International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:1, No:5, 2007

Dynamical Analysis of Circadian Gene Expression

Carla Layana and Luis Diambra

***Abstract*—**Microarrays technique allows the simultaneous measurements of the expression levels of thousands of mRNAs. By mining this data one can identify the dynamics of the gene expression time series. By recourse of principal component analysis, we uncover the circadian rhythmic patterns underlying the gene expression profiles from Cyanobacterium Synechocystis. We applied PCA to reduce the dimensionality of the data set. Examination of the components also provides insight into the underlying factors measured in the experiments. Our results suggest that all rhythmic content of data can be reduced to three main components.

***Keywords*—**circadian rhythms, clustering, gene expression, PCA.

I. Introduction

MOLECULAR biology has traditionally focused on the study of individual genes considered in isolation as a method for determining gene function. However, given the availability of complete genomes from an ever-increasing list of organisms and in order to determine the principles underlying complex biological processes, it became necessary to simultaneously investigate the expression patterns of large numbers of genes, taking into consideration temporal, as well as, anatomical patterns. In this sense, the study of gene expression has been greatly facilitated by microarray technology. The anticipated flood of biological information produced by these experiments will open new perspectives into genetic analysis. Expression patterns have already been used for a variety of inference tasks such as identify gene clusters based on co-expression [1], [2], define metrics that measure a gene's involvement in a particular process [3], predict gene regulatory circuits [4]. Data on large-scale temporal gene expression patterns may provide for inferences on the causal links between genes expressed over the course of phenotypic change [5]-[7].

One of the challenges of bioinformatics is to develop effective ways to asses global gene expression data. A rigorous approach to gene expression analysis must involve an up-front characterization of the data structure. In addition to a broader utility in analysis methods,principal component analysis (PCA) can be valuable tool in obtaining such a characterization. Furthermore, the gene expression data is currently rather noisy, and PCA can detect and extract small signals from noisy data. PCA is an exploratory multivariate statistical technique for simplifying complex data sets [8], and gene expression data are well suited to analysis using PCA. Given m observations on n variables, the goal of PCA is to reduce the dimensionality of the data matrix by finding r new variables, where r is less than n. Termed principal components, these r new variables together account for as much of the variance in the original n variables as possible while remaining mutually uncorrelated and orthogonal. Each principal component is a linear combination of the original variables, and so it is often possible to ascribe meaning to what the components represent. Principal components analysis has been used in a wide range of biomedical problems, including the analysis of microarray data in search of outlier genes [9] as well as the analysis of other types of expression data [10], [11]. DNA microarray data sets are currently appearing in the literature available, where most initial analysis have focused on characterizing the waveform of gene expression over time, and in clustering genes based on this waveform or other features. When clustering genes based on expression information, it can be important to determine if the experiments have independent information or are highly correlated.

The paper is organized as follows: in Sec. II we describe the IT technique for modeling expression and establish a model based distance between two temporal profiles. We applied the method to online available data in Sec. III. Sec. IV discusses and summarizes the results obtained.

II. Methods

A. Experimental data

Kucho *et al*. monitored genome-wide mRNA levels, for 3,070 Cyanobacterium Synechocystis chromosomal genes simultaneously, over two circadian cycle period, 48 hours, at 4-hours intervals [12]. RNA samples were isolated from two independent cyanobacterial cultures. Each RNA sample was used for three independent microarray experiments. Thus, a maximum of six data points per gene was obtained for each time point of a biological replicate (i.e., three technical

C. L. Author is with the Centro Regional de Estudios Genómicos, Universidad Nacional de La Plata, CP 1888, Argentina. (corresponding author to provide phone/fax: +5411-4275-8100; e-mail: clayana@creg.org.ar).

L. D. Author is with the Centro Regional de Estudios Genómicos, Universidad Nacional de La Plata, CP 1888, Argentina. (e-mail: ldiambra@creg.org.ar).

replicates x two spots). Each biological replicate was treated independently with the same procedure until the final step of the cycling gene characterization of their rhythmicity and phase. Spots meeting any of the following criteria were flagged and not used for the analysis: (i) the GenePix Pro did not find the spot area automatically, (ii) the net signal intensity was ≤ 0, (iii) the percentage of saturated pixels in the spot area was ≥ 25, and (iv) severe noise was present. Genes carrying fewer than one unflagged data at any time point were removed from our analysis. The data are available from KEGG database http://www.genome.ad.jp/kegg/expression. We normalize each gene expression time series to media zero and maximum of 1.0; as it is sometimes recommendable when attempting PCA on measurements that are not on a comparable scale.

