FRED HUTCHINSON
CANCER RESEARCH CENTER

Periodic genes of the yeast *Saccharomyces cerevisiae*: A combined analysis of five cell cycle data sets.

Tata Pramila[1], Wei Wu[1,2], William Stafford Noble[2], Linda Breeden[1]

[1] Fred Hutchinson Cancer Research Center, Seattle, WA, USA (tpramila@fhcrc.org, lbreeden@fhcrc.org)

[2] Department of Genome Sciences, University of Washington, Seattle, WA, USA (noble@gs.washington.edu)

We have carried out three microarray experiments across the cell cycle of budding yeast, using spotted cDNA arrays and alpha factor to induce synchrony. Two data sets (called alpha30 and alpha38) are dye swap technical replicates with a sampling interval of 5 minutes and a total of 25 time points. The third data set (alpha26) has a sampling interval of 10 minutes and a total of 13 time points. These data sets have been processed using the error model of the Rosetta Resolver v. 3.2 Expression Data Analysis System and can be downloaded at the links below. The raw data sets are also available from GEO Database, where they have been deposited for permanent public accessibility. Also available are six RNA measurements in which the same RNA was labeled with both dyes and hybridized to itself for purposes of error estimation (same vs same). We have used the Resolver normalized data from experiments alpha30, alpha38, and three other data sets from the public domain that were generated by alpha factor, cdc15 (Spellman et al., (1)) and cdc28 (Cho et al. (2)) synchronization to identify periodic transcripts.

All five data sets have also been analyzed by a permutation-based method (PBM5) published by de Lichtenberg et al (4). This method ranks each transcript by combining two permutation-based statistical tests for periodicity and magnitude of regulation, respectively. The scoring penalizes genes that only display one property, i.e., high amplitude fluctuations with no periodicity or very low amplitude periodic oscillation. The method also computes a gene specific number called the "peak time" for each transcript in each data set, describing when in the cell cycle the gene is maximally expressed. In addition, a combined peak time is calculated as a weighted average from all five data sets, and the error associated with that calculation is also provided. Peak times are expressed as percent of the cell cycle, and zero is set at the M/G1 boundary.

These values and the heat maps of the data from our three new data sets can be visualized using a variant of the PRISM program (6) via the links provided below. For comparison, the linked tables also include the results of previous efforts to identify periodic transcripts from the three public domain data sets (1, 5). The Prism display enables you to select the information to be viewed and to sort the data based on rankings or expression times.

Prism Visualization of three alpha-factor synchronized yeast cell cycle microarray data sets

- **Data set alpha 30 (5 min interval): row-normalized, original, tab-delimited text**
- **Data set alpha 38 (5 min interval): row-normalized, original, tab-delimited text**
- **Data set alpha 26 (10 min interval): row-normalized, original, tab-delimited text**

Documentation

Data Sets

Yeast strain: W303a: *ade2-l, trp1-1, can1-100, leu2-3, -112, his3-11, -15, ura3*
Growth media: YEP glucose

All three data sets are alpha-factor synchronized microarray time series spanning two cell cycles. Data set alpha 26 has a sampling interval of 10 minutes, while data set 30 and 38 have a sampling interval of 5 minutes. Data sets alpha 30 and alpha 38 are dye swap techincal replicates, but the data has been adjusted so that all three data sets have a consistent value (and color) for peaks and troughs. All values are $\log_{10}$.

Data Set		**Labeling convention**
26		t_0/SS, t_{10}/SS ... t_{120}/SS: Cy5/Cy3
Dye Swap Technical Replicates	**30**	t_0/SS, t_5/SS ... t_{120}/SS: Cy3/Cy5
	38	t_0/SS, t_5/SS ... t_{120}/SS: Cy5/Cy3
SS: steady state, t: cell cycle time points		

If you are visiting this page for the first time, you can click "Click here" to view the microarray data sets with default display options.

Alternatively, you may enter a session identifier, which is assigned the first time you view the page, into the textfield and press the button labeled "Retrieve data." The data will then be displayed with previous options you have specified (See 4: Selecting output options).

