2012 International Conference on Biological and Life Sciences
IPCBEE vol.40 (2012) © (2012) IACSIT Press, Singapore

Non-metric Multidimensional Scaling for Biological Characterization of Reduced Yeast Cell Cycle

Julie Ann Acebuque Salido [+] and Jhoirene Clemente

Algorithm & Complexity Lab Department of Computer Science University of the Philippines Diliman Diliman 1101 Quezon City, Philippines

Abstract. Gene function discovery remains a computational challenge for both biologist and computer ccience. In this paper, we use Non-metric Multidimensional Scaling (nMDS) to visualize the set of time series gene expression data taken from a synchronized population of yeast. From the visualization, we propose a methodology for identifying gene function of uncharacterized genes based from the confidence intervals made by the set of genes identified in a certain biological function. We focus on the set of genes involved in the growth phases (G1 and G2) of a cell. Based from the characterizations, on one of the growth phase, 3 out of 4 (75%) of identified genes are in its accurate biological function.

Keywords: Gene Function discovery, RYCC, nMDS

1. Introduction

Genes are the basic hereditary units 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. However, not all not all genes are associated with their biological functions. Previous works in [1,2] identified yeast genes that are periodically expressed at specific phases of cell cycle. Using standard laboratory techniques, they produced a list of genes with corresponding biological functions associated with the processes identified by a specific cell cycle phase.

Synthesis of proteins is dependent on the expression of genes in an organism. Gene expression at a specific time point can be measured using microarray technology. Computationally, these expression levels are quantitative values showing how genes are expressed compared to others. Pattern discovery in gene expression data, i.e. clustering and data mining, can lead to identification of biological function of genes [2]. Incorporating these procedures can minimize or ultimately eradicate tedious wet laboratory experiments.

Yeast have been subjected to a number of high throughput investigations such as gene expression analysis [1,5,6,7,8], protein-protein interaction mapping [2] and synthetic genetic interaction analysis [2]. In literature, [10] used nMDS to analyze the periodicity of gene expression in human fibroblast serum. The work shows that nMDS visualization captures the temporal pattern of gene expression data.

In this paper, we will present a methodology that suggest possible function of genes using nMDS visualization. We used genes that are identified to be involved in the cell cycle regulation of yeast, specifically for those involved in the growth phases of the cell. We used the identified biological characterization presented in [1,2] as reference to the true biological function of the genes.

2. Definitions and Basic Notations

2.1. Reduced Yeast Cell Cycle (RYCC)

[+] Corresponding author.
E-mail address: salidojulieann2@gmail.com

The Reduced Yeast Cell Cycle (RYCC) [5] dataset used in this research came from a synchronized population of yeast. We can describe it as a normalized data matrix with 384 rows and 17 columns. Each row corresponds to a gene and each column is a specific time point. Each time point have equal ten-minute interval which covers nearly two full cell cycles (170 min). The time series dataset is shown to exhibit periodicity [5,12]. Each gene in the the dataset is characterized depending on the cell cycle identified in [1,2]. We only used the set of genes associated to the growth phases of the cell, that is G1 and G2. In the dataset, not all genes characterized in a specific cell cycle phase are associated with its specific function. Below shows the summary of the total number of genes with known and unknown functions.

Table 1: Number of characterized and uncharacterised genes in growth phases

Cell Cycle Phase	Biosynthesis	DNA Replication	Uncharacterized
G1	17	3	55
G2	-	5	28

2.2. Non-metric Multidimensional Scaling

nMDS is used for the purpose of visualizing a highly dimensional data in 2 or 3 dimensional Euclidean space. Let O be the set of n objects and E is the Euclidean space. The goal of nMDS is to find a mapping from O to E such that the dissimilarity between the objects in O is consistent as much as possible with the distances of the objects in the Euclidean space. The distance between two object in O, say x_i and x_j where $1 \leq i, j \leq n$ is computed to obtain the data set's dissimilarity matrix D, let that be defined in the set O x O. Each object in D is computed using the Euclidean distance.