B. Principal component analysis technique

We considered a set of m time series $S_1,\ldots, S_j,\ldots, S_m$, each one corresponding to the temporal expression of each gene. The time series corresponding to gene j is defined by $S_j = x_{1,j},\ldots,x_{i,j},\ldots,x_{n,j}$, where $x_{i,j}$ is the expression level of the i-th gene in the j-th time point. Thus $x_{i,j}$ can be considered as the elements of a matrix X of size mxn.

Principal component can ve consider as a linear transformation, known as Karhunen--Loève expansion, of the expression data from the genes x array space to the reduced *eigenarrays* x *eigengenes* space of size rxr, where r is the rank of the matrix X. The equation for singular value decomposition of X is the following:

$$X = UDV^t \tag{1}$$

where U is a mxn matrix, D is an nxn diagonal matrix, and V^t is also an nxn matrix. The columns of U are called the left singular vectors, $\{u_k\}$, and form an orthonormal basis for the array expression profiles, named *eigenarrays*. The rows of V^t contain the elements of the right singular vectors, $\{v_k\}$, and form an orthonormal basis for the *eigengenes*. The elements of D are only nonzero on the diagonal, and are called the singular values. Furthermore, $d_k > 0$ for $1 \le k \le r$, and $d_i = 0$ for $(r + 1) \le k \le n$. By convention, the ordering of the singular vectors is determined by high-to-low sorting of singular values, with the highest singular value in the upper left index of the D matrix. Note that for a square, symmetric matrix X, singular value decomposition is equivalent to diagonalization, or solution of the eigenvalue problem.

We calculate the PCA first calculate V^t and D by diagonalizing the product $X_t X$ as follows

$$X_t X = VD^2V^t \tag{2}$$

and we can compute $U=XVD^{-1}$, where the $(r+1),...,n$ columns of V for which $d_k = 0$ are ignored in the last matrix multiplication. The remaining n-r singular vectors in V or U were calculated using the Gram-Schmidt orthogonalization process.

The diagonal elements of D correspond to *eigenexpression levels* and, following Alter *et al*. [13], one can define the fraction of eigenexpression as:

$$p_l = d_l^2 / \sum_{k=1}^{r} d_k^2 \tag{3}$$

indicates the relative significance of the *l-th eigengene* and *eigenarray* in terms of the fraction of the overall expression captured. The *eigengenes* and *eigenarrays* are unique, except in degenerate subspaces, defined by subsets of equal eigenexpression levels, and except for a phase factor of ±1, such that each *eigengene* (or *eigenarray*) captures both parallel and antiparallel gene (or array) expression patterns.

III. Results

The PCA procedures were applied to genes expression time series from both cyanobacterial cultures independently. One culture, named arbitrarily 1, the criterions described in Experimental data section lead to 2542 genes expression time series, while in the other culture, hereafter named 2, we obtained 2908 genes expression time series, in both cases of length 12 time points.

Considering the culture 1, our analysis indicates that the circadian expression data can be summarized in two variables. Fig. 1a displays the 12 *eigengenes* in 12 arrays sorted by the corresponding *eigenexpression* levels, while in Fig. 1b we show a bar chart of the fractions of *eigenexpression*. We can see that first *eigengene* v_1 captures about 35% of the overall expression. This *eigengene* describes an initial transient which increase and then decrease in the dataset expression. On the other hand, *eigengenes* v_2 and v_3 show oscillating circadian expression periods during the array. The expression fraction of the both circadian *eigengenes* represent around 31% (Fig. 1b). The phases difference between them is almost 6 hs, or equivalently $T/4$ where T is the period of the oscillation. This means that every gene with circadian expression can be described in terms of these *eigengenes*, consequently, we interpret that v_2 and v_3 represent circadian oscillations in the expression dataset. In Fig. 1c we depict a line-joined graphs of the expression levels of *eigengenes* v_2 (red) and v_3 (blue), the fit dashed graphs of periodic functions corresponds to trigonome tricfunctions

$$0.07\cos[\pi x/T] + 0.34\cos[2\pi x/T] - 0.20\sin[2\pi x/T] \tag{4}$$

$$0.16\cos[\pi x/T] + 0.05\sin[\pi x/T] - 0.16\cos[2\pi x/T] - 0.21\sin[2\pi x/T] \tag{5}$$

