

Estimating Cell Cycle Phase Distribution of Yeast from Time Series Gene Expression Data

Jhoirene Clemente¹, Julie Ann Salido¹ and Raissa Relator²

¹ Algorithms and Complexity Laboratory, Department of Computer Science;

² Institute of Mathematics, University of the Philippines Diliman

Abstract. Two standard methods for estimating the cell cycle phase distribution of a yeast population are budding index analysis and one that uses fluorescence-activated cell sorter (FACS). However, both of these methods will require wet lab procedures. In this research, we propose a method that estimates the cell cycle phase distribution from the time series gene expression data.

Keywords: Cell cycle phase distribution, FACS, budding index analysis, time series gene expression data, spectral clustering, kernel k -means clustering.

1. Introduction

The development of microarray technology has supplied a large amount of data to the field of Bioinformatics. This technique is a key technology that facilitates genome-wide analysis of gene expression levels for gene function discovery and biomedical applications. This data in its raw form is difficult to understand, thus there is a need to apply data mining techniques to aid in analysis. Furthermore, analysis without biological significance is useless. So to assert the results made by in silico procedures, information about the samples are needed. This information is usually obtained by performing specific laboratory techniques.

Gene expression data is highly dependent on the state of the sample. The state may be the current cell cycle phase, phenotypic trait, or the tissue where the samples were taken. A single sample may have different gene expressions through time, thus this leads to the analysis of time series gene expression data.

Examples of time series data analysis are the studies made by [1, 7] where their purpose was to identify sets of genes that are periodically expressed at specific phases of a cell cycle in yeast. It was also necessary for both studies to specify the cell cycle phase at each timepoint. To identify the said information, the method used in [1] considered the size of the buds, the cellular position of the nucleus, and standardization of more than 20 transcripts whose mRNA fluctuations have been previously reported. Identification of the cell cycle phase distribution of yeast cells is also presented in [6] where two methods were proposed: one based on FACS and the other using budding index analysis, both of which require wet lab procedures. Details of these two methods will be discussed in Section 2.3.

Gene expression analysis results are highly dependent on basic information about samples, and not all available time series gene expression data include these information. In this paper we intend to develop a method of approximating the cell cycle phase distribution based on the time series gene expression data.

2. Definitions and Basic Notations

2.1. Gene Expression Data

Genes are the basic hereditary unit of living organisms. These are encoded in the chromosomes of an individual and dictate the biological processes which are carried out by proteins in a cell. Protein synthesis is dependent on the gene expression of an organism and gene expressions are measured using DNA microarrays. A microarray is a tiny square array with thousands of *probes*, whose expression values are measured through their luminosity. Each probe corresponds to a specific gene of interest. A microarray chip

contains the expression levels of approximately 20,000 gene transcripts [3]. The amount of gene expressed dictates how much proteins are synthesized and therefore responsible for the biochemical interactions taking place inside the cell.

Data mining techniques for gene expression data has long been studied in the literature. However, interpretations are still based from biological significance and are dependent on biological information of the sample.

2.2. Yeast Cell Cycle

The eukaryotic cell cycle is an ordered and periodic set of events that includes growth and synthesis of biochemical substances essential for living. The stages of the cycle are ordered as follows: pre-synthetic gap (G1), Synthesis (S), post-synthetic gap (G2), and Mitosis (M). During the G1 phase, the cell grows and takes in nutrients in preparation for the synthesis stage, and produces enzymes needed for DNA replication. The synthesis, also called the DNA synthesis phase, is marked by DNA replication, transcription and protein synthesis. The amount of DNA is effectively doubled after the synthesis phase. At the second gap phase, the cell grows and takes in nutrients as preparation for mitosis. Enzymes necessary for the production of microtubules are synthesized, along with other proteins needed for the cell division. The cell division stage is further characterized by 4 events: prophase, metaphase, anaphase and telophase. After the last phase, two daughter cells are formed.

