Identification of Candidate Gene Function Group of Yeast Cell Cycle Genes using Gene Expression Data

Julie Ann Salido, *Member, IAENG*

***Abstract*—In various gene expression repositories such as for *Saccharomyces Cerevisiae*, it is a usual occurrence that cell cycle genes are not assigned with biological functions. This is due to current researches focus on gene function discovery using wet laboratory which is time consuming and tedious. This research will focus on the identification of genes' biological functions of yeast, *Saccharomyces Cerevisiae* involved in cell cycle using time series gene expression data. A method for identifying gene functions that uses Nonmetric Multidimensional Scaling with confidence intervals of 95% and confidence ellipse of 95% is proposed of the computing method for identifying the goodness of fit per group. The results are cross validated as a comparison with the three known databases of *Saccharomyces Cerevisiae*, Comprehensive Yeast Genome Database of Munich Information Center for Protein Sequences, Kyoto Encyclopaedia of Genes and Genomes and protein Basic Local Alignment Search Tool of National Center for Biotechnology Information. Using sensitivity analysis of result of identified candidate biological function group compared with the three databases. This method shows a good identification of genes biological functions based on the main characterization of biological phase, using sensitivity analysis associated with confidence interval. The method were able to give candidate biological function groups to 97.77%, 175 of the 179 unclassified genes.**

***Index Terms*—Gene expression, Biological function, nMDS, Confidence Interval, Saccharomyces cerevisiae.**

I. Introduction

CEll cycle is associated with numerous biological changes, making it an attractive model for the genome wide regulation of gene activity. 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 the genome wide analysis of gene expression levels for gene function discovery and biomedical applications. However, this huge amount of data has no meaning without doing significant data mining and other exploratory techniques. Identification of gene functions is carried out by doing specific laboratory techniques which are often very tedious.

Studies have been made identifying sets of genes that are periodically expressed at specific phases of cell cycle in yeast and the cell cycle phase at each time point[4], [20]. The group of Cho[4] identified the cell cycle phase based on the size of the buds, the cellular position of the nucleus and standardization to more than 20 transcripts whose mRNA fluctuations are used as reference.

Yeast genome have been subjected to a number of high throughput investigations such as gene expression analysis[1], [6], [13], [25], computational methods for estimating cell cycle distribution[17], functional analysis [18] and identification of cell cycle regulated genes by microarray hybridization[20] and identification of cell cycle phases using gene expression by[3] among others.

Gene based clustering is the most common technique used in clustering gene expression data. The most common techniques used are: K-means, self-organizing map (SOM), hierarchical clustering, graph-theoretical approach, model based clustering, and density based hierarchical approach[6], [25].

Genes are the basic hereditary unit of living organisms and 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 deoxyribonucleic acid (DNA)[24] microarrays.

The 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 are taken. A sample may have different gene expressions through time, and this sample leads to the analysis of time series gene expression data.

The amount of gene expressed dictates how much proteins are synthesized and therefore responsible for the biochemical interactions taking place inside the cell and gene expression[24] analysis results are highly dependent on basic information about samples and not all available time series gene expression data include these information. This research endeavoured to develop a method for identifying the candidate biological functions of unclassified yeast *Saccharomyces cerevisiae* genes.

II. Basic Definition and Notations

A. Gene Function

In this research gene function refers to the genes biological function as characterized by Cho[4] in his study. The genes biological functions are: cell cycle regulation (CCR), directional growth (DG), DNA replication (DNAR), mating pathway (MP), glycolysis replication(GR), biosynthesis (BIO), chromosome segregation (CS), repair and recombination (RR), transcriptional factors (TF) and miscellaneous (MIS).

Manuscript received April 14, 2015; revised December 4, 2015. This work was supported in part by the Commission on Higher Education (CHED), in collaboration with the Office of the Vice Chancellor for Research and Development of UP Diliman, for funding support through the CHED and Engineering Graduate Scholarship Program and Aklan State University.

J.A. Salido is with the Aklan State University College of Industrial Technology Kalibo, Aklan, Philippines e-mail: salidojulieann2@gmail.com

(Advance online publication: 29 February 2016)

B. Data Set

This study will focus on the 384 x 17 normalized data set of *Saccharomyces cerevisiae* from the data set of Yeung[25] as the reduced yeast cell cycle(RYCC).

The group of Cho in 1998[4] were able to characterized genes to thier cell cycle phases. The characterized genes are further classified according to their biological functions as discussed in Section II-A and summarized in Table I. There are 179 unclassified gene from the 384 normalized data set[25].

