Gene expresseion oscillation analysis

redacted

11/02/2020

Introduction

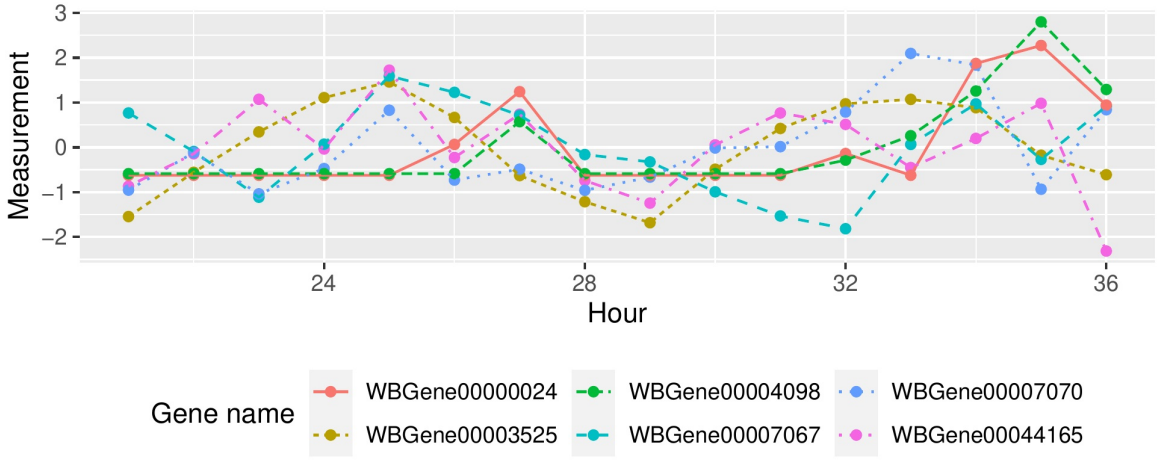
The data, which is from the field of biology, are collected by gene expression measures. In particular 20182 genes are measured hourly from the 21-st hour to 36-th hour, hence data is 20182×16.

The analysis of the dataset is in the following way:

1. To see whether a gene oscillates with period of 8 hours, compute spectral density of the measurements (after subtracting the least-square linear trend), and check if the density at frequency $\omega_2 = \frac{2\pi}{8}$ is one of the 2 largest densities among that at $\omega_j = \frac{2\pi j}{16}$ for $j = 1, \ldots, 8$. If not, then the gene is considered as randomly perturbing without oscillations, at least not the 8-hour-period oscillation this project concerns. This is to check whether oscillation with period 8 is the leading oscillation.
2. In addition, fit two linear models for the remaining genes, one with sinusoidal terms (denoted as M_1) and one without (as M_0). Since the two models are nested, check the BIC's of the two models, and only if the BIC of M_1 is smaller the gene is considered to be oscillating. This is to check whether the oscillation is statistically significant.
3. For these considered oscillating genes, 10449 genes in total, normalize the 16-hour measurements to be zero mean and unit variance.
4. Use k-means algorithm to cluster the genes according to Euclidean distance of the normalized M_1 coefficients. The number of clusters can be determined by elbow method or BIC criterion.
5. Summary the clusters and analyze the oscillation phase and amplitude by fitting linear models again.

Data visualization

As stated above the data is pre-processed to be zero mean and unit variance. And below is a glimpse of the 6 example genes after normalization.



Oscillation check

Leading oscillation check

The leading oscillation can be justified by the spectral density. In particular, denote Y_t as the measurement at time t and assume r_t is the residual of Y_t regressed on t, the spectral density could be computed by

$$w_j = \frac{1}{T}|\sum_{t=1}^{T} r_t \exp(-i \cdot 2\pi jt/T)|^2,\ j = 1, \ldots, T-1,$$

where $T = 16$ is the length of time, i is the imaginary unit, and $|\cdot|$ is the module of complex numbers. Indeed $w_j = w_{T-j}$ hence only w_j for $j = 1, \ldots, T/2$ need to be computed.

The spectral density represents that the $\frac{w_j}{\sum w_k}$ $(j = 1, \ldots, T/2)$ proportion of the vector is contributed by the sinusoidal series with period $\frac{T}{j}$.

Then check one of the two criteria:

1. w_2 is one of the 2 largest numbers in $\{w_1, \ldots, w_{T/2}\}$;
2. $w_2/\sum w_j$ exceeds certain threshold Δ.

According to the explanation above, only those genes that w_2 dominates (or almost dominates) all w_j will be considered to oscillate with a period of 8.

I used the first criterion here and eliminated all those genes that the leading sinusoidal term does not have period 8.