2.3. Estimation of Cell Cycle Phase Distribution

The first of the two computational methods for estimating the cell cycle phase distribution of a budding yeast cell population presented in [6] uses a device called fluorescent-activated cell sorter (FACS). This method is based on the fact that the cell cycle phase is dependent on the amount of the DNA present in the cell. For instance, if the amount of DNA during the G1 phase is n , then the DNA amount during the G2 is $2n$, while the amount of DNA during S is somewhere in between the amounts present during G1 and G2. The main purpose of the FACS device is to measure the DNA content in a cell [6]. Hence, this method produces an age distribution of the cell population as an output. Based from a histogram, the cell cycle phase is identified by manually marking the range for each phase. Thus, the variability of the DNA amount in the S phase makes it difficult to find a good estimate.

The second distribution estimation method, budding index analysis, applies an automated image analysis method. The goal is to detect cells that are focused relatively well, and to completely ignore cells that are poor in focus. This method determines the total number of cells and the size of the bud of each cell. There are 3 classifications of the cells: cells without buds, cells with small buds, and cells with large buds. These classes correspond to the cell cycles G1, S and G2/M respectively. Based on the image analysis, there should emerge from the estimation distribution a similar alignment in the cell cycle phase of a cell.

2.4. Data Set

The Reduced Yeast Cell Cycle (RYCC) data set used in this research is from [8]. It is a data matrix with 384 rows and 17 columns. Each row represents a gene with 17 dimensions where each dimension corresponds to a point in the time series. It contains 384 genes that are grouped based on the five phases of the cell cycle: G1/M, G1, S, G2, and M. It is shown in [1] that the data set exhibits periodicity and some relational patterns with respect to the cell cycle. The microarray samples, collected at 17 timepoints taken in ten-minute intervals, cover nearly two full cell cycles (170 min).

2.5. Clustering Methods

In this paper, we make use of two relatively new approaches in data clustering: kernel and spectral methods. For data which are not linearly separable, using these methods should allow us to cluster data using classifiers in the form of hypersurfaces. We give an overview of the two clustering algorithms used. For a deeper understanding, readers are referred to [4, 5].

Clustering methods that use spherical or elliptical metric to group data may not work well when clusters are non-convex. Spectral clustering was introduced to address this problem. The main idea in spectral clustering is to construct similarity graphs that represent the local neighborhood relationships between

observations. To start with, we consider an $n \times n$ similarity matrix whose entries s_{ij} is the similarity between observation i and observation j . The data is then represented in an undirected *similarity graph* $G = (V, E)$, where the n vertices represent the observations and two vertices v_i and v_j are connected by an edge if $s_{ij} \geq 0$ or is greater than some predefined threshold. These edges are weighted by the s_{ij} 's. The goal is to partition the graph such that the edges between different groups have low weights, i.e. data points have high dissimilarity, and edges within a group have high weights, i.e., data points have high similarity.

As an alternative, we can also consider a fully connected similarity graph whose edges are weighted by $w_{ij} = s_{ij}$ and construct the *adjacency matrix* $\mathbf{W} = \{w_{ij}\}$. The *degree* g_i of vertex i is given by the sum of the weights of the edges connected to it, that is $g_i = \sum_j w_{ij}$.

If \mathbf{G} is the diagonal matrix whose entries are g_i , the *graph Laplacian* is given by $\mathbf{L} = \mathbf{G} - \mathbf{W}$. Spectral clustering method finds the respective eigenvectors corresponding to the m smallest eigenvalues of \mathbf{L} to form a matrix $\mathbf{Z}_{n \times m}$. Then the rows of $\mathbf{Z}_{n \times m}$ are clustered, in this case using kernel k -means, resulting in a clustering of the original data points.

A *kernel* is a function K such that for all data points x_1, x_2 in an input space X ,

$$K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle,$$

where ϕ is a mapping from X to an inner product space \mathcal{F} called a *feature space*. Hence, kernel methods are a class of algorithms that perform by mapping the input data into a high dimensional feature space. This is done using the so-called “kernel trick” which is primarily based on Mercer's theorem. According to the theorem, any continuous, symmetric, positive semidefinite kernel function $K(x,y)$ can be expressed as a dot product in a high dimensional space. Therefore, any algorithm that employs the inner product operation can be applied with the kernel trick.

