An Improved Method for Identifying Cell Cycle-regulated Genes in Yeast

Eric Bair

December 6, 2001

1 Introduction

In a 1998 paper, Spellman *et. al.* (1991) attempted to create "a comprehensive catalog of yeast genes whose transcript levels vary periodically within the cell cycle." To identify these genes, they grew several cultures of yeast under various experimental conditions. They used DNA microarrays to measure the expression levels of 6187 known or predicted genes. After the data was collected, they used a periodicity and correlation algorithm to identify cell cycle-regulated genes.

Although Spellman *et. al.*'s paper represented an important advance in this area, I feel that their methodology is subject to criticism. I will discuss several criticisms of their method, and I will describe a new method to address these criticisms.

2 Description of the Problem

We hope to discover yeast genes whose expression level varies periodically within the cell cycle. Such genes should be overexpressed at certain times during the cell cycle and underexpressed at other times. Moreover, when we plot the expression level of the gene versus time, we would expect the plot to have a sinusoidal pattern. Figure 1 is a plot of the expression level of the *rfa1* gene versus time. Brill and Stillman (1991) have shown that the *rfa1* gene is cell cycle regulated through traditional methods. Note the sinusoidal pattern in the data. Compare this expression pattern with the expression pattern of the *efb1* gene, shown in Figure 2. This gene is not known to be cell cycle-regulated, and we observe that there is little or no indication of a sinusoidal pattern in the expression data. The problem is therefore to distinguish between these two expression patterns.