TABLE I: Summary of unclassified and biologically characterized genes in each biological functions for all cell cycle phases.

Biological Function	No. of Genes Early G1	No. of Genes Late G1	No. of Genes S	No. of Genes G2	No. of Genes M	Total
Characterized/Classified						
Cell Cycle Regulation (CCR)	3	7	0	1	4	14
Directional Growth (DG)	1	8	1	6	3	19
DNA Replication (DNAR)	6	17	4	0	1	28
Mating Pathway (MP)	3	5	0	0	2	10
Glycolysis Respiration (GR)	9	1	1	0	2	13
Biosynthesis (BIO)	3	3	8	5	2	21
Chromosome Segregation (CS)	0	10	12	3	6	31
Repair and Recombination (RR)	0	13	1	1	1	16
Transcription Factors (TF)	0	3	8	1	5	17
Miscellaneous (MIS)	5	14	5	7	4	35
Total Characterized/Classified	30	81	40	24	30	205
Total Unclassified	37	54	35	28	25	179

C. Non-metric Multidimensional Scaling

The motivation of using the dimensionality reduction technique called Non-metric Multidimensional Scaling (nMDS) is from article of Yeung[25] and Taguchi[23]. Taguchi in his study[23] uses nMDS to analyse the cell cycle periodicity of the human fibroblast serum and showed the effectiveness of this technique in visualizing the temporal patterns of gene expression level. nMDS is used for the purpose of visualizing a highly dimensional data in a two dimensional or three dimensional space. Projecting the data into a lower dimension makes it easy for domain experts in analysing thier data.

1) Algorithm: Let O be the set of n objects and E be 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 are 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 such that $1 \leq i, j \leq n$ 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.

$$[D]_{ij} = \delta_{ij}^2$$

$$\delta_{ij}^2 = (x_i - x_j)^T (x_i - x_j)$$

From the dissimilarity matrix D, define an inner product matrix $B = X^T X$, where each element in B is

$$[B]_{ij} = x_i^T x_j$$

From the known squared distances in D, find the inner product matrix B, and then from B to the unclassified 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[5]. An iterative implementation of nMDS minimizes the stress, the minimum stress computed serves as its goodness of fit.

D. Confidence Intervals

1) Confidence Band: A confidence interval with a confidence coefficient (1 - α), 0 $\leq$ α $\leq$ 1, is random interval whose endpoints are statistics called confidence limits. A 100 (1 - α) $\%$ confidence interval is given a 100 (1 - α) $\%$ confidence to contain the true value of the parameter estimated.

The confidence intervals may be extended to curve estimation where the confidence limit for every value of α on the curve is plotted along with the estimated curve. A confidence band encloses an area that one can be 100 (1 -α) % certain contains the true curve. It gives a visual sense of how well the data define the best-fit curve. The best-fit curve is constructed with the confidence band is extended above and below the curve by

$$\sqrt{c}\sqrt{\frac{SS}{DF}} t_{\alpha}(DF)$$

where c = G $|x$ x σ x $G'|x$, $G|x$ is the gradient vector of the parameters at a particular value of x, $G'|x$ 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_{α} (DF) is the value x's t critical value based on the confidence level and the degrees of freedom DF.

2) Confidence Ellipse: Confidence ellipse is another plot related to the confidence band. It uses intervals for both X and Y. The interval is projected horizontally and vertically respectively. The confidence ellipse is formed by the following equation

$$\bar{Z} \pm R \text{ x } I$$

where $\bar{Z}$ is the mean of either X or Y, R is the range of either X or Y, I is the confidence level 1-α. These form the minor and major axes of the ellipse. The ellipse is given a 100(1-α)% confidence to contain the data points it bounds.

E. Characterizing Classes of Outliers

The paper[19], described potential outliers as points found near or at the periphery of a region occupied by a cluster in the 2-dimensional visualization. The potential outliers are classified into (1) absolute potential outliers; (2) valid potential outliers; and (3) ambiguous potential outliers through the use of confidence bands and confidence ellipses.

1) Absolute potential outliers.
   An absolute potential outlier is 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.
2) Valid potential outliers.
   A valid potential outlier is a point lying outside the confidence ellipse but is still within the confidence band. This point is no longer bounded by the confidence ellipse but is still represented by fitted curve.
3) Ambiguous potential outliers.
   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.

