Identification and Clustering of Genes Expressed In Circadian Rhythms

Abstract:

Some genes are expressed in patterns following circadian rhythms. The ARSER and COSOPT algorithms both identify many genes which appear to be expressed in roughly 24-hour periods in a set of expression profiles for mouse livers, but the ARSER algorithm identifies more cycling genes. We use the k-means, bottom-up hierarchical, and CLICK algorithms to cluster genes which appear to be expressed in circadian rhythms, and evaluate them by measuring their stabilities. There appears to be a cluster-able signal within the data, and k-means and CLICK are the most stable for the dataset. The stability of the clusterings across different clustering algorithms is low, showing that the algorithms cluster the data differently.

Introduction:

Circadian rhythms are patterns that repeat roughly once every 24 hours. Many genes are expressed in patterns that follow circadian rhythms, such that they are more expressed at certain times of the day and less expressed at others. In living things, circadian rhythms are controlled both by external cues and internal regulatory mechanisms. Complex regulatory networks for genes expressed in circadian rhythms have been shown to exist, and a gene called *Mop3* has been identified as a key regulator of circadian rhythms in mammals (Bunger et al., 2000). It would be useful to understand these regulatory networks because they may yield insight into the regulation of circadian behaviors in humans, and may help develop

treatments for things such as sleeping disorders.

To learn about regulatory networks that controll genes which are expressed in circadian rhythms, we identify genes which appear to have circadian expression patterns, and arrange them into groups of genes which appear to be regulated similarly. Using microarray chips, we can measure a gene's relative expression levels over time. A gene's expression profile is the series of expression measurements over time. We use an algorithm called ARSER (Yang and Su, 2010), which computes the spectrum for each gene to identify potential periods, and then uses a harmonic regression to fit a sinusoid to the gene's expression profile.

To identify genes which are regulated similarly, we use and evaluate three different clustering algorithms to cluster the genes based upon the parameters of their sinusoids (period, amplitude, and phase.) We implement each clustering algorithm for a set of 24,034 gene expression profiles collected from mouse livers over three days, and evaluate the results of each algorithm by measuring the stability of each clustering when the input dataset is subsampled.

Approach:

Given a set of gene expression profiles, we first attempt to identify genes which appear to be expressed following circadian rhythms. Several methods have been proposed for this problem. The method which is currently the most wildly-used is a pattern-matching algorithm called COSOPT (Straume, 2004) COSOPT first removes the linear trend from the time series, then attempts to fit a series of 101 cosine waves to the data and returns the pattern that most closely matches the input data. Pattern-matching algorithms such as COSOPT are efficient to compute, but often fail to identify patterns that are not sinusoidal.

A newer method called ARSER has recently been proposed (Yang and Su, 2010)

which combines time-domain and frequency-domain approaches to the problem. ARSER uses spectral analysis to identify periods in the expression profiles, and then uses a harmonic regression to fit a sinusoid to the expression profile.

ARSER uses autoregressive spectral analysis to identify genes which appear to be expressed periodically. First, an autoregressive model is fit to the time series. After computing the autoregressive model, ARSER estimates the spectrum of the model using the parameters of the model instead of the original data, and considers periods between 20 and 28 hours (a too-wide period period window inhibits ARSER's ability to identify periods.) Prominent periods in the gene expression time series appear as peaks in the autoregressive spectrum. If there is no periodicity in the data there will be no peaks in the spectrum, and if there are multiple periods within the window there will be multiple peaks present in the spectrum. ARSER calculates p and q values for each gene, which describe the confidence that the gene's expression profile exhibits periodicity.

After the autoregressive spectral analysis step, ARSER computes a harmonic regression for the expression profile using the periods found in the spectrum. A mean value, phase, and amplitude are estimated for the expression of each gene. ARSER also computes an R² value which represents the goodness-of-fit of the harmonic regression.

The ARSER process is summarized in Figure 1. Unlike COSOPT, ARSER is able to identify periodicity in genes whose expression is not sinusoidal. An important limitation of the ARSER algorithm is that it can only use expression profiles where the time points are evenly spaced.

To identify genes which appear to be expressed similarly, we implement and evaluate three different clustering algorithms for the parameters (period, amplitude, and phase) of the

sinusoids of our cycling genes. The amplitude is clustered on a log scale. The clustering task consists of arranging items into groups such that there is some sort of similarity between the items in each group. Different clustering algorithms exploit different mathematical properties of the data. The algorithms that we use are k-means, bottom-up hierarchical, and CLICK.

