*}(n,a) - \hat{\boldsymbol{\theta}}(n,a))^2}$.
Given the extension of the present work analysis results are not shown, but it is worthwhile to mention that for every pair of normalization methods, subsequently correlation analyses show superb increases from standard to robust gene-rhythmicity measure, see Figure 1.

References

Gautier, L. *et al.* (2004). Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307 – 315.

Hubbell, E. *et al.* (2002). Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585 – 1592.

Huber, W. *et al.* (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Genome Biol.*, **18**, 96 – 104.

Hughes *et al.* (2010). JTK-CYCLE: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J. Biol. Rhythms*, **25**, 372 – 380.

Irizarry, R.A. *et al.* (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249 – 264.

Larriba, Y. *et al.* (2016). Order restricted inference for oscillatory systems for detecting rhythmic genes. *NAR*, **44**, e163.

Li, C. and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *P. Nati. Acad. Sci. USA.*, **98**, 31 – 36.

Oliva, A. *et al.* (2005). The cell cycle-regulated genes of Schizosaccharomyces pombe. *PLoS Biol.*, **3**, 1239 – 1260.

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