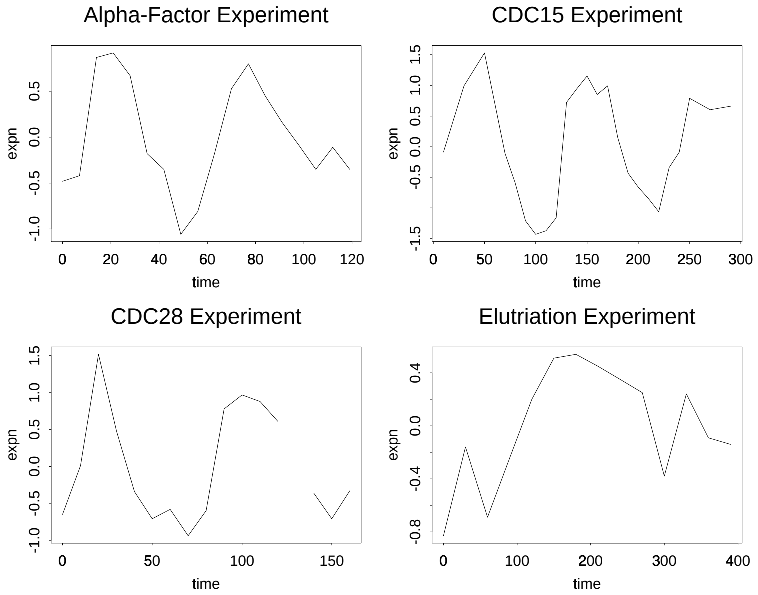

Figure 1: Expression profile of the *rfa1* gene

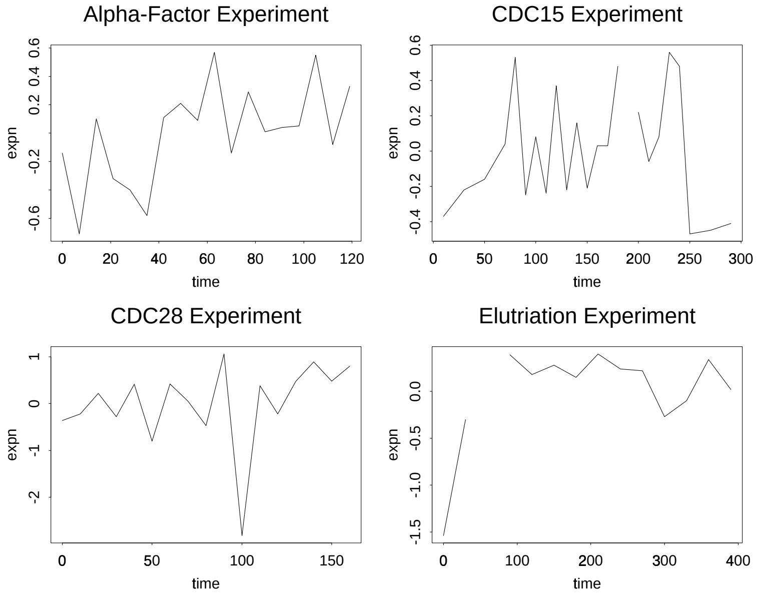

Figure 2: Expression profile of the *efb1* gene

3 Spellman *et. al.*'s Method

The following is Spellman *et. al.*'s description of their method for identifying cell cycle-regulated genes: "Data for each gene in the α factor time series were extracted from the database and were normalized so that the average $\log_2$(ratio) over the course of the experiments was equal to 0. A Fourier transform (see (1) and (2)) was applied to the data series for each gene, and the resulting vector (C) was stored for each gene, where ω is the period of the cell cycle, t is the time, Φ is the phase offset, and ratio(t) is the ratio measurement at time t. We found that the magnitude of the Fourier transform (4) was unstable for small variations of ω, so we averaged the vectors of the transform over a range of 40 values, which were evenly spaced around the estimated division time for the experiment (66 ± 11). We initially set the value of Φ to 0.

$$A = \sum \sin(\omega t + \Phi) \log_2(\text{ratio}(t)) \tag{1}$$

$$A = \sum \cos(\omega t + \Phi) \log_2(\text{ratio}(t)) \tag{2}$$

$$C = \langle A, B \rangle \tag{3}$$

$$D = \sqrt{A^2 + B^2} \tag{4}$$

"The expression profile of each gene across the experiments was then correlated to five different profiles representing genes known to be expressed in G1, S, G2, M, and M/G1 using a standard Pearson correlation function. The profiles for known gene classes were identified by averaging the $\log_2$(ratio) data for each of the genes known to peak in each of the five time periods. The peak correlation score was defined as the highest correlation value between the data series for each gene and each of the profiles. The vector calculated by the Fourier transform was scaled by the peak correlation value.

"The above process was repeated for the *cdc15* experiment (ω varying between 60 and 80) and for the *cdc28* data (ω varying between 80 and 100) from Cho *et. al.* (1998). The *cdc28* data set was first converted to ratio style measurements by dividing each measurement by the average value of the measurements for that gene. Before this step it was necessary to exclude some data points that appeared to be aberrant. Any data value where the two values on either side were threefold different in the same direction were excluded. Each gene thus had three vector scores (one for each of the three analyzed data series).

"To generate a single vector for each gene, we added the vectors for each experiment together. However, the value of Φ for the three experiments should not be the same, because the experiments start at different points in the cell cycle. Therefore, before combining the vectors for the three experiments, constants

Φcdc15 and Φcdc28 (relative to the α factor experiment), were calculated for the *cdc15* and *cdc28* experiments, respectively, that maximized, for the known genes, the average magnitude of the summed vectors. The elutriation data were not included, because it was not possible to calculate a Φ that maximized the values of more than a handful of the known genes. The α factor and cdc15 vectors were multiplied by 0.7, so that they would not unduly contribute to the final 'aggregate CDC score,' which was calculated by taking the magnitude of this final vector.

"Genes were ranked by their aggregate CDC scores, and the list was examined to identify the positions of known cell cycle genes within it. We selected a threshold CDC score that was exceeded by 91% of known cell cycle-regulated genes. Altogether 800 genes met or exceeded this CDC score."

4 Criticisms of Spellman *et. al.*'s Method

Spellman *et. al.*'s methodology represents an important first step in the search for cell cycle-regulated genes. However, their method is subject to criticism on several grounds. First, their algorithm is difficult or impossible to reproduce. The descriptions of many of the steps of their algorithm are vague. For example, they state that "...constants, Φcdc15 and Φcdc28 were calculated for the *cdc15* and *cdc28* experiments, respectively, that maximized, for the known genes, the average magnitude of the summed vectors." However, they do not describe how these constants were determined. And although their raw data is available on the Internet, the source code for their algorithm is not. Without a detailed description of the gene-finding algorithm and without the source code, their results would be difficult to reproduce. This could be problematic if someone wanted to repeat the experiment, or if someone wanted to use their method to identify cell cycle-regulated genes in a different organism.

Second, there is no rigorous mathematical or biological justification for their algorithm. Intuitively, one would expect that a cell cycle-regulated gene should have larger Fourier coefficients and be correlated with known cell cycle-regulated genes. However, there is no guarantee that this will be the case. In particular, one can imagine a gene that is cell cycle-regulated, but whose mRNA transcript levels show only small variation with respect to time. The Fourier coefficients associated with such a gene would probably be small because the corresponding expression levels are small, in spite of the fact that the expression level is periodic with respect to time. Such a gene may be misclassified using Spellman *et. al.*'s method.