The k-means and bottom-up hierarchical algorithms consider the distance between elements when computing clustering. We use the Euclidean distance metric,

$$D_{i,j} = \sqrt{\sum_{x=1}^{N} (i_x - j_x)^2}$$

where *i* and *j* are the vectors representing the two data points, *N* is the number of dimensions in *i* and *j*, and $D_{i,j}$ is the distance between *i* and *j*. The Euclidean metric represents the shortest geometric distance between the two points.

K-means clustering is initialized by randomly placing k number of centroids on the same manifold as the data. Iteratively, each data point is assigned to the nearest centroid and centroids are re-computed to be the mean of their assigned data points until convergence. K-means clustering is fast to run and tends to produce evenly-sized, circle-shaped clusters.

For bottom-up hierarchical clustering, each point is initially assigned to its own cluster. Iteratively, the two closest clusters are merged until there is only one remaining cluster, forming a hierarchy of clusters. We use the average link metric to compute the distance between two clusters, where the mean value for each cluster is found, and then the Euclidean distance between the mean values of each cluster is computed.

The third clustering algorithm that we use is the CLuster Identification via Connectivity Kernels (CLICK) algorithm (Sharan and Shamir, 2000). The CLICK algorithm represents the

data points as a complete graph where vertices represent individual genes and edges are weighted to reflect the similarity between each pair of gene expression profiles. Minimumweight cut operations are used to identify tight groups of highly-similar genes, referred to as kernels, and then the kernels are expended into a full clustering by iteratively assigning singletons (genes not in a kernel) to them.

For the CLICK algorithm, the vectors representing each gene are normalized so each field has a mean of 0 and standard deviation of 1. The similarities between each pair of genes is measured using the dot product of the two vectors. We estimate the mean μ_T and standard deviation σ_T similarities of genes who are mates (in the same cluster), as well as the mean μ_F and standard deviation σ_F similarities between non-mates by dividing the similarities into two groups using the k-means algorithm. The probability that any two vertices are mates, p_{mates} , is estimated using the number edges in each cluster.

The process by which kernels are identified is referred to as basic-CLICK. In the basic-CLICK algorithm, a complete graph *G* is constructed representing all genes to be clustered. For efficiency, edges with a similarity lower than μ_F are trimmed from *G*. The edges are given weights that reflect the probabilities that they are between mates:

$$w_{i,j} = \log \frac{p_{mates} f(s_{i,j} | i, j \text{ are mates})}{(1 - p_{mates}) f(s_{i,j} | i, j \text{ are not mates})}$$

where $w_{i,j}$ is the weight of the edge between genes *i* and *j*, and *f* is the mates and nonmates probability density function. A minimum weight cut *C* is calculated, which is the set of edges in *G* with the lowest total weight that disconnects the *G* when the edges are removed.

To determine whether the current graph is a kernel or not, we calculate the probability that all of the edges in *C* belong to mates, and the probability that all edges belong to non-

mates. For edges that were trimmed out, μ_F is used as the similarity. If the probability that all edges are mates is higher, we consider *G* to be a kernel. Otherwise, we run basic-CLICK on each of the partitions of the cut to recursively find kernels in the graph. If a single gene is split off from *G* at any point, we add it to a list of singletons.

After kernels are identified with basic-CLICK, an adoption step adds singletons to kernels when the similarity between the kernel and singleton is higher than $\mu_T - \sigma_T$. Iteratively, basic-CLICK and the adoption step are run using the remaining singletons until there is no change.

After convergence of the basic-CLICK and adoption step, similar clusters are merged together. Center points for each cluster are computed, and if the similarity between the center points for two clusters is higher than $\mu_T - \sigma_T$, the clusters are merged. After the merging step, one last adoption is performed.

Empirical evaluation:

We use a dataset from the Chris Bradfield lab to evaluate ARSER and the three clustering algorithms. Samples were taken from a population of mice over a duration of 3 days. At 4-hour intervals, for a total of 12 samples, 3 mice were sacrificed and samples were taken from their livers. RNA from the livers at each time interval were pooled and measured with a microarray. The values for each time point in the dataset are the log₁₀ ratio of the gene's expression for that time point over the average expression for all time points. In total, expression levels for 24,034 genes were measured.