F. Sensitivity of Implementations

The sensitivity is used in statistics to describe diagnostic tests, and include medical tests, medical signs or symptoms [27]. The terms used with the description of sensitivity are true positive (TP) and false negative (FN). TP is considered if a disease is proven present in a patient, the given diagnostic test also indicates the presence of the disease. The true positive and true negative are also considered as standard of truth. FN is considered if the diagnosis test suggest the disease is absent for a patient with disease.

$$\begin{aligned} Sensitivity &= \frac{TP}{(TP+FN)} \\ &= \frac{Number\ of\ true\ positive\ assessment}{Number\ of\ all\ positive\ assessment} \end{aligned}$$

The numerical values of the sensitivity represents the probability of a diagnostic test identifies positive test. The higher the numerical value of the sensitivity, the less likely it returns a false-positive results. The numerical value of accuracy represents the proportion of the true positive results in a given population.

III. Computing Framework

In this study, the set of classified genes with known biological functions is used for the analysis especially in validating and assessing the quality of our visualizations for the unclassified genes. The RYCC will be used as a data set.

A. Identification of gene biological function groups

The computing framework for the identification of candidate gene function group of yeast biological functions will have these steps and begin by identification of groups cell cycle phases, followed by computing framework for identification of genes candidate group biological function:

1) Compute for the nMDS to have a 384 x 2 data matrix for its appropriateness on projecting the characteristics of the normalized gene expression as discussed in Section II-C1.
2) Visualize the result of 384 x 2 data matrix using a scatter plot graph for each cell cycle phases per biological funtion GE_{pb}, where GE is a set of genes, p = (1, 2, ..., 5) is a set of phases and b = (1, 2, ..., 9) is a set of biological functions, with genes based on Table I, periodic genes classified as discussed in Section II-A and II-B. Set the input as active input for graph analysis, graph using 2D scatterplot SP_{pb}, where SP_{pb} = (SP_{11}, SP_{12} ,...,SP_{59}) set the variable of for 2 as nMDS y, graph type as regular, no computation for regression band and fit. Each point in SP_{12} represents a gene.
3) Build a confidence ellipse, with 95% level of confidence as discussed in Section II-D per biological functions E_{p_b}, b = 1, 2, 3,...,10, identified in Table I, and confidence bands as discussed in Section II-D1 of 95% level of confidence for both linear and polynomial to compute for the goodness of fit.
4) For the fitted curve C_f, compute for the linear C_l, quadratic C_q, cubic C_c, quartic C_{q4} and quintic C_{q5}.
5) Compute for the root mean squared (RMSD) of all fitted curve C, and compare for the least RMSD.
6) Identify the best goodness of fit based on least RMSD. Construct the best fit curve, and the confidence band above and below the curve.
7) Identify the genes GE that are at its true function GE_{tf}, genes that are within one confidence ellipse and band.
8) Identify the genes that are potential outliers and classify according to absolute potential outliers GE_{ab}, valid potential outliers GE_v and ambiguous potential outliers GE_{am} as described Section II-E.
9) Compare the results of the identified genes function groups with 3 known databases as a cross validation. The 3 databases are CYGD of MIPS [11], KEGG [14] and BLAST of NCBI [22], a validation based on the results generated on the study.
   In CYGD MIPS, the gene name is used to search the database with its functional classification.
   In KEGG, the gene name is used to search the database, where the primary gene name shown first can be used as an alternative identifier (in place of the accession number) to retrieve the entry. And the hierarchical classification of gene functions according to the KEGG Ontology (KO) system, in which the third level corresponds to each KEGG pathway map or BRITE functional hierarchy. The BRITE hierarchy link will display additional hierarchies, especially for protein families.
   In PBLAST, Standard protein-protein BLAST (blastp) is used for both identifying a query amino acid sequence and for finding similar sequences in protein databases. Like other BLAST programs, blastp is designed to find local regions of similarity. When sequence similarity spans the whole sequence, blastp will also report a global alignment, which is the preferred result for protein identification purposes. The $AA\ seq$, the number of amino acids and the sequence data is used from [14]. The AA seq link generates the sequence data in the FASTA format. The DB search link is used for sequence similarity search by BLAST or FASTA against various databases, and copied for search in [11].
   a) Set the entry sequence ($FASTAsequence$). Insert sequence data ($FASTAsequence$) from KEGG[14].
   b) Set the search to database reference proteins to ($refseq\_protein$).
   c) Based the program selection algorithm to protein-protein BLAST ($blastp$).
   d) Algorithm parameters, is set to max target sequences, the scoring parameters matrix to $BLOSUM62$.
   e) Identify using the descriptions on sequences producing significant alignments in the query of 100% and maximum identity of 100%.
10) Measure the sensitivity of the methods used by comparing the results of the MIPS search with the can-

didate biological functions identified and discussed in Section II-F.