Significancy check

Then I consider if including the sinusoidal term with period 8 in linear model actually enhance the model's interpretability. This is done by computing the BIC of two linear models (with or without sinusoidal term of period 8):

$$M_0 : Y_{it} = C_i + k_i t + e_{it},$$
$$M_1 : Y_{it} = C_i + k_i t + \alpha_i \cos(\frac{2\pi}{8}t) + \beta_i \sin(\frac{2\pi}{8}t) + e_{it}.$$

If the BIC increases after adding the sinusoidal term, i.e. $BIC(M_1) > BIC(M_0)$, the gene will be classified as **NOT** oscillating.

There are now altogether 10449 genes that are considered to have sufficient oscillation pattern with period 8.

Clustering

Then the genes are clustered according to the linear model coefficients. The linear models are defined as

$$Y_{it} = C_i + k_i t + \alpha_i \cos(\frac{2\pi}{8}t) + \beta_i \sin(\frac{2\pi}{8}t) + e_{it},\ e_{it} \sim N(0, \sigma_i^2),$$

where i is the gene index, $t \in \{21, \ldots, 36\}$ represents the hour, and e_i is the innovation noise. In particular this could be further simplified to

$$Y_{it} = C_i + k_i t + A_i \sin(\frac{2\pi}{8}t + \theta_i) + e_{it},\ e_{it} \sim N(0, \sigma_i^2),$$

where

$$A_i = sign(\beta_i)\sqrt{\alpha_i^2 + \beta_i^2},\ \theta_i = \arctan(\frac{\alpha_i}{\beta_i}).$$

Then a coefficient vector is extracted:

$$c_i = (C_i, k_i, A_i, \theta_i),\ c_i^0 = (C_i/3, 10k_i, A_i, \theta_i).$$

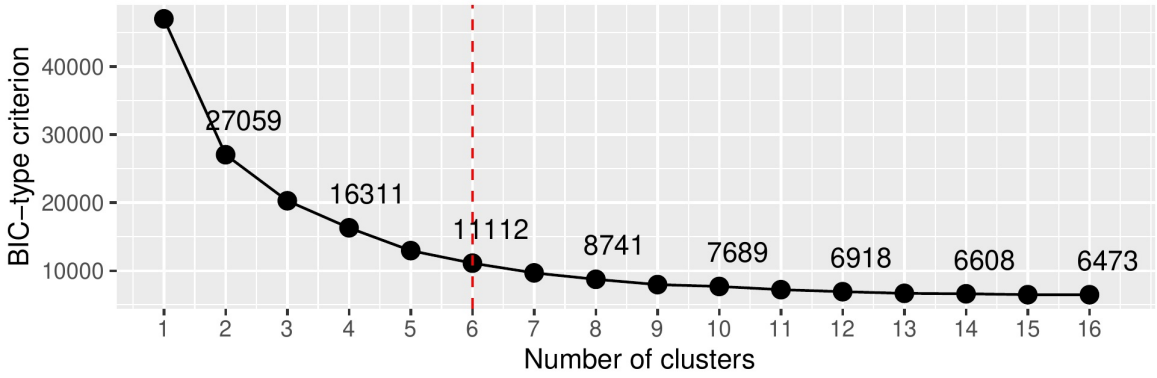
The vector c_i^0 is transformed to force the elements in c_i^0 have similar magnitudes and equally considered in the following clustering step.

K-means algorithm is applied to the partially normalized coefficient vectors $\{c_i^0\}_{i=1}^n$ where the number of clusters is adaptively determined by elbow method.