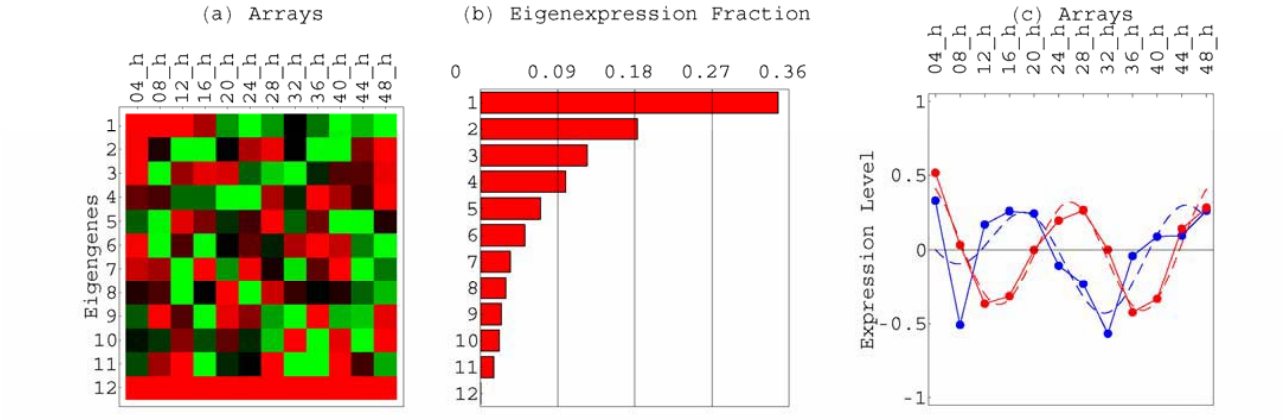

Fig. 1 (a) Raster display of the expression of 12 eigengenes v_i, with overexpression (red), no change in expression (black), and underexpression (green) derived from culture 1. (b) Bar chart of the fraction of eigenexpression p_1 of each eigengene. (c) Line-joined graphs of the expression levels of v_2 (red) and v_3 (blue) in the 12 arrays, and dashed graphs of trigonometric function (4) and (5) of period T~24 hs.

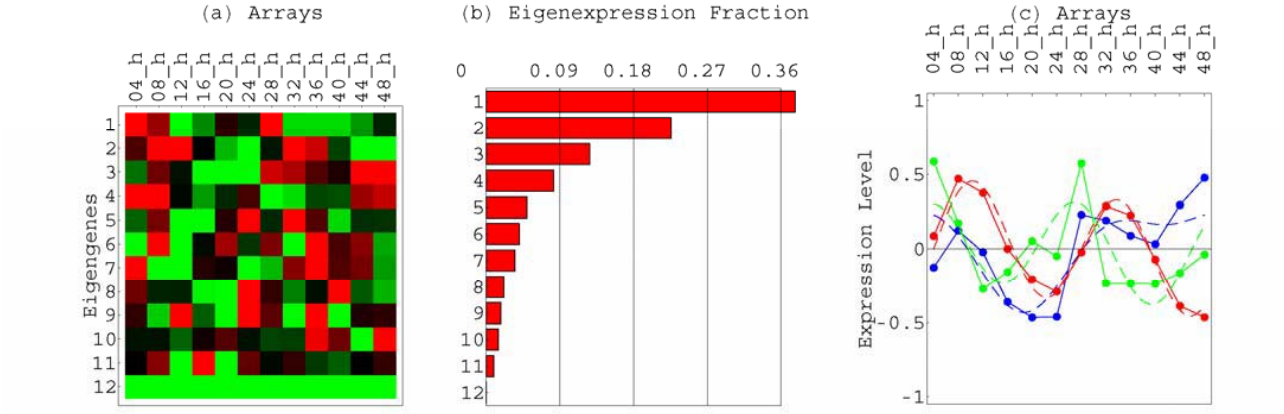

Fig. 2 (a) Raster display of the expression of 12 eigengenes v_i, with overexpression (red), no change in expression (black), and underexpression (green) derived from culture 2. (b) Bar chart of the fraction of eigenexpression p_1 of each eigengene. (c) Line-joined graphs of the expression levels of v_1 (green), v_2 (red) and v_3 (blue) in the 12 arrays, and dashed graphs of their respective fit trigonometric function.