Kernel-based algorithms have come a long way since their introduction. Aside from the fact that kernel functions have provided algorithms a bridge between linearity and nonlinearity, their performance have been proven comparable to, if not better than, existing algorithms in various areas where they have been applied. Thus, recent trends involve “kernelizing” algorithms, that is these algorithms have been extended to use kernels by applying the kernel trick. As of present, many kernel methods for clustering have already been developed, most of them are kernel versions of known clustering algorithms, such as k -means, SOM, and neural gas. However, we will only consider using kernel k -means in clustering our data.

Given a set of unlabeled data and the number of groups we wish to partition the data, k -means aims to find the clusters and their centers by moving the cluster centers iteratively until the total variance within a cluster is a minimum. Thus, the algorithm iterates the following steps until convergence:

1. for each center, find the set of data points closer to it than any other center;
2. identify the new center of each cluster by computing for the means of the data points it contains.

Applying the kernel trick in this algorithm results to kernel k -means. For a set of observed data, kernel k -means involves projecting this data set in some feature space \mathcal{F} by means of a nonlinear map ϕ . Then, the algorithm for the usual k -means is employed. This means that clustering is done in the feature space \mathcal{F} . As a result, the data mapped in \mathcal{F} are now separated by hyperplanes, which when mapped back to the original input space form nonlinear partitionings.

3. Methodology

3.1. Data Preprocessing

The RYCC data set from [8] were initially derived from [1], where a set of 416 genes in yeast were identified to be dependent on the cell cycle. Further filtering of the data was done in [8] by eliminating genes that are associated to more than one phase of the cell cycle and genes that have negative gene expression values, thus resulting to a 384×17 (genes \times sample) data matrix. In addition to these, we disregarded the set of genes identified to be outliers in [2].

3.2. Clustering of Samples

We divide the data into their respective groups as indicated in [1] and clustering using the two algorithms discussed in the previous section is performed for each group. Thus, five clusters per group were obtained, each cluster corresponding to a one of the cell cycle phases. However, indexing of the clusters of a group may not be the same for all groups. To address this problem, center means of each resulting cluster of each group are computed and sorted, thus obtaining a cluster index that can be applied to all groups and allowing us to obtain visualizations of the clustering results.

It is important to note that a clustering method such as the k -means is dependent on initialization, which in this case involves finding the 5 initial centers of each cluster using heuristics. Therefore, to approximate the true clustering of each group, several trials are made and the final visualizations are obtained by getting the average of the center means of the most expressed cluster at a specific time interval for each group. These values are then compared to the average of the center means of each cluster over all trials. The expressed group at a certain timepoint is assigned to a cluster if the distance between their center means is the minimum.

4. Analysis and Results

We graph the time domain data for each group and align the cell cycle phase distribution from FACS and budding index analysis. Based from the alignment, the following behaviours were observed. Groups M/G1, G1 and M have expression levels that peak in their corresponding phases identified by the reference cell cycle distribution. Since the S phase is characterized by synthesis of a lot of proteins needed by the cell, we can expect a peak of the gene expression level that doesn't conform to the corresponding cell cycle phase.

The results of the two clustering methods used are shown in Figure 1 and Figure 2, along with the estimated cell cycle phase distribution and the reference distribution based from using FACS and budding index analysis. The estimation of the cell cycle phase distributions are obtained by computing for the correlation of the center means for each pair of adjacent timepoints. We assigned a similar cell cycle phase to two adjacent timepoints if their cluster means are significantly correlated. Since the *alpha factor-based method* used to synchronize the sample population starts with M/G1 as the initial phase, assignment of the cell cycle phases are done in a similar manner. However, different confidence levels may lead to different distributions. To address this problem, we tested a range of confidence levels (70% - 95%), and computed an index that will measure the goodness of the estimates. We used the *Hamming distance* or the edit distance to measure the consistency of the estimates with respect to the reference distribution. As much as possible we want our edit distance to be minimum. A summary of the edit distances obtained for each confidence level is given in Table 1. Figures 1 and 2 are visualizations of the clustering results and best estimates obtained using the two algorithms.

