

Statistical Resynchronization and Detection of Periodic Transcripts

Xin Lu^{1,3}, Wen Zhang^{1,2,3}, Zhaohui S. Qin¹, Jun S. Liu*¹

¹Department of Statistics, Harvard University, Cambridge, Massachusetts 02138 USA

²Department of Biology, Kunming Medical College, Kunming 650031 China

³ Joint first authors

* To whom correspondence should be addressed: jliu@stat.harvard.edu

Abstract

A Periodic-Normal Mixture (PNM) model is proposed to fit the yeast cell cycle data of Cho et al. and Spellman et al. and to resynchronize the transcription profiles of cell-cycle related genes. Subsequently a two-component Beta mixture model is used to approximate the distribution of the PNM model fitting residues so as to compute the posterior probability for each gene to be cell-cycle related. Our result suggested that ~32% genes in yeast genome are likely to be periodically expressed, among which 822 genes met our stringent criterion. Of the 822 genes, 282 are absent from the list of 800 genes reported by Spellman et al., including many known important cell-cycle related functional genes. Phase matching of the 822 resynchronized expression profiles implied that the three synchronization methods might have brought cells to the same phase at the time of release.

1. Introduction

Transcription profiles of periodically expressed genes in *Saccharomyces cerevisiae* (budding yeast) observed from microarray experiments usually display a relatively clear pattern with a high and sharp peak (or trough) within the first cell cycle, and much flatter or even undetectable bumps in the ensuing cycles. This progressive synchrony decay in the sampled yeast cell population is caused by the diversity of the individual cell's growth rate.

2. Material and methods

The yeast cell cycle experiments examined in this study are those synchronized by *cdc28* [1], *alpha*, and *cdc15* [2], respectively. We deleted in each dataset 1000 genes with least variation and those with 25% or more missing values, 5510 genes passed the initial screening in at least two data sets and were used in the study. The remaining genes were normalized by row such that each row has mean zero and standard deviation one.

The cell cycle rate ρ of the yeast population is assumed to follow a Normal distribution: $\rho \sim N(\mu, \sigma^2)$. The level of transcription V of a periodic gene can be approximated by a linear combination of sinusoids through Fourier decomposition.

$$V(t\rho) = \sum_{k=1}^K (a_k \cos(kt\rho) + b_k \sin(kt\rho)) \quad (1)$$

Then the observed gene expression level at time t is a mixture of expression by the rate distribution, with additive noises:

$$Y(t) = \int V(t\rho)\varphi(\rho)d\rho + \varepsilon \\ = \sum_{k=1}^K (a_k \cos(kt\mu) + b_k \sin(kt\mu))e^{-\frac{1}{2}k^2t^2\sigma^2} + \varepsilon \quad (2)$$

We chose $K=3$ here to avoid overfitting.

To estimate the synchrony decay, we started from a selected set of periodically expressed genes identified by traditional methods [2]. The mean μ and standard deviation σ of cell cycle rate were inferred from the

expression level of these genes by minimizing:

$$\sum_g e_g^2 = \sum_g \sum_t (Y_g(t) - \int V_g(t\rho)\varphi(\rho)d\rho)^2 \quad (3)$$

where e_g^2 is called the model fitting residue sum of squares (RSS), and eqn (3) is the total RSS.

Based on the estimated μ and σ from the initial periodically expressed gene set, all genes in the three data sets were fitted by the PNM model, and the top 100 genes from each data set with the least RSS were selected to re-estimate μ and σ . This process is repeated until μ , σ and the top 100 genes become stable. The final μ and σ were used to calculate the Fourier decomposition parameter a_{gk} , b_{gk} and the RSS e_g^2 for all preprocessed genes.

The RSS of either a periodic or an aperiodic transcript g should follow a Beta distribution because of the standardization pre-processing of each gene's transcription profile and the Gaussian assumption. Therefore, the distribution of the RSSs in the whole data set could be approximated by a mixture of two Beta distributions. If we assume that the proportion of periodic transcripts was a fixed value across the three data sets, then the Maximum Likelihood Estimation (MLE) could be achieved by maximizing the total log likelihood function:

$$\sum_{i=1}^3 \sum_g \log(\gamma\beta(e_{gi}^2 | \theta_{1i}) + (1-\gamma)\beta(e_{gi}^2 | \theta_{2i})) \quad (4)$$

The posterior probability of a gene g that belongs to the periodic group can be formulated in the classical Bayesian manner:

$$p_g = \frac{\gamma \prod_{i=1}^3 \beta(e_{gi}^2 | \theta_{1i})}{\gamma \prod_{i=1}^3 \beta(e_{gi}^2 | \theta_{1i}) + (1-\gamma) \prod_{i=1}^3 \beta(e_{gi}^2 | \theta_{2i})} \quad (5)$$

The inter-experimental phase-shifts could be estimated by minimizing the matching errors of the profiles.

$$\sum_g \int_0^{2\pi} \{ [V_{g,28}(s) - V_{g,15}(s+m_1)]^2 + [V_{g,28}(s) - V_{g,6}(s+m_2)]^2 \} ds \quad (6)$$

3. Results and discussion

The cell cycle period T and the progression rate of synchrony decay estimated are listed in Table 1.

Table 1. The cell cycle period and synchrony decay in three microarray experiments

	PNM	Spellman[2]	Aach[3].	Zhao[4]
cdc28	83.2 ± 8.5	85		85
alpha	59.5 ± 5.2	66 ± 11	67.5 ± 6.5	58
cdc15	115.7 ± 11.1	110	119.0 ± 14.0	115

The two Beta distributions, corresponding to the periodic and the aperiodic transcripts, respectively, overlapped substantially. We limit our list of periodically expressed genes to those with 0.95 or higher posterior probabilities according to eqn (5). This stringent criterion yielded 822 genes. Among them 282 are absent from the list of 800 genes detected by Spellman et al., including many known important cell-cycle functional genes.

The relative phase shifts estimated by eqn (6) are: $m_1=2.1\%$ and $m_2=8.6\%$ of one cell cycle. These results imply that in the three experiments yeast cells might have restarted their cell cycles from roughly the same phase.

Saccharomyces Genome Database (SGD) Gene Ontology Term Finder analysis of these 822 genes showed that almost all of the top significant terms are involved in chromosome cycle, spindle cycle and bud cycle. On the other hand, far less periodically expressed genes were involved in cytoplasmic processes except for some transportation, signaling or protein modification functions.

References:

- [1] R.J. Cho, M.J. Campbell, E.A. Winzeler, et al. "A genome-wide transcriptional analysis of the mitotic cell cycle". *Mol. Cell*, **2**, 1998, pp. 65-73
- [2] P.T. Spellman, G. Sherlock, M.Q. Zhang, et al. "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization". *Mol. Biol. Cell*, **9**, 1998, pp. 3273-3297
- [3] J. Aach, and G.M. Church. "Aligning gene expression time series with time warping algorithms". *Bioinformatics*, **17**, 2001, pp. 495-508
- [4] L.P. Zhao, R. Prentice and L. Breeden, "Statistical modeling of large microarray data sets to identify stimulus-response profiles". *Proc. Natl. Acad. Sci. USA*, **98**, 2001, pp. 5631-5636