$$[D]_{ij} = \delta_{ij}^{\,2} = (x_i - x_j)^{\mathrm{T}} (x_i - x_j)$$

From the dissimilarity matrix D, we define an inner product matrix $B = X^T X$, where each element in B is $[B]_{ij} = x_i^{\,\mathrm{T}} x_j$. From the known squared distances in D, we can find the inner product matrix B, and then from B to the unknown coordinates X. Since B is symmetric, positive semi-definite, with rank p therefore B has p non-zero eigenvalues and $(n - p)$ zero eigenvalues. Given the properties of B we can get X from B using its spectral decomposition [10]. An iterative implementation of nMDS minimizes the stress. The minimum stress computed serves as its goodness of fit.

2.3. Confidence Intervals

2.3.1. Confidence Band

A confidence interval with a confidence coefficient $(1 - \alpha)$, $0 \leq \alpha \leq 1$, is a random interval whose end points are called confidence limits. A $100(1 - \alpha)$% confidence interval contains the true value of the parameter estimated, e.g. the $[D]_{ij}$.

The confidence band encloses an area of $100(1-\alpha)$% that contains the true curve. It visualizes the best fit curve, and the confidence band is constructed as the best fit curve is constructed. The confidence band is extended above and below the curve by

Where $c = G \mid x \, \Sigma \, \mathrm{x} \, G' \mid x$, $G \mid x$ is the gradient vector of the parameters at a particular value of x. $G' \mid x$ is the transposed gradient vector, Σ is the variance-covariance matrix. SS is the sum of squares for the fit, DF is the degrees of freedom and $t_\alpha(DF)$ is the value x's t critical value based on the confidence level α and the degrees of freedom DF.

2.3.2. Confidence Ellipse.

Confidence ellipse uses intervals for both X and Y. The interval is projected horizontally and vertically, respectively. Confidence ellipse is formed by, $Z \pm R$ x I, where Z is the mean of either X or Y, R is the range either X or Y, I is the confidence level $(1 - \alpha)$. These form the minor and major axis of the ellipse. The confidence ellipse is given a $100(1 - \alpha)$% confidence to contain the points it bounds.

2.4. Characterizing Classes of Outliers

Potential outliers are points found near or at the periphery of a region occupied by a cluster in the 2-dimensional visualization [11]. The potential outliers are classified into (1) absolute potential outliers; (2) ambiguous potential outliers through the use of confidence bands and confidence ellipses.

Absolute potential outliers are a point lying outside the confidence band and confidence ellipse. This point is no longer bounded by the confidence ellipse and is not represented by fitted curve.

An *ambiguous potential outlier* is a point that is bounded by two different confidence ellipses or two different confidence bands, or a point that is within the confidence ellipse but outside the confidence band. It is unclear as to which cluster should this point be identified with.

A method is also described to validate cluster membership of identified ambiguous potential outliers as defined in section 2.3.

3. Methodology

In this study, the dataset described in Section 2.1 is used. The methodology for identifying possible gene function is as follows.

Step 1. Compute the 2D representation of genes using nMDS.

Step 2. Visualize the output of step 1 using a scatter plot, and color each gene (represented by a point) according to the associated biological classification.

Step 3. Build a confidence ellipse and confidence band with 95% level of confidence per cluster based on the known biological functions identified in [1,2].

Step 4. Record information on the genes which are bounded by more than one confidence ellipse and genes which are not bounded by confidence ellipse and confidence bands. Based on the visualization, scrutinize each cluster. Re-color each gene based on the known biological function of each gene and assign another color for genes with unknown function.

Step 5: Remove all absolute potential outliers and ambiguous potential outliers as described in [11] at section 2.3.

