Quantifying cyclicity

In order to objectively analyze cell-cycle regulated gene expression, a numerical cyclicity score can be constructed for the expression profile of each gene. Suppose that the data have already been preprocessed and normalized, and let $Z_1, \ldots, Z_m$ denote the expression levels across the m time points for a given gene under consideration. Also suppose that the culture is sampled evenly at interval Δ minutes and has a nominal interdivision time of τ minutes. We describe two methods that have been used in several studies to identify cyclically-expressed genes. (In one published study [52] results of the two methods were combined).

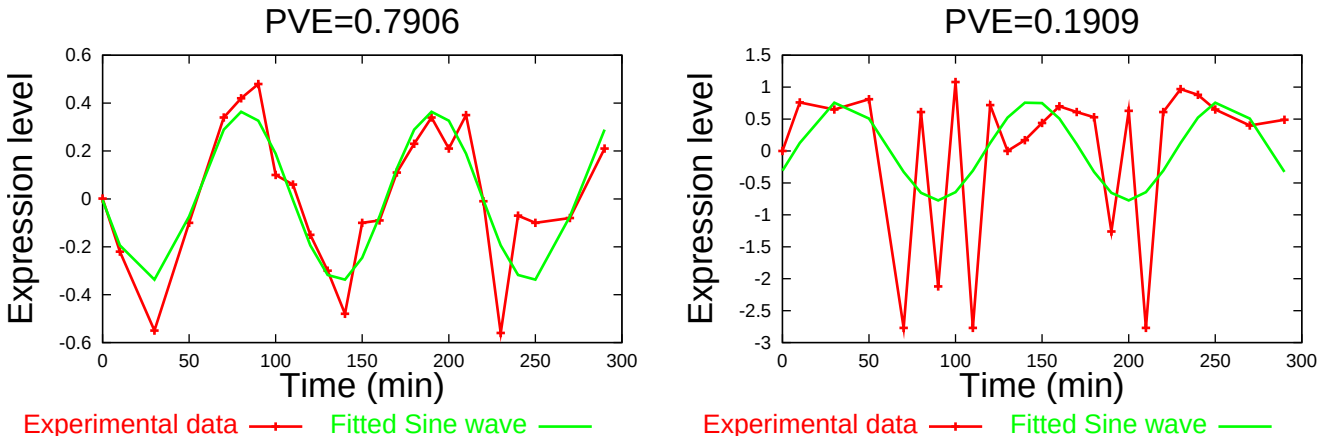
Fourier approach

For given values of μ (baseline), α (amplitude), and phase (ϕ), the following sine wave can be constructed

$$f_k(\mu, \alpha, \phi) = \mu + \alpha \sin(2\pi\Delta k/\tau + \phi),$$

and the fit $F(\mu, \alpha, \phi) = \sum_k (Z_k - f_k)^2$ to the experimental data can be determined. For each gene, optimal values μ^*, α^*, and ϕ^* can be numerically determined such that $F(\mu^*, \alpha^*, \phi^*)$ gives the smallest possible value for F. Then the "fitted waveform" $\hat{Z}_k = f_k(\mu^*, \alpha^*, \phi^*)$ can be calculated, and writing $R_k = Z_k - \hat{Z}_k$ yields the decomposition $Z_k = \hat{Z}_k + R_k$, where $\hat{Z}_k$ is perfectly cyclic and R_k is perfectly non-cyclic. This leads to the variance decomposition $\mathrm{var}(Z) = \mathrm{var}(\hat{Z}) + \mathrm{var}(R)$, where $\mathrm{var}(X) = \sum (X_i - \bar{X})^2/n$ is the statistical variance. Since all three variances must be non-negative, smaller values of $\mathrm{var}(R)$ relative to $\mathrm{var}(Z)$ imply greater cyclicity in Z. This suggests using the variance ratio $\mathrm{var}(\hat{Z})/\mathrm{var}(Z)$ (known as PVE or "proportion of variance explained" [by a sine wave]) as a numerical measure of cyclicity.

The PVE falls between zero and one, and values closer to one indicate greater cyclicity. The following figure shows two examples of yeast expression data at two PVE levels, where the experimental data Z (red) and the fitted sine wave $\hat{Z}$ (green) are superimposed.



Correlation approach

To identify cyclic gene expression using correlation, a set of genes agreed upon as cell-cycle regulated is identified, and the genes peaking in a given cell cycle phase are averaged. Correlation coefficients are then computed between the average expression profile for genes peaking in each cell cycle phase, and the expression levels of the gene under consideration. The largest among these correlation coefficients is used to quantify the cyclicity of the gene under consideration.

Scaling and magnitude of expression

It is important to note that Fourier PVE and correlation do not take account of the magnitude of change in gene expression, but rather are normalized to the total variation of expression in a given gene. Put a different way, the expression levels Z_i may be scaled by any nonzero constant without changing the results of either analysis. Thus a gene that varies by 50% relative to its mean level and a gene that varies by 500% relative to its mean level will have the same cyclicity score as long as they exhibit the same pattern of internal cyclicity relative to their mean levels. More generally, the amplitude of variation and the PVE are not coupled. For instance, in the above plots, the gene on the left has greater PVE (more perfect fit to a sine wave) but lower amplitude of cyclicity compared to the gene on the right. The opposite relationship can occur as well.