On the other hand, when considering culture 2, the PCA analysis indicates that there are three *eigengenes* related to circadian expression. Fig. 2a displays the 12 *eigengenes* in 12 arrays, while in Fig. 2b we show a bar chart of the fractions of *eigenexpression*. We can see that the first 3 *eigengenes* v_1 (green), v_2 (red) and v_3 (blue) captures about 70% of the overall expression. These *eigengenes* describes oscillatory behavior during the array. These *eigengenes* are not the same *eigengenes* that obtained from culture 1, however they differ in the phases.

We consider the expression of the 2542 genes in the subspace spanned by *eigengenes* v_2 and v_3, which is inferred to approximately represent all circadian expression oscillations. We associated to each gene the parameter r as

$$r = \sqrt{\text{correlation with } v_2^{\,2} + \text{correlation with } v_3^{\,2}} \qquad (6)$$

Indicating the distance from origin of the (v_2, v_3) plane to the point representing the gene expression in that plane. One may expect that genes that have almost all of their expression in this subspace with $r \sim 1$, where; have a circadian rhythm, and that genes that have almost no expression in this subspace, with $r \sim 0$, are not regulated by the circadian clock at all. If we consider genes $r \geq 0.9$ we are able to identify 78 genes which exhibited strong circadian rhythm. Fig. 3 show the correlation of each gene with v_2 and v_3. The 78 selected circadian genes (red circles) are out of the dashed circle with,

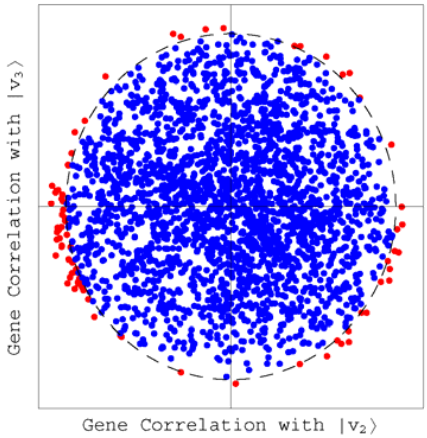

Fig. 3 Correlation of each gene with v_2 and v_3, for all genes in the subspace associated with the circadian cycle. Red dots correspond to the genes which have high projection onto this subspace.

while $r = 0.9$ *no*-circadian genes are in blue.

PCA also can aids in data visualization. In this sense, we also have sorted all circadian genes according to the phase, defined by arctan (v_2/ v_3). Fig. 4 depict the normalized gene expression profile for these circadian genes. All these genes were also classified by Kucho *et al*. as circadian genes [12].

IV. Discussion and Conclusions

We have shown that PCA provides a useful mathematical framework to process and to model genome-wide expression data, in which both the mathematical variables and operations may be assigned biological meaning. Application of PCA to the Cyanobacterium Synechocystis dataset reveals that PCA can identify periodic patterns in time series data. In conclusion, we propose that PCA should be added to the current arsenal of technique being used for processing and visualization data.

V. Acknowledgement

CL is fellow of ANPCyT (Argentina) and LD is researcher of CONICET (Argentina).