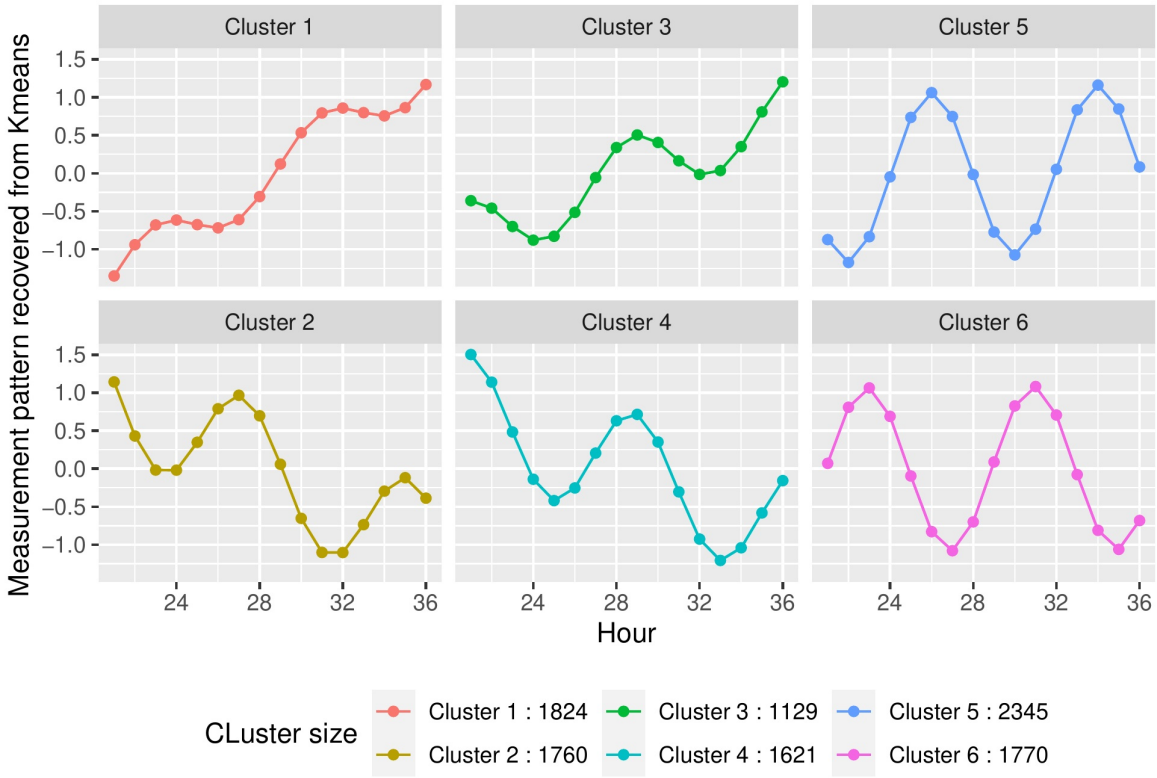
For example a BIC-type criterion is introduced, which is the total within-cluster variance adding a penalty term:

$$BIC(K) = \sum_{j=1}^{K} \sum_{m \in C_j} \|\overrightarrow{c_{j,m}^0} - \overrightarrow{\mu}_j\|^2 + 2\log(n) \cdot k \cdot 16$$

where K is the total number of clusters, C_j is the index set of the j-th cluster, and n is the total number of genes. I set K from $1, \ldots, 16$, and have the following plot:



Then I chose the model with 6 clusters and had the following normalized measurements recovered from the clusters' center coefficients vectors:



From the above, which is how the measurements fluctuate in each cluster, the following conclusions may be made:

1. The genes in clusters 1 and 3 have increasing trends, i.e. tend to be more active in the 16 hours;
2. The genes in clusters 2 and 4 have decreasing trends, i.e. tend to be less active in the 16 hours.
3. In cluster 5 and 6, the genes displayed somehow standard oscillation patterns with ignorable linear trends, but their phases are almost opposite.

Oscillations

Then for all the genes that are classified in the same cluster, fit the following linear model:

$$Y_{it} = C_{j(i)} + k_{j(i)}t + \alpha_{j(i)}\cos(\frac{2\pi}{8}t) + \beta_{j(i)}\sin(\frac{2\pi}{8}t) + e_{it},\ e_{it} \sim N(0, \sigma^2_{j(i)}),$$

where i is the gene index, $t \in \{21, \ldots, 36\}$ represents the hour, $j(i)$ is the label for the cluster of gene i, and e_i is the innovation noise.

Simplify the model to

$$Y_{it} = C_{j(i)} + k_{j(i)}t + A_{j(i)}\sin(\frac{2\pi}{8}t + \theta_{j(i)}) + e_{it},\ e_{it} \sim N(0, \sigma^2_{j(i)}),$$