Primary output

The primary output page displays a heat map representing the expression matrix. Columns in the matrices correspond to columns in the data set. Each row in the matrix corresponds to a single gene, and the corresponding gene ID, gene-specific scores from various computational methods (described below), and annotation appears to the right. '-' indicates that no data were available. The gene ID is linked to the Saccharomyces Genome Database. Optionally, the user may configure the page to link instead to GenBank, UniGene or the Comprehensive Yeast Genome Database. Clicking on the matrix itself zooms in on a particular gene (See 3: Zooming in on a gene).

Scores from the following methods have been estimated:

- PBM5: Permutation Based Method analysis combining the data sets alpha 30,

alpha 38, alpha, cdc15 and cdc28. Ranking of the genes, peak times (calculated from each of these data sets as well as the combined peak) and the associated errors are provided.
- PNM5 posterior: Periodic Normal Mixture (PNM) model integrating the three public domain data sets plus data sets alpha 30 and alpha 38. All genes with a posterior probability above 0.95 are considered periodic and displayed in red.
- Spellman CDC: Aggregate Fourier analysis (CDC) score developed by Spellman et.al.
- shape-invariant: periodic genes are indicated with a list of data sets from which they were identified (5).

Note that the top of the page lists a numeric session identifier the first time you view this page. You should keep track of this number, because you may change the display options, and later you can use the session number to view the data set according to your own customized display options.

There is a button labeled "Change Display Options" at the top of the heat maps, you may click it to go to the page where you can specify your own display options (See 4: Selecting output options).

Zooming in on a gene

Clicking on the heat map matrix will take you to a gene-specific page. This page plots the expression level of this gene across two cell cycles. Flagged data are marked with green crosses.

Default output options

At the top of the right frame, there are several options that control the format of the output. These include the following:

- **Gene ID column** allows you to specify which column in your input file contains the gene IDs. This column will be used to search the catalog for corresponding annotations.
- **Filter** allows you to remove all genes that do not meet a specified criterion. By default, the top 1000 most periodic genes (ranked by PBM5) are included in the output.
- **Sort** allows you to sort the primary Prism output on one of the columns from the data matrix.
- **Secondary Sort** allows you to sort the primary Prism output on another column of the data matrix to break ties resulting from the primary sorting.
- **Color** allows you to select the colors representing high and low expression, respectively.
- **Flag value** allows you to specify what value is used to indicate corrupt or valid data.
- **Aliases** allows you to specify whether the alternate names of a given gene are included as part of its annotation.
- **Sampling interval** allows you to specify the time (in minutes) between data points, if this is time series data. This value is only used to make gene-specific plots of expression as a function of time. By default, this value is zero, and the plots are indexed by experiment number rather than time.

Once you have selected these options (or left them with their default values), press the button labeled "Go."

Visualizing a subset of the data

If you wish to visualize a specific subset of transcript profiles, you should download the tab-delimited text file of the data set, select the subset of profiles for the genes of interest, and save that subset as a tab delimited text file, with the same column headings, on your local computer. Then, go to http://noble.gs.washington.edu/prism and upload the data file. There you can use Prism to visualize the data as you wish.

References

1. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, and Futcher B. "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization" *Molecular Biology of the Cell* 9: 3273-3297, 1998. (Website)
2. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Garielian AE, Landsman D, Lockhart DJ and Davis RW. "A genome-wide transcriptional analysis of the mitotic cell cycle." *Molecular Cell* 2:65-73, 1998.
3. Lu X, Zhang W, Qin ZS, Kwast KE, and Liu JS. "Statistical resynchronization and Bayesian detection of periodically expressed genes" *Nucleic Acids Research*. 32:447-455, 2004.
4. de Lichtenberg U, Jensen LJ, Fausbøll A, Jensen TS, Bork P, Brunak S. "Comparison of computational methods for the identification of cell cycle regulated genes." *Bioinformatics* 21(7):1164-1171, 2005. (Website)
5. Luan Y and Li H. "Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data". *Bioinformatics*. 20:332-339, 2004. (Website)
6. Wu W and Noble WS. "Genomic data visualization on the web." *Bioinformatics.* 20(11):1804-1805, 2004.

Acknowledgment

This work was funded by NIH GM41073 to LLB and NIH HG003070 and NSF BDI-0243257 to WSN.

Fred Hutchinson Cancer Research Center
1100 Fairview Ave. N. PO Box 19024 Seattle, WA 98109