IV. Results and Discussion

This research classified the set of previously identified unclassified genes as shown in Table I with respect to their functional classification from [4] and from the MIPS [11], KEGG[14] and BLAST database[22], then relate the extracted functions of genes to its classification based on the criteria set in the methodology. The number of data set per biological function are identified for analysis with respect to the confidence ellipse and confidence bands. Biological functions with number of identified genes lower than 3 are excluded from analysis since at least 3 points are needed to compute for the confidence ellipse. And biological functions with number of classified genes lower than 4 are excluded from analysis since at least 4 points are needed to compute for the goodness of fit.

The visualization of the normalized data set using nMDS per biological function per cell cycle phase is shown in Section IV-A. IV-B, IV-C, IV-D and IV-E. Table II shows the number of classified true functions identified on all phases of cell cycle. The 23.46% of the previously unclassified genes as of year 2006[26] were given a candidate true functions for classification. And 133 are considered as potential outliers, 74.30% are considered ambiguous genes, since this set of genes were identified in multiple biological functions according to the criteria set. And there are 97.77% of previously unclassified genes in[25], were given a candidate biological functions based on the methods discussed in III-A as shown in Table III as of May 2014 with respect to the 3 databases. Assessed the given method using the defined criteria in sensitivity based on the paper of Zhu[27] as discussed in Section II-F.

TABLE II: The number of candidate true functions genes identified on each cell cycle.

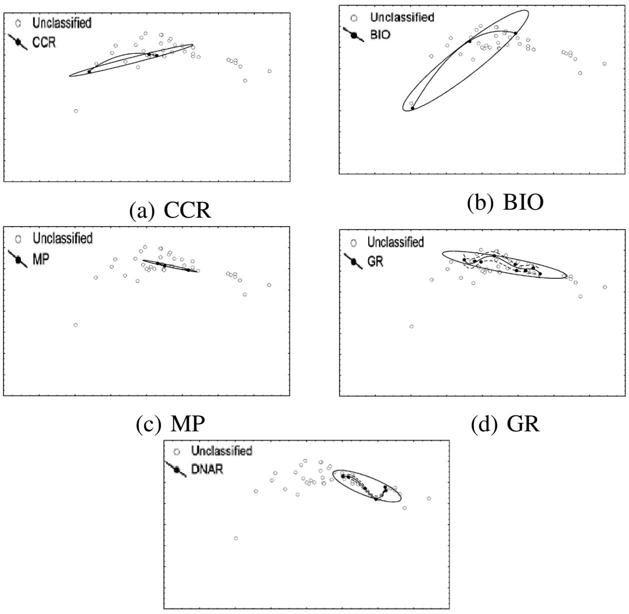
Phases	No. of Identified True Functions	No. of Unclassified
Early G1	16	37
G1	2	54
S	7	35
G2	12	28
M	5	25
Total	42	179

TABLE III: The summary of the number of candidate identified biological functions per cell cycle phases.

Phases	No. of one Candidate	No. of Ambiguous & Valid Outliers	Total No. of Genes with Candidate
Early G1	16	18	34
G1	2	51	53
S	7	28	35
G2	12	16	28
M	5	20	25
Total	42	133	175

A. Candidate Biological Functions for Early G1

The visualization of early G1 of the classified gene, that have enough number of genes to generate a confidence ellipse and regression bands, with respect to the unclassified genes of the RYCC data set[26] as discussed in Section II-B and Table I. The nMDS visualization of RYCC data set with 95% confidence ellipse and 95% goodness of fit of classified genes with respect to the unclassified genes in early G1 are shown in Figure 1. Table IV shows the candidate biological function of genes based on the classification as discussed in Section III-A for the true candidate functional classification.