Furthermore, Spellman *et. al.* applied their method to 104 yeast genes that are already known to be cell cycle regulated. They calculated the "CDC score" for each of these 104 genes, and took the 10th percentile of these scores to be their significance threshold. Although this idea seems reasonable, it makes it difficult to verify that the algorithm is producing meaningful results. Any reasonable algorithm should correctly identify a high percentage of these genes as cell cycle-regulated. If an algorithm fails to do so, we would conclude that the algorithm is fatally flawed. However, if one applies an arbitrary procedure to these 104 genes and chooses a certain percentile of the output to be the significance threshold, the procedure will always identify a fixed percentage of these genes as cell cycle regulated, even if the algorithm is complete nonsense. For example, if I simply assigned a random number to each gene, and chose the 10th percentile of these random numbers as my threshold, I would still correctly identify 90% of these genes as cell cycle-regulated!

Finally, even if Spellman *et. al.*'s method is completely valid, it is difficult to quantify the uncertainty in the results. Spellman *et. al.* attempted to estimate the global false positive rate for their method by permuting the expression data for each gene. Presumably, this would destroy any periodicity that existed in the data. If their algorithm produced a significant "CDC score" when applied to this permuted data, it is obviously a false positive. After permuting the data for all 6,187 genes, they calculated the "CDC scores" for each permuted data set. Of the 6,187 permuted data sets, 75 produced significant "CDC scores." Thus, they estimated that about 75 of the 800 genes they identified as cell cycle-regulated were false positives, and so the estimated false positive rate is slightly less than 10%.

However, there are problems with this procedure. As noted above, the yeast cultures were grown under different conditions, and the expression levels of the genes were recorded separately for each set of conditions. In the above procedure, Spellman *et. al.* apparently permuted all data for a given gene. In other words, data from different experiments were combined. In addition to removing the periodicity of the data, such a permutation could produce a smaller "CDC score" because the data from different experiments may not be of comparable orders of magnitude. Thus, they would underestimate the false positive rate. A better estimate of the false positive rate would only permute the data within the same experimental conditions.

Additionally, with Spellman *et. al.*'s method, there is no way to make a confidence statement about an individual gene. They estimate the global false positive rate, but there is no way to determine the probability that a given gene is a false positive. Such information may be important for future research. For example, one can imagine that a researcher might like to perform further analysis on the 100 genes that are most likely to be cell cycle-regulated. Without making confidence statements about individual genes, such an analysis would be impossible.

5 My Proposed Alternative

In order to overcome these shortfalls of Spellman *et. al.*'s method, I devised an alternative. My method is outlined below. S-Plus code for my algorithm, as well as the complete data set that I used, are available at http://www–stat.stanford.edu/ ebair/cellcycle.tar.gz.

1. We examine the first gene on the array, and begin by considering the data for only one of the four experiments. We perform a Fast Fourier Transform on this data. (Prior to performing the transformation, we discard any missing values in the data.)

2. If there is an underlying periodic signal in the data, we should be able to represent this signal with a small number of high-order Fourier coefficients. Thus, we set each of the Fourier coefficients from step 1 to 0 except for the first three (including the constant term). Then we invert the transformation.

3. Step 2 gives us an approximation of the original data. We compute the sum of squared errors of this approximation.

4. Randomly permute the data 1,000 times and repeat steps 1, 2, and 3 for each permutation. This will produce a vector of length 1,000 consisting of the sum of squared errors that result from applying step 2 to each permuted vector.

5. As discussed earlier, permuting the data should destroy any periodicity present in the data. Moreover, the Fourier approximation that we obtain in step 2 should be a good approximation if the data is periodic, but should be a poor approximation otherwise. Thus, if the gene is cell cycle-regulated, we would expect the error associated with the original data to be much smaller than the the errors associated with the permuted data. Otherwise, we would expect all the errors to be approximately the same. We can formulate the problem as a hypothesis test as follows: Let the null hypothesis be that there is no periodicity in the data, and let the alternative hypothesis be that the data is periodic. Using the permutation test procedure described in Efron and Tibshirani (1993), we can compute a p-value for this hypothesis test: Simply count how many times a permuted data set produces a smaller sum of squared errors than the original data, and divide this number by 1,000. In this manner, we obtain a p-value for each set of experimental conditions for each gene.