Step 6. Cross validation with 3 known databases, MIPS CYGD (Munich Information Center for Protein Sequences Comprehensive Yeast Genome Database) [4], KEGG (Kyoto Encyclopedia of Genes and Genomes) [9] and BLAST (Basic Local Alignment Search Tool) [3], a validation based on the results generated on the study. The identified genes are cross validated using the database of MIPS CYGD and KEGG are tabulated. The FASTA file format used as input for sequence alignment programs derived from KEGG genes is used for the protein-protein BLAST search and tree view search of the set of genes identified.

Step 7. Cross validation.

4. Results and Analysis

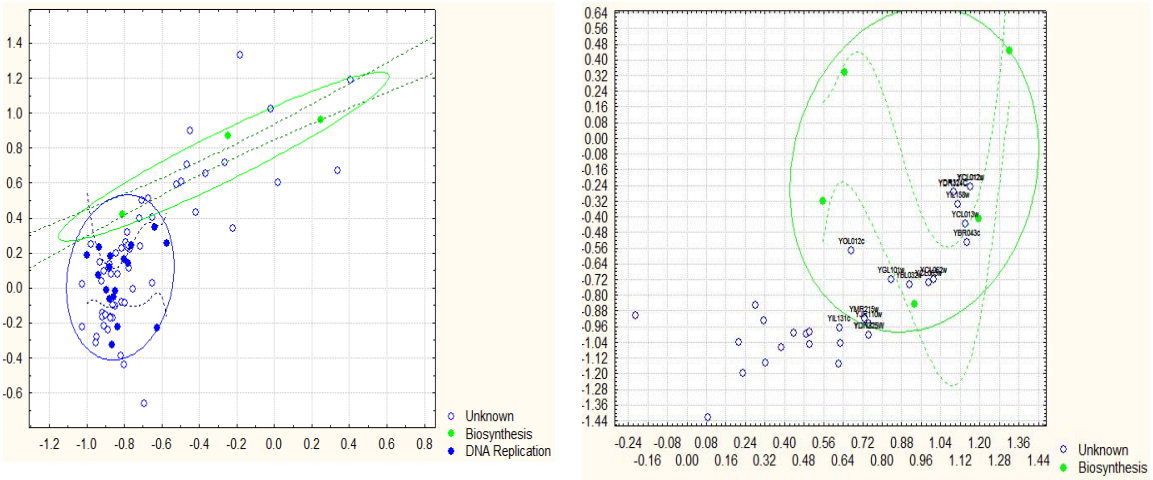
The visualization of the growth phases G1 and G2 of the biosynthesis and DNA replication discussed in Section 2 as shown in figure 1.a and 1.b. Table 2 show the fitted curves per biological functions, in DNA replication of G2 there is no fitted curve since there are not enough set of genes to construct the curves. 95% confidence ellipses are constructed based on the points of different groups of genes with 95% confidence bands for every curve.

Table 2: Curves fitted per biological function

		Goodness of Fit	
	Biological function	**G1**	**G2**
1	DNA Replication	-9.9604-42.472*x-58.3034*x^2-26.0324*x^3	
2	Biosynthesis	0.9515+0.1921*x-0.5718*x^2	-16.4637+62.3588*x-75.4942*x^2+28.7899*x^3

Using this set of graphs for analysis of gene expression data as a visualization technique that summarizes this set of graphs such that properties of each gene will still be visible. The summary of the identified genes as seen in Table 3 of both growth phases as characterized by Section 2.3. Identified genes are those genes

that are within the confidence ellipse and confidence band. Ambiguous genes are identified as genes that lie outside the confidence band but within the confidence ellipse. Absolute potential outlier genes are those genes outside of the confidence band and confidence ellipse.



(a) (b)

a) DNA Replications and Biosynthesis Genes in G1, (b) Biosynthesis with in G2.

Fig. 1: nMDS visualization of

Table 3: Summary of Classifications of Genes on G1 and G2.