(a) CCR (b) BIO (c) MP (d) GR (e) DNAR

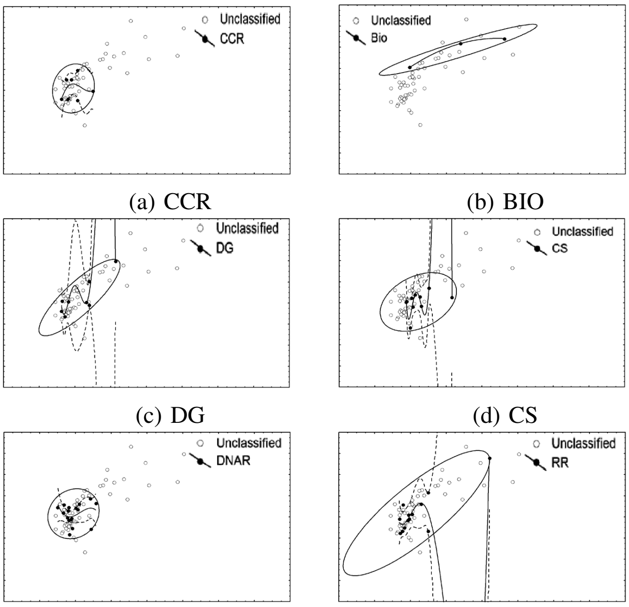
Fig. 1: nMDS visualization of RYCC data set with 95% confidence ellipse and 95% goodness of fit of classified genes with unclassified genes in early G1, (a) Cell cycle regulation (b) Biosynthesis (c) Mating pathway (d) Glycolysis respiration and (e) DNA replication.

TABLE IV: The identified candidate true biological functions of unclassified genes during the early G1 phase of cell cycle, using 95% confidence ellipse and 95% confidence band.

No.	Gene Name	Candidate Gene True Classification
1	YML109w	BIO
2	YBR054w	BIO
3	YPR002w	GR
4	YBR158w	BIO
5	YDL117w	BIO
6	YPL066w	BIO
7	YBR052c	BIO
8	YHR022c	BIO
9	YBR053c	BIO
10	YKL163w	GR
11	YDR511w	MIS
12	YLR254c	MIS
13	YBR231c	MIS
14	YDR368w	MIS
15	YLR050c	MIS
16	YLR051w	MIS

B. Candidate Biological Functions for the First Growth Phase (G1)

The visualization of G1 of the classified genes, that have enough number of genes to generate a confidence ellipse



(a) CCR (b) BIO

(c) DG (d) CS

(e) DNAR (f) RR

Fig. 2: nMDS visualization of RYCC data set with 95% confidence ellipse and 95% goodness of fit of known genes with respect to the unclassified genes in G1. (a) CCR (b) BIO (c) DG (d) CS (e) DNAR (f) RR

and regression bands, with respect to the unclassified genes of the RYCC data set[26] as discussed in Section II-B and Table I. The nMDS visualization of RYCC data set with 95% confidence ellipse and 95% goodness of fit of classified genes with respect to the unclassified genes in G1 are shown in Figure 2. Table V shows the candidate biological function of genes based on the classification as discussed in Section III-A for the true candidate functional classification.

TABLE V: The identified candidate biological functions of unclassified during the G1 phase of cell cycle, using 95% confidence ellipse and 95% confidence band.

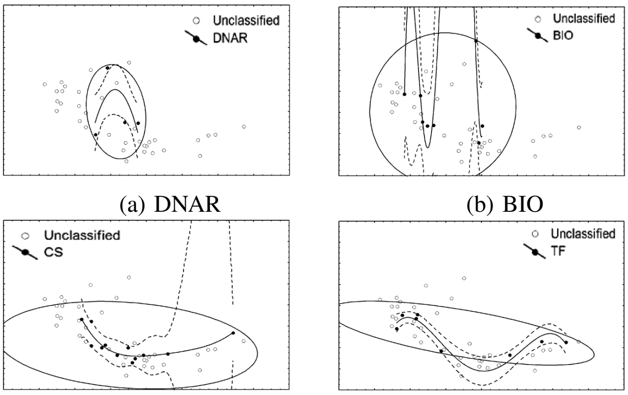
No.	Gene	Candidate Gene Classification
1	YJR043c	MP
2	YDR493w	BIO

C. Candidate Biological Functions for Synthesis (S)

The visualization of S of the classified genes, that have enough number of genes to generate a confidence ellipse and regression bands, with respect to the unclassified genes of the RYCC data set[26] as discussed in Section II-B and Table I. The nMDS visualization of RYCC data set with 95% confidence ellipse and 95% goodness of fit of classified genes with respect to the unclassified genes in S are shown in Figure 3. Table VI shows the candidate biological function of genes based on the classification as discussed in Section III-A for the true candidate functional classification.