The dataset was run through both ARSER and COSOPT implementations to identify genes with periodic behavior. The dataset is not evenly-sampled with respect to time, so we

use linear interpolation to estimate points at regular time intervals for ARSER. ARSER produces a q value that can be used to filter cycling genes, while COSOPT produces a p value. Yang and Su (2010) use ARSER's q value and COSOPT's p value to compare the two algorithms.

Table 1 shows the number of cycling genes in the sets produced by ARSER and COSOPT at different p and q values, as well as the number of genes which are in both sets, and genes which are in only one of the sets. ARSER consistently identifies more cycling genes than COSOPT, but also consistently misses many of the genes identified by COSOPT.

Figures 2-4 depict plots of k-means, hierarchical, and CLICK clusterings of sinusoids computed for the mouse liver dataset, all with 5 clusters. Visually, the most of the clusters in each of the clusterings appear to be well-separated.

For the CLICK clustering, the distribution of edge similarities should be bimodal so edges between mates and non-mates can be separated. The distribution in our set appears to be more unimodal (Fig. 5). Using k-means to separate the edges into mates and nonmates still appears to produce a CLICK clustering which is comparable to k-means and hierarchical clusterings.

It is often difficult to determine whether a clustering is representative of true order in the data, or if it is arbitrary output of the clustering algorithm. To evaluate the clusterings of our data, we measure the stabilities of the clusterings. If a clustering changes minimally when the underlying data is perturbed, that clustering is stable, and is likely indicative of true order within the data (Ben-Hur et al., 2002). To determine the number of clusters in a dataset, we can cluster subsamples of the data using different k values, and choose the k value with the

highest stability.

To measure the stabilities of the clusterings, we first take random subsamples of the mouse liver dataset, such that the subsamples contain 75% of the genes in the original dataset. Each of the subsamples are clustered using each clustering algorithm, and the clusterings of the subsamples are compared to each other. The average distance between each clustering is a measurement of stability for that clustering method.

The distance between two clusterings is a measurement of the difference between the two clusterings. To compute the distance between two clusterings, we use the CDistance algorithm (Coen et al., 2010). CDistance frames the problem of measuring cluster similarity as the optimal transportation problem, which asks *what is the cheapest way to move a set of masses from sources to sinks, who are some distance away?* Cost is defined as the total mess times distance moved.

The naïve solution to the transportation problem is that the sources do not cooperate, and all of the sources distribute their masses proportionally to each of the sinks. The optimal solution is that the sources cooperate, and all agree to the transport their masses with a globally minimal cost. The cost of the optimal solution is computed with a linear program.

The similarity distance between two sets of points indicates the degree to which cooperation reduces the cost of moving from a source onto a sink, and is defined as the cost of the optimal solution over the cost of the naïve solution. The clustering distance (CDistance) is an application of the similarity distance to clusterings, which considers the two clusterings to be the sources and sinks.

The k-means and CLICK algorithms are both more stable than hierarchical clustering, and both have an average distance between subsample clusterings of about 0.31. Table 2

shows that the clusterings appear to be more stable around 5 clusters, indicating that there are about 5 true clusters present in the data.

The stability across clusterings is very low. The average distance between clusterings of the full dataset using the three clustering algorithms and k = 5 is very high, at 0.8 (Table 2).

Discussion:

Both ARSER and COSOPT identify a sizable number of genes which appear to be expressed in following circadian rhythms. ARSER identifies many genes which COSOPT fails to find, which may be explained by ARSER's ability to detect non-sinusoidal patterns that COSPOT cannot. COSOPT also finds many genes that ARSER misses, and we have not yet identified the cause of the discrepancy.

ARSER can only use datasets in which there is a fixed amount of time between samples, and our mouse liver dataset is not evenly-sampled. We use linear interpolation to estimate some points so we can use ARSER. Linear interpolation can "smooth out" the data, causing the oscillation to appear to have a lower amplitude than it actually has. To mitigate this, interpolation via smoothing splines could be used instead of linear interpolation. A newer algorithm called LSPR (Yang et al., 2011) has recently been published by the authors of ARSER, which could be used to detect circadian expression in our data which is not evenly-sampled.