Phase	Biological Function	Absolute Outlier(Gene Name)	Ambiguous Genes	Identified Genes
G2	Biosynthesis	YDR451c, YCR085w, YMR003w, YNR009w, YOR073w, YDR325W, YLL047w, YCR086w, YIL169C, YKL053W, YKL069W, YPL264C	YCL012w, YDR324C, YIL158w,YCL013w,YOL012c, YGL101w, YMR215w, YJR110w	YBL032w, YBR043c, YCL062w, YCL063w
G1	Biosynthesis	YPR120c, YPL267w, YDL156w, YOL017w, YLL022c, YLR236c, YCL022c, YCL024w, YLR121c, YHR154w, YHR110w, YDL010w, YPR174c, YJL181w, YLR183c, YOL007c, YIL026c, YJR043c, YOR144cYLR326w, YKL108w, YLR381w, YNL273w, YJL018w, YBR089w, YLR349w, YCL061c, YKR013w, YKL161c, YDL018c, YKR083c, YDR013w, YGR238c, YHR113w, YDL124w, YHR039c, YNL309w, YPL208w, YLR376c, YNL300w, YAR003W, YCL060c, YDL103c, YNL303w, YNL310c,	YDL119c, YDR493w, YDL105w,YJL078c, YDR309c	YKR077w, YOL017w, YBR071w, YPL014w, YKR083c, YDR383c, YGL028c
	DNA Replication	YKR077w, YOL017w, YBR071w, YPL014w, YKR083c, YGL028c, YDL119c, YDR493w, YJL078c, YDR309c, YNL300w, YJR043c, YLR349w, YKR083c, YKL161c, YHR039c, YDL124w, YHR113w	YDR383c, YDL105w, YLR361w, YPR120c, YDL156w, YCL022c, YCL024w, YLR121c, YHR110w, YPR174c, YOL007c, YIL026c, YOR144c, YLR326w, YKL108w, YNL273w, YJL018w, YKR013w, YGR238c, YPL208w, YLR376c, YAR003W, YNL303w,	YPL267w, YLL022c, YLR236c, YHR154w, YDL010w, YLR183c, YCL061c, YDL018c, YDR013w, YCL060c, YDL103c, YNL310c, YNL309w, YJL181w, YBR089w

As seen in Table 4, there are 4 identified genes in biosynthesis of both growth phases, in G1 75% of identified genes cross validated with the 3 known database in yeast genes are involved in biogenesis which is involve in metabolism which is also refers to biosynthesis. Identified genes in biosynthesis and DNA replication, of the unknown genes identified in the growth phase have unknown classification in tree view of BLASTP, but are all a leaf of *ascomycetes*.

Table 4: G1 Identified Genes in Biosynthesis with KEGG, MIPS and BLASTP

G1	Gene Biosynthesis	KEGG	MIPS	BLASTP
1	YKR077w	Msa2p	Protein of unknown function localised to cytoplasm and nucleus	Msa2p, hypothetical protein
	YOL017w	Esc8p	Protein involved in telomeric and mating-type locus silencing , CELL CYCLE AND DNA PROCESSING, DNA processing, DNA restriction or modification, DNA conformation modification	ESC8; Esc8p
2	YBR071w	hypothetical protein	Protein of unknown function localised to cytoplasm	hypothetical protein
3	YPL014w	hypothetical protein	Protein of unknown function localised to cytoplasm and nucleus	hypothetical protein
	YKR083c	Dad2p	Outer kinetochore protein - part of Dam1 complex, chromosome segregation/division, CELL TYPE DIFFERENTIATION, CELL FATE, BIOGENESIS OF CELLULAR COMPONENTS, CELL CYCLE AND DNA PROCESSING,	DAD2, HSK1; Dad2p; K11567 DASH complex subunit DAD2
4	YDR383c	Nkp1p	Non-essential Kinetochore Protein	Nkp1p, hypothetical protein
5	YGL028c	Scw11p (EC:3.2.1.-)	Cell wall protein, may play a role in conjugation during mating, C-compound and carbohydrate metabolism, cytokinesis (cell division) /septum formation and hydrolysis	SCW11; Scw11p (EC:3.2.1.-)