D. Candidate Biological Functions for the Second Growth Phase (G2)

The visualization of G2 of the classified genes, that have enough number of genes to generate a confidence ellipse and regression bands, with respect to the unclassified genes of the RYCC data set[26] as discussed in section II-B and Table I. The nMDS visualization of RYCC data set with 95% confidence ellipse and 95% goodness of fit of classified genes with respect to the unclassified genes in G2 are shown in Figure 4.



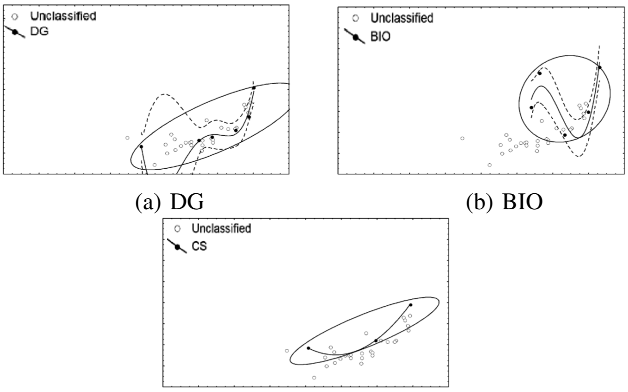
(a) DNAR (b) BIO

(c) CS (d) TF

Fig. 3: nMDS visualization of RYCC data set with 95% confidence ellipse and 95% goodness of fit of known genes with respect to the unclassified genes in S phase (a) DNAR (b) BIO (c) CS (d) TF .

TABLE VI: The identified candidate biological functions of unclassified during the synthesis phase of cell cycle, S using 95% confidence ellipse and 95% confidence band.

No.	Gene	Candidate Gene Classification
1	YOL019w	DNAR
2	YDR252w	CS
3	YMR048w	BIO
4	YGR189C	BIO
5	YFR026C	BIO
6	YKL066W	BIO
7	YNL072W	BIO



(a) DG (b) BIO

(c) CS

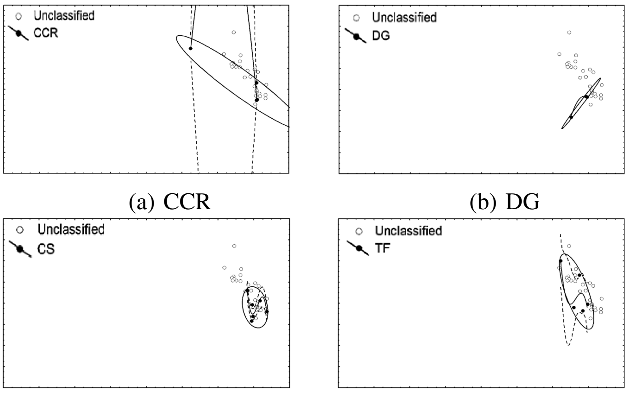
Fig. 4: nMDS visualization of RYCC data set with 95% confidence ellipse and 95% goodness of fit of known genes with respect to the unclassified genes in G2 phase (a) DG (b) BIO (c) CS.

E. Candidate Biological Functions for Mitosis (M)

The visualization of M of the classified genes, that have enough number of genes to generate a confidence ellipse and regression bands, with respect to the unclassified genes

TABLE VII: The identified candidate biological functions of unclassified during the G2 phase of cell cycle, using 95% confidence ellipse and 95% confidence band.

No.	Gene	Candidate Gene Classification
1	YIL131c	DG
2	YDR451c	DG
3	YCR085w	DG
4	YMR003w	DG
5	YOR073w	DG
6	YLL047w	DG
7	YDR366c	DG
8	YCR086w	DG
9	YDR325W	DG
10	YKL069W	DG
11	YPL264C	DG
12	YKL053W	MIS



(a) CCR

(b) DG

(c) CS

(d) TF

Fig. 5: nMDS visualization of RYCC data set with 95% confidence ellipse and 95% godness of fit of known genes with respect to the unclassified genes in M phase (a) CCR (b) DG (c) CS (d) TF.

of the RYCC data set[26] as discussed in section II-B and Table I. The nMDS visualization of RYCC data set with 95% confidence ellipse and 95% goodness of fit of classified genes with respect to the unclassified genes in M are shown in Figure 5. Table VII shows the candidate biological function of genes in G2 based on the classification as discussed in Section III-A for the true candidate functional classification. Table VIII shows the candidate biological function of genes in M based on the classification as discussed in Section III-A for the true candidate functional classification.