where

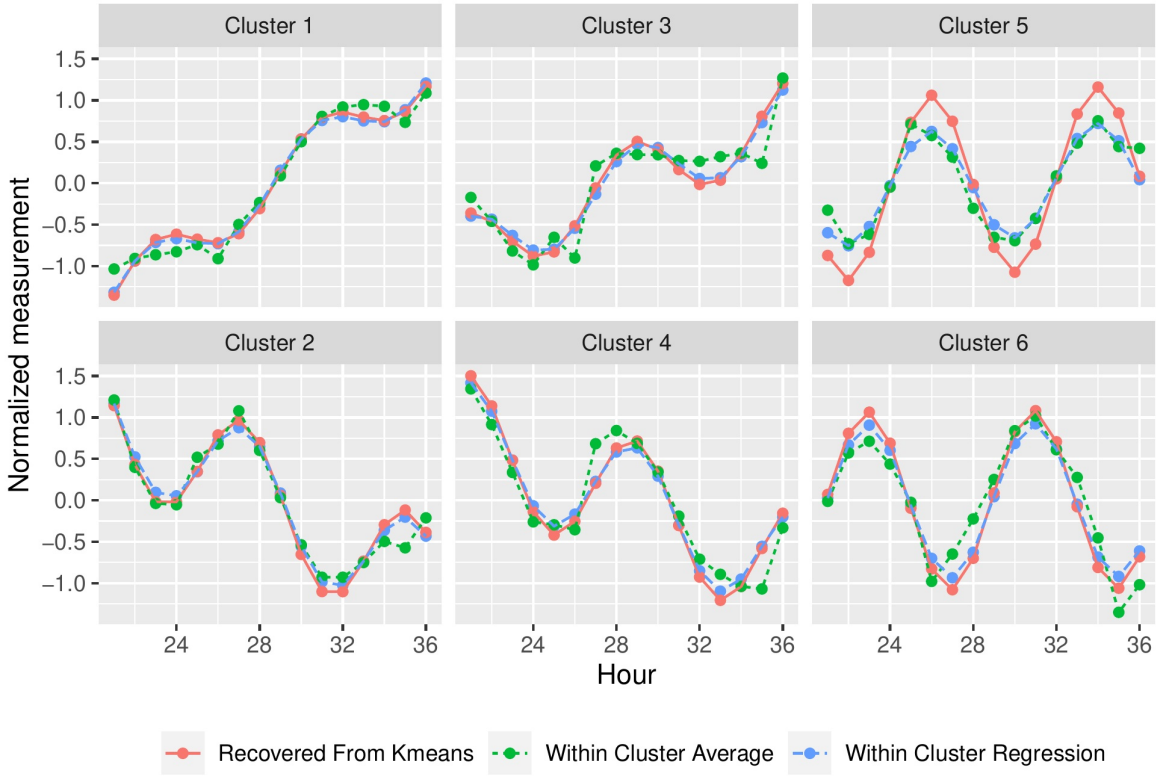
$$A_{j(i)} = sign(\beta_{j(i)})\sqrt{\alpha^2_{j(i)} + \beta^2_{j(i)}},\ \theta_{j(i)} = \arctan(\frac{\alpha_{j(i)}}{\beta_{j(i)}}).$$

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
C	-5.25	3.86	-3.08	2.81	-0.35	-0.06
k	0.18	-0.14	0.11	-0.10	0.01	0.00
α	0.17	-0.57	-0.32	-0.52	0.04	0.62
β	-0.26	0.37	-0.27	-0.42	0.67	-0.69
σ	0.56	0.72	0.80	0.68	0.83	0.72
A	-0.31	0.68	-0.42	-0.67	0.67	-0.93
θ	-0.56	-1.00	0.87	0.89	0.05	-0.73

Just as expected:

1. In clusters 1 and 2, the absolute values of slopes are significantly large relatively to those of the other clusters;
2. The absolute values of slopes in clusters 3 and 4 are slightly less than the ones in clusters a and 2, but larger than those in clusters 5 and 6;
3. The difference of phases in the last 2 clusters is approximately $\theta_5 - \theta_6 \approx -\frac{3\pi}{4}$. This could be verified if the sign-difference of A_5 and A_6 are moved into the phase, i.e. if $A_i < 0$ then set $A_i = -A_i$ and $\theta_i = \theta_i + \pi$.

For the fitted results, the following plot is provided.



Conclusions & Discussions

1. The decision whether a gene has oscillation of period 8 is determined by (1) checking the spectral density for the leading frequency, or (2) checking BIC of linear models with/without sinusoidal terms. However the first step of decision is still manually made by fixing a proportion threshold or giving the number of the largest frequencies to be checked. There ought to be some deterministic data-free approaches to be applied.

2. The same issue exists for finding the optimal number of clusters, K. The choice of K was again determined by directly looking at the elbow plot. It is possible to play with different choices of K and analyze the results for better performances on individual genes.

3. The manually normalization of the coefficient vector c_i to c_i^0 for clustering could be replaced by standardizing each element. For instance, normalize $\{C_i\}_{i=1}^n$ to have zero mean and unit variance, then normalized $\{k_i\}_{i=1}^n$ and so on. After k-means clustering the centers could be transformed back according to the original mean and variance for analysis.

4. If another oscillation period of interest is fixed and known beforehand, this approach could be applied with minor modifications.

5. If the oscillation period is unknown and to be determined, the first step of checking leading frequency according to spectral density could be modified to finding the leading frequencies (with large w_j), and the model's procedure could be then continued by working on these found leading frequencies.

6. It is possible that some of the genes in the cluster may not perfectly match the average cluster pattern, but it is still guaranteed that a gene is closest to the cluster where it is classified. When conducting individual analysis on a specific gene, a close look is necessary.