As an alternative to the PVE, the unnormalized variance $\mathrm{var}(\hat{Z})$ is used to form the composite cyclicity score for the yeast analysis [52]. This measure gives cyclic genes with greater amplitude higher scores than equally cyclic genes with lower amplitude. However this is not a perfect solution to the scaling problem, since writing $\mathrm{var}(\hat{Z}) = \mathrm{PVE} \cdot \mathrm{var}(Z)$ reveals that a highly variable gene can receive a high cyclicity score even if it is only slightly cyclic. Rather than forming the product or some other composite of PVE and variance, we recommend considering PVE and variance as distinct factors characterizing the level of cyclic variation in a gene. Both values must be high in order for a gene to be cyclically expressed in a biologically meaningful way.

Specification of the nominal doubling time

The Fourier approach requires the specification of a numerical value for the nominal doubling time τ. There are examples of disagreement in the literature about how these values are specified. Aach and Church [1] propose alternate doubling times for the yeast data [52], and it is possible that the doubling time for human fibroblasts [9] may be too short. A related issue is that the results of the Fourier analysis may not be stable when calculated for a single value of τ, so the yeast data Fourier analysis was carried out for a range of values of τ and the results were averaged [52].

The use of a sine-wave as the ideal waveform

The use of the sine wave as the ideal waveform for assessing cyclicity can be criticized in that it does not arise from fundamental biological considerations. In fact, truly cyclic gene expression is not expected to exhibit sine-like fluctuations (i.e., fluctuations that are symmetric above and below the mean value) since there is a floor of zero in the measurements. A logarithmic transform produces much more symmetrical cycles. A measure of robustness is provided by the fact that the Fourier PVE will be high for cyclic patterns that are not sinusoidal. For example, the "boxcar" function which is 1 when the sine-wave is positive and -1 when it is negative has PVE 0.83. A sine-wave raised to the fifth power, which has much more localized peaks than an ordinary sine-wave, has PVE 0.79.

Statistical Analysis

The assessment of statistical significance in a microarray experiment is complicated by the fact that large numbers of genes are measured. For example, while a PVE of 0.7 would be considered highly significant in an assay of a single gene, if 5000 genes are assayed at 12 time points, 22 genes would be expected to have cyclicity scores exceeding 0.7 even if all expression levels are completely random.

Randomization analysis

A standard technique for assessing whether the level of cyclicity in an experiment is greater than expected by chance is to produce an artificial data set in which all cyclicity arises as an artifact of the noise, and then compare the level of measured cyclicity in the artificial data set to the level of measured cyclicity in the actual data set. The artificial data set can be constructed by permuting or resampling the observed values, or by simulating independent values from a distribution such as the normal distribution.

Let S_p be the p^{th} quantile of the cyclicity scores for the artificial data set (so that fraction p of the cyclicity scores are less than or equal to S_p and fraction $1-p$ of the cyclicity scores are greater than S_p), and let T_p be the corresponding quantile for the actual data. It is common to state the ratio S_p/T_p or the ratio $S_p/(S_p+T_p)$ as the "false positive rate". If T_p is larger than S_p for p close to 1, then the observed cyclicities are stronger than expected by chance. This can be visualized by forming a scatterplot of S_p against T_p, and noting whether the

upper tail falls below the 45° diagonal. Such a quantile-quantile plot is called a *QQ* plot and there are published examples of this type of analysis [42,49,50].

In a more sophisticated analysis, multiple artificial data sets can be constructed so that error bars can be placed around the S_p values. In fact, the upper quantiles of the artificial cyclicity scores are quite variable, so it is important to repeat the simulation several times and take the mean quantile as S_p.

While widely used, it should be noted that there are several limitations to this approach to statistical inference. One limitation is that the artificial data set is constructed independently across the genes, while in truth there are significant correlations. If these correlations tend to be positive, randomization methods may overstate the significance level. Similarly, for each gene the artificial data set is constructed independently across the time points. In fact, due to the slowly varying nature of the cell culture (even for non-cell-cycle genes), there are significant serial correlations in the expression levels of a given gene over time. These autocorrelations may be positive or negative, depending on the sampling interval, Δ. Thus a consequence of correlation across genes and across time points may be to impute significance to data in which the cycles are actually due to experimental noise and non-cell-cycle related biological variation.

Influence of sampling interval on statistical significance

It is important to note that the number of samples per cycle Δ/τ has an important influence on statistical significance, with greater values of Δ/τ giving greater significance for a given PVE. For example, the probability of observing PVE $= .5$ by chance when 12 time points are measured has order 10^{-2}, while when 24 time points are measured the same probability has order 10^{-4}.

Cut points and threshold values

The cyclicity measures described here are found to vary continuously over their range in experimental data. That is, there is no sharp cut-off such that some proportion of genes fall distinctly above the cutoff while the remainder fall below the cutoff. In practice it is possible to select a threshold that incorporates most known positives while excluding as many negatives as possible, but it has always been found that a few known positives are so far down the list that lowering the threshold to include them would also include many genes that are highly unlikely to be cell-cycle regulated. Alternatively, it is possible to appeal to statistical principles such as the "false discovery rate" (FDR) to define a cutoff. For most applications, however, it is not crucial to specify a precise value such that only genes with cyclicity exceeding the value are said to be cyclic.