TABLE VIII: The identified candidate biological functions of unclassified during the M phase of cell cycle, using 95% confidence ellipse and 95% confidence band.

No.	Gene	Candidate Gene Classification
1	YGL201c	CCR
2	YOL137w	CCR
3	YPL186c	CCR
4	YOL014w	MIS
5	YGR230w	MIS

F. Sensitivity of the computing framework

The true positives were identified and compared from the unclassified genes of [4] in 1998 and [25] in 2001, with the database search in MIPS[11] as of May 2014. The true positives and false positives of the result in the candidate biological functions identified by the methods used are shown in summary in Table IX. The rating for each phases are summarized. The numerical values of the sensitivity represents the probability of a diagnostic test identifies positive test. The higher the numerical value of the sensitivity, the less likely it returns a false-positive results.

TABLE IX: The summay of true positive and false negative identified on cell cycle phases.

Phases	No. of True Positive	No. of False Positive	Total No. of Subjects	Rate of Sensitivity
Early G1	15	6	21	0.7143
G1	26	8	34	0.7647
S	10	12	22	0.4545
G2	9	6	15	0.6000
M	2	0	2	1.0000
Total	62	32	94	0.7067

V. Conclusion

This research was able to develop a method to identify candidate genes biological functions group using regression and visualization for unclassified genes. Cross validation was achieved through comparison to verify the correctness of the identified candidate biological functions with the three known databases in yeast; with CYGD of MIPS[11], KEGG[14] and BLASTP of NCBI[22].

From the methods and tools used, this research was able to achieve the following:

1) The identified genes through the confidence interval ellipse and confidence bands, exhibit similarity of biological functions based on the defined biological functions per phase as enumerated in Section II-B and tabulated in Section IV.
2) Identified in unclassified genes are proteins that has not yet been isolated and its amino acid sequence is predicted from the DNA sequence available and suggested biological functions start characterizing these genes.
3) The methods are able to give candidate biological functions to 97.77%, 175 of the 179 unclassified genes in (Cho,1998).
4) Based on the sensitivity rating of 0.7067 of the methods used, this can be increased since most genes can exhibit one or more biological functions.

VI. Recommendations

With the results of this research and through domain expert validation, we recommend that:

- Further analysis of domain experts on the set of outlier genes detected to the set of genes with proteins of unknown functions from [4] and MIPS database.
- A set of wet laboratory be done on possible identifications of biological functions for genes for Tables IV, V, VI, VII and VIII.
- Consider visualizing another gene expression data in time series using nMDS visualization.

Acknowledgment

The author acknowledges the Commission on Higher Education (CHED), in collaboration with the Office of the Vice Chancellor for Research and Development of UP Diliman, for funding support through the CHED and Engineering Graduate Scholarship (SEGS) Program. The author also wishes to express gratitude to Aklan State University, Dr. Danilo E. Abayon, particularly the College of Industrial Technology, Dr. Ersyl T. Biray for the administrative support. My sincere gratitude to Dr. Jaime D.L. Caro, Ms. Jasmine A. Malinao and Dr. Henry N. Adorna for the continuous support of my researches, for all your patience, motivation, enthusiasm, and immense knowledge. Ms. Salido would like to thank Erlo, Jhoirene, Rich , Francis and Jeffrey for the outpouring support and Algorithm and Complexity Lab for the use of equipments.

REFERENCES

[1] Califano, A., Stolovitzky, G. and Tu, Y.(2000). Analysis of Gene Expression Microarrays for Phenotype Classification. IBM Computational Biology Center. NY 10598.

[2] Califano, A.(1999). SPLASH: Structural Pattern Localization Algorithm by Sequential Histograming. Bioinformatics. IBM press.

[3] Clemente. J. and Salido, J.A.(2011). Non Metric Multidimensional Scaling and Vector Fusion Visualization of Cell Cycle Independent Gene Expressions for Gene Function Analysis. Philippine Information Technology Journal. Vol. 4. No. 1. February 2011. pp. 25-31.

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

[5] Cox, T.T. and Cox, M.A. (2001) .Multidimensional Scaling, Chapman & Hall/CRC. 2nd Ed.

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

[7] Dumbgen, L. and Johns, R.B. (2004). Confidence bands for Isotonic Median Curves using sign test. Journal of Computational and graphical statistics. Vol. 13. No. 2. pp. 519-533.

[8] Fitzgibbon, A., Pilu, M. and Fisher, R.B. (1999). Direct Least Square Fitting of Ellipses. IEEE Transaction in pattern analysis and machine intelligence. Vol. 21. No. 5. pp. 476 - 480.