Table 5: G2 Identified Genes in Biosynthesis with KEGG, MIPS and BLASTP

G2	Gene Biosynthesis	KEGG	MIPS	BLASTP
1	YBL032w	Hek2p, HEK2, KHD1; Hek2p	BIOGENESIS OF CELLULAR COMPONENTS, Heterogeneous nuclear RNP K-like gene, CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES	unnamed protein product [Saccharomyces cerevisiae]
2	YBR043c	Qdr3p	Multidrug transporter, function as a quinidine, barban, cisplatin, and bleomycin resistance determinant , CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES, CELL RESCUE, DEFENSE AND VIRULENCE	QDR3, AQR2; Qdr3p,
3	YCL062w	Vac17p	Vacuole-specific receptor of Myo2p, BIOGENESIS OF CELLULAR COMPONENTS, REGULATION OF METABOLISM AND PROTEIN FUNCTION	sce:YCL063W VAC17, YCL062W; Vac17p (A)
4	YCL063w	Vac17p	Vacuole-specific receptor of Myo2p , BIOGENESIS OF CELLULAR COMPONENTS, REGULATION OF METABOLISM AND PROTEIN FUNCTION	VAC17, YCL062W; Vac17p

5. Recommendations

We would like to recommend further analysis on the identified genes for specific biological functions as seen in the summary of identified genes. And perform specific wet laboratories on the suggested identified genes with respect to its biological functions.

6. Acknowledgements

The authors would like to thank Mr. Reynand Jay Canoy for his suggestions in microarray technology. Ms. Salido would like to thank Aklan State University and CHED FDP2-SEGS for funding her education in UP Diliman. Ms. Clemente would like to thank ERDT for funding her Masters degree in UP Diliman.

7. References

[1] Cho, R., Campbell, M. et. al.. A Genome-wide Transcriptional Analysis of the Mitotic Cell Cycle. Molecular Cell, Vol. 2, 1998, 65-73.

[2] Spellman, P., Sherlock, G. et. al.. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization , Molecular Biology of the Cell, Vol. 9 3273-3297, 1998.

[3] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman , "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", (1997), Nucleic Acids Res. 25:3389-3402.

[4] Güldener U, Münsterkötter M, Kastenmüller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martinez J, Perez-Ortin JE, Michael H, Kaps A, Talla E, Dujon B, Andre B, Souciet JL, De Montigny J, Bon E, Gaillardin C, Mewes HW CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Research Jan 1;33 Database issue:D364-8 (2005).

[5] Yeung, K. Y., "Cluster Analysis of Gene Expression Data", Department of Computer Science and Engineering, Ph.D. Dissertation: Computer Science Department at University of Washington, 2001.

[6] Alamgir, Md., Erukova, V. , Jessulat, M., Azizi, A., and Golshani, A., "Chemical-genetic profile analysis of five inhibitory compounds in yeast", BMC Chemical Biology 2010, 10:6, p15.

[7] Domany, E., "Cluster Analysis of Gene Expression Data", Journal of Statistical Physics, 2003, Vol 110, Nos 3-6.

[8] Niemisto, A., Matti Nykter et. al, "Computational Methods for Estimation of Cell Cycle Phase Distributions of Yeast Cells", EURASIP Journal of Bioinformatics and System Biology, Volume 2007.

[9] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M.; KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res. 40, D109-D114 (2012).

[10] Taguchi, Y. H. and Oono, Y. "Relational patterns of gene-expression via non-metric multidimensional scaling analysis", 2005.

[11] E.R. Oquendo, J. Clemente, J. Malinao, and H. Adorna; Characterizing Classes of Potential Outliers through Traffic Data Set Data Signature 2D nMDS Projection, In Philippine Information Technology Journal, Volume4, Number 1, 2011.

[12] Clemente, J. and Salido, J.A.. Non-Metric Multidimensional Scaling and Vector Fusion Visualization of Time Series Gene Expression Data for Gene Function Analysis. PSITE, (2010).