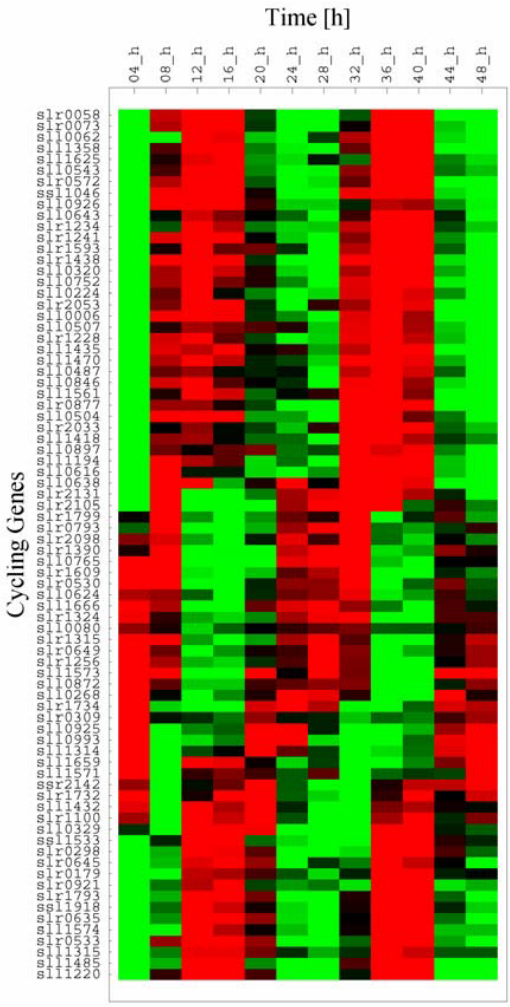

Fig. 4 Expression profile of circadian genes. Relative expression at each time point was normalized to mean zero and amplitude one. Overexpression (red), no change in expression (black), and underexpression (green).

References

[1] M.B. Eisen, P.T. Spellman, , P.O. Brown, D. Botstein, ¨Cluster analysis and display of genome-wide expression patterns¨. *Proc. Natl Acad. Sci.* vol. 95, pp. 14863-14868, Dec. 1998.

[2] G.S. Michaels, D.B. Carr, M. Askenazi, S. Fuhrman, X. Wen, R. Somogyi, ¨Cluster analysis and data visualization of large-scale gene expression data¨. *Pacific Symposium on Biocomputing* vol. 3 pp. 42-53, 1998.

[3] P.T. Spellman, G. Sherlock, M.Q . Zhang, V.R . Iyer, K . Anders, M.B Eisen, , P.O. Brown., D. Botstein, B. Futcher , ¨Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization¨. *Mol Biol Cell* vol. 9 pp. 3273-3297, Dec. 1998.

[4] P. D'haeseleer, S. Liang, R. Somogyi, ¨Genetic network inference: From co-expression clustering to reverse engineering¨. *Bioinformatics* vol. 16 pp. 707-726, 2000.

[5] R.J. Cho, et al., ¨A genome-wide transcriptional analysis of the mitotic cell cycle¨. *Mol. Cell* vol. 2 pp. 65-73, 1998.

[6] S. Chu, J. DeRisi, et al., ¨The transcriptional program of sporulation in budding Yeast¨. *Science* vol. 282 pp. 699-705, 1998.

[7] J.L. DeRisi, V.R. Iyer, P.O. Brown, ¨Exploring the metabolic and genetic control of gene expression on a genomic scale ¨. *Science* vol. 278 pp. 680-686, Oct. 1997.
[8] A. Basilevsky, ¨Statistical Factor Analysis and Related Methods¨, *Theory and Aplications, John Wiley & Sons, New York* , 1994.
[9] S.G. Hilsenbeck, W.E. Friedrichs, R. Schi®, P. O'Connell, R.K. Hansen, C.K. Osborne, S.A.W. Fuqua, ¨Statistical analysis of array expression data as applied the problem of tamoxifen resistance¨. J *Natl Cancer Institute* vol. 91 pp. 453-459, 1999.
[10] J. Vohradsky, X.M. Li, C.J. Thompson, ¨Identification of prokaryotic developmental stages by statistical analyses of two-dimensional gel patterns¨. *Electrophoresis* vol. 18 pp. 1418-1428, 1998.
[11] J.C. Craig, J.H. Eberwine, J.A. Calvin, B. Wlodarczyk, G.D. Bennett, R.H. Finnell, ¨Developmental expression of morphoregulatory genes in the mouse embryo: an analytical approach using a novel technology¨. *Biochem Mol Med* vol. 60 pp. 81-91, 1997.
[12] K. Kucho, K. Okamoto, Y. Tsuchiya, S. Nomura,, M. Nango, M. Kanehisa, M. Ishiura, ¨Global Analysis of Circadian Expression in the Cyanobacterium Synechocystis sp. Strain PCC 6803¨. *Journal Of Bacteriology* vol. 187 pp. 2190-2199, Dec. 2004.
[13] O. Alter, P.O. Brown, D. Botstein, ¨ Singular value decomposition for genome-wide expression data processing and modeling¨. *Proc. Natl Acad. Sci. USA* vol. 97 pp. 10101-10106, June 2000.