6. A slight difficult arises from the fact that there are four sets of experimental conditions, and four data sets for each gene. In other words, the above procedure will give us four p-values for each gene. It is unclear how we should classify a gene if some of the p-values are significant but others are not. To resolve this problem, we examine the smallest of the four p-values. Under the null hypothesis, each p-value should be uniformly distributed

on the interval [0,1], so the minimum of the four p-values should have a Beta(1,4) distribution. (See Rohatgi (1976) for a derivation of this result.) We use this fact to compute a final p-value for each gene: Denote the smallest of the four p-values by x. Under the null hypothesis, x should have a Beta(0,1) distribution. To compute the p-value, we compute the probability that a Beta(0,1) is less than or equal to x.

Applying this procedure to each gene in the yeast genome, we obtain a p-value for the null hypothesis that the gene is not cell cycle-regulated.

The only remaining question is to decide how small the p-value for a gene must be in order to be considered "significant." This significance threshold must be high enough that we correctly identify most or all of the cell cycle-regulated genes, but not so high that the algorithm produces a large number of false positives. To reconcile these competing goals, I set the maximum acceptable false positive rate to a fixed value α. I then found the largest possible significance threshold such that the associated false positive rate does not exceed α. Benjamini and Hochberg (1995) show how to find this threshold: Let $p_{(1)} \leq \cdots \leq p_{(6187)}$ be the ordered, observed p-values for the 6187 hypothesis tests. Let

$$\hat{k} = \max \{ k \mid p_{(k)} \leq k/6187 \cdot \alpha \}$$

and reject $p_{(1)}, \cdots p_{(\hat{k})}$. This procedure maximizes the number of significant genes identified while maintaining a false positive rate less than or equal to α.

6 Results and Conclusions

As noted above, Spellman *et. al.* identified 800 genes as cell cycle-regulated with an estimated false positive rate of 10%. They correctly identified 95 of the 104 genes that are known to be cell cycle regulated. When I set the false positive rate to be 10%, my procedure identified 3037 genes as possibly cell cycle regulated, including 98 of the 104 known genes. I set the false positive rate to 1% and repeated my procedure. The algorithm found 997 genes that may be cell cycle-regulated.

I believe these numbers demonstrate that my algorithm has much greater power than the procedure that Spellman *et. al.* used. If anything, the increase in power is even greater than the above results indicate. Recall that Spellman *et. al.* estimated the false positive rate of their procedure by permuting all the data points associated with a particular gene, whereas I only permuted the data within each set of experimental conditions. Thus, my estimated false positive rate is more conservative than Spellman *et. al.*'s estimate.

Furthermore, with my method, one can make confidence statements about individual genes. Each gene has an associated p-value, so we have an idea of the probability that a given gene is cell cycle-regulated. Thus, if a researcher wanted to restrict her or his attention to the "most significant" genes, it would be easy to do so. With Spellman *et. al.*'s method, however, there is no obvious way to do this.

We must use caution when interpreting the results of the two methods. The p-values and estimated false positive rates for my method, as well as Spellman *et. al.*'s method, do not directly measure the probability that a gene is truly cell cycle-regulated. They measure how well a given gene's expression data fits our model. (i.e. The data should be periodic and sinusoidal.) However, there is no biological requirement that the expression level of a cell cycle-regulated gene should follow this pattern. It is also possible that a gene could have a sinusoidal expression pattern even if it is not cell cycle-regulated. Thus, it is incorrect to say that 90% of the 3037 genes I identified must be cell cycle-regulated. We have only shown that about 90% of these 3037 genes show evidence of sinusoidal expression patterns.

This distinction is important because both my method and Spellman *et. al.*'s method may be overestimating the number of cell cycle-regulated genes. Price *et. al.* (1991) estimated that some 250 cell cycle-regulated genes might exist. Obviously, we cannot be certain whether this estimate is accurate. However, it does seem improbable that 3037 yeast genes are cell cycle-regulated. This is nearly half of the yeast genome. My method, or any other computational method, cannot be expected to perfectly classify each gene in the entire genome. However, I feel that it represents a quick and relatively simple procedure to identify such genes before resorting to laborious and expensive laboratory experiments.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. *Journal of the Royal Statistical Society, Series B* **75**: 289–300.

[2] Brill, S.J. and Stillman, B. (1991). Replication factor-A from Saccharomyces cerevisiae is encoded by three essential genes coordinately expressed at S phase. *Genes and Development* **5**, 1589–1600.

[3] Cho, R.J., *et. al.* (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **2**, 65–73

[4] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall: New York.

[5] Price, C., Nasmyth, K., and Schuster, T. (1991). A general approach to the isolation of cell cycle-regulated genes in the budding yeast, *Saccharomyces cerevisiae. Journal of Molecular Biology*, **218**: 543-556.

[6] Rohatgi, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics.* John Wiley and Sons: New York.

[7] Spellman, P., *et. al.* (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273–3297.