The stability measurements for the clusterings using each of the single clustering algorithms are low, indicating that there is a cluster-able signal within the data (Table 2). K-means and CLICK clustering have the highest similarities, indicating that they are the most suitable for clustering our dataset.

The average distance between clusterings of the full dataset using different clustering

algorithms (with 5 clusters) is very high at 0.8 (out of 1,) indicating that the clusterings are not stable across algorithms. The algorithms all appear to cluster the data differently. The high value may also be attributed to CLICK's notion of singletons, which is not present in the other two clustering algorithms. For the distance measurements between clusterings, we create a cluster of singletons.

Clustering is an important first step for learning about gene expression regulatory

networks, as it shows us which genes may be regulated together. Network models and

knockout experiments can be implemented to learn more about the regulatory networks.

References:

- 1. Ben-Hur, A., Elisseeff, A., and Guyon, I. A stability based method for discovering structure in clustered data. In Pacific Symp. on Biocomputing, 2002.
- Bunger, M. K., L. D. Wilsbacher, S. M. Moran, C. Clendenin, L. A. Radcliffe, J. B. Hogenesch, M. C. Simon, J. S. Takahashi, and C. A. Bradfield. 2000. Mop3 is an essential component of the master circadian pacemaker in mammals. Cell. 103:1009-1017.
- 3. Coen, Michael H., M. Hidayath Ansari, and Nathanael Fillmore. "Comparing Clusterings in Space." Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel. 2010. Print.
- 4. Sharan, Roded, and Ron Shamir. "CLICK: A Clustering Algorithm with Applications to Gene Expression Analysis." ISMB (2000): 307-16. Print.
- 5. Straume, M. 2004. DNA Microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning, Methods Enzymol, 383, 149-166.
- Yang, R., and Z. Su. 2010. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. Bioinformatics. 26:i168-i174. doi: 10.1093/bioinformatics/btq189.
- Yang, Rendong, Chen Zhang, and Zhen Su. "LSPR: an Integrated Periodicity Detection Algorithm for Unevenly Sampled Temporal Microarray Data." Bioinformatics (2011). Print.

Tables and figures:



Figure 1: **Diagram of ARSER process.** First the data is detrended and smoothed, then fitted with an AR spectral model to find the period. A harmonic regression is computed, and sinusoid parameters are returned. (Source: Yang and Su, 2010)

ARSER q, COSOPT p	Total genes in ARSER set	Genes only in ARSER set	Genes in both sets	Genes only in COSOPT set	Total genes in COSOPT set
0.005	327	149	178	62	240
0.01	568	200	368	116	484
0.05	2423	774	1649	395	2044

Table 1: **Comparison of Cycling genes identified by ARSER and COSOPT.** *Counts of cycling genes for both ARSER and COSOPT and overlap counts for different ARSER q and COSOPT p values. ARSER consistently finds more cycling genes than COSOPT, but also misses many of COSOPT's genes.*

k	k-means	hierarchical	CLICK	All algorithms
3	0.38	0.85		
4	0.4	0.6		
5	0.31	0.47	0.31	0.80
6	0.33	0.48		
7	0.38	0.48		
8	0.38	0.45	-	
9	0.35	0.42		
10	0.34	0.42		

Table 2: **Stability measurements for subsample clusterings.** Stability measurements by clustering algorithms and number of clusters. The number of clusters is automatically determined by the CLICK algorithm. The "All algorithms" value is calculated using different clusterings of the full dataset instead of subsamples. Lower values are more stable.

K-means clusters in wild type (q ≤ 0.01), k = 5



Figure 2: Plot of k-means clustering of sinusiods for cycling genes (q <= 0.01) in wild type mice. Colors represent cluster assignment.



Agglomerative hierarchical clusters in wild type (q <= 0.01), k = 5

Figure 3: Plot of hierarchical clustering of sinusiods for cycling genes $(q \le 0.01)$ in wild type mice. Colors represent cluster assignment.

CLICK sinusiod clusters in wild type



Figure 4: Plot of CLICK clustering of sinusiods for cycling genes (q <= 0.01) in wild type mice. Colors represent cluster assignment. Singletons are depicted as a cluster.



Figure 5: Distribution of CLICK similarity measurements between genes. Colors indicate edges divided into mates and non-mates by k-means.