Spectral Clustering						
Confidence Level	70%	75%	80%	85%	90%	95%
Index (Edit Distance)	15	7	14	14	14	14

Kernel k -Means						
Confidence Level	70%	75%	80%	85%	90%	95%
Index (Edit Distance)	9	4	3	8	5	11

Table 1: Edit distances obtained for each confidence level



Fig 1: Spectral clustering result with an estimate cell cycle phase distribution and the reference distribution based from FACS and budding index analysis

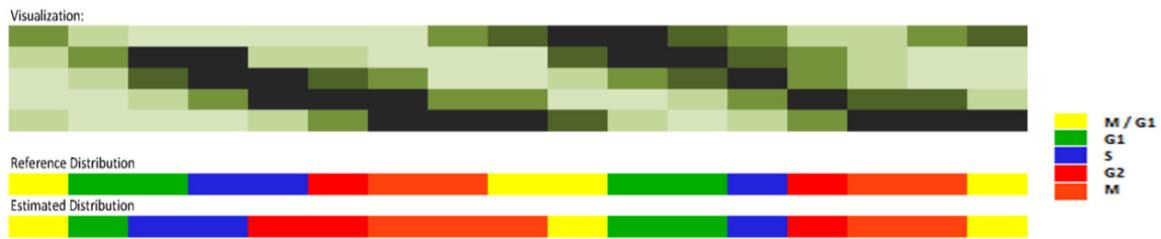


Fig 2: Kernel k-means clustering result with an estimate cell cycle phase distribution and the reference distribution based from FACS and budding index analysis

5. Conclusion

Given the time series gene expression data of a synchronized population of yeast, we obtained an estimate of the cell cycle phase distribution that approximates the result of FACS and budding index analysis by up to 82.35%. The method employs kernel k -means to cluster the set of samples of each group and a correlation confidence level of 80% is used for the cell cycle phase assignment. Based from the computed edit distances, kernel k -means has a better approximation in all levels of confidence compared to the spectral clustering algorithm.

6. Recommendations

For further studies, we recommend using our method in a different time series gene expression data with smaller time interval. We also wish to extend the study to asynchronous population and prokaryotic data.

7. Acknowledgements

The authors would like to thank Ms. Jasmine Malinao for her insights and advise, Erlo Robert Oquendo for the statistical background, Dr. Anactleto Argayosa and Angelo Dela Tonga for preliminary biological interpretations. Ms. Clemente would like to thank ERDT for funding her Master's degree in UP Diliman. Ms. Salido would like to thank Aklan State University for funding her studies in UP Diliman.

8. References

- [1] Cho, R.J., et al. (1998). A Genome-wide Analysis of the Mitotic Cell Cycle. *Molecular Cell* **2** 65-73.
- [2] Clemente, J. and Salido, J.A. (2010). Non-Metric Multidimensional Scaling and Vector Fusion Visualization of Time Series Gene Expression Data for Gene Function Analysis. *Proc. of the Phil. National Conference on Information Technology Educators*.
- [3] Domany, E. (2003). Cluster Analysis of Gene Expression Data. *Journal of Statistical Physics* **110** Nos. 3-6 1117-1139.
- [4] Filippone, M., et al. (2008). A survey of kernel and spectral methods for clustering. *Pattern Recognition* **41** 176-190.
- [5] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag Series in Statistics, New York.
- [6] Niemisto, A., et al. (2007). Computational Methods for Estimation of Cell Cycle Phase Distributions of Yeast Cells. *EURASIP Journal of Bioinformatics and System Biology*.
- [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.
- [8] Yeung, K.Y. (2001). Cluster Analysis of Gene Expression Data. Dissertation, Department of Computer Science and Engineering, University of Washington.