[9] Gene. (2011). In Merriam-Webster.com. Retrieved October 20, 2011, from http://www.merriam-webster.com/dictionary/gene.

[10] Golub, T. R. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression Monitoring. Science. 286(5439): 531-537.

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

[12] ICMG Ltd. (n.d). Cell Cycle Research. *What is Cell Cycle?*. Retrieved May 15, 2012. from http://www.cellcycles.org.

[13] Jiang, D., Tang, C., and Zhang, A. (2004). Cluster Analysis for Gene Expression Data: A Survey. IEEE Transactions on Knowledge and Data Engineering. Vol 16. No. 11. pages 1370-1386.

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

[15] Malinao, J.A. and Juayong, R.A.B. and Becerral, J.G. and Cabreros, K.R.C. and Remaneses, K.M.B. and Khaw, J.G. and Wuysang, D.F. and Corpuz, F.J.O. and Hernandez, N.H.S. and Yap, J.M.C. and Adorna, H.N. (2010). *Patterns and Outlier Analysis of Traffic Flow Using Data Signatures via IDIRBrG Method and Vector Fusion Visualization*, 2010 3rd International Conference on Human-Centric Computing (HumanCom). 1-6, 10.1109/HUMANCOM.2010.5563344.

[16] National Center for Biotechnology Information. (n.d.). Genetics Home Reference, *Glossary*. October 2011, from http://ghr.nlm.nih.gov/glossary.

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

[18] Oliver, Stephen G., Winson, Michael K., Kell, Douglas B. and Baganz, F. (1998). Systematic Functional Analysis of the Yeast Genome. Trends Biotechno. p373-378.

[19] Oquendo, E.R., Clemente,J., Malinao,J. and Adorna, H. (2011). Characterizing Classes of Potential Outliers through Traffic Data Set Data Signature 2D nMDS Projection. Philippine Information Technology Journal. Volume 4. Number 1.

[20] 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.

[21] Spellman, P.T., Zhang, M.Q., Brown, P.O., Botstein, D. & Futcher, B.(1998). Yeast Cell Cycle Analysis Project, Retrieved from http://genome-www.stanford.edu/cellcycle/data/rawdata/.

[22] Altschul, S.F. , Madden, T.L., Schffer, A.A. , Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J.. (1997). *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res. 25:3389-3402.

[23] Taguchi, Y. H. and Oono, Y.. (2005). Relational patterns of gene-expression via non-metric multidimensional scaling analysis. Bioinformatics. Vol. 21. No 6. pages 730-740.

[24] US Library of Medicine. National Institute of Health. Bethesda. MD 20894. *Genetics Home Reference*. July 2011. Retrieved from http://ghr.nlm.nih.gov/.

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

[26] Yeung, K.Y. & Ruzzo, W.L. (2006). Principal Component Analysis for clustering gene expression data. Retrieved from http://faculty.washington.edu/kayee/pca/.

[27] Zhu, W., Zeng, N. and Wang N. (2010). Sensitivity, Specificity, Accuracy, Associated Confidense Interval and ROC Analysis with Practical SAS Implementations. Health Care and Life Science. NESUG 2010.

Julie Ann Acebuque Salido (M14) This author became a Member (M) of IAENG in 2014, Born in Mandurriao, Iloilo City, Philippines on September 15, 1977. Master of Science in Computer Science, University of the Philippines Diliman, Department of Computer Studies, Algorithm and Complexity Laboratory, Philippines, 2015, bioinformatics, information technology, applied computer science.

She is CHAIR, MONITORING AND EVALUATION, Aklan State University from August 2014 up to present, August 2008 May 31, 2010; ICT COORDINATOR, ASSISTANT PROFESSOR in Aklan State University, June 2008 up to present. She is a recipient of the Science and Engineering Government Scholarship Program of Commission and Higher Education, June 2010- May 2012, in University of the Philippines Diliman, Quezon City Philippines. Published researches: Vision-Based Size Classifier for Carabao Mango Using Parametric Method, International Research Conference in Higher Education (IRCHE), Manila, Philippines, October 3-4, 2013. Non-metric Multidimensional Scaling for Biological Characterization of Reduced Yeast Cell Cycle, Published in the International Proceedings of Chemical, Biological & Environmental Engineering, IPCBEE vol.40 (2012), Singapore. Estimating Cell Cycle Phase Distribution of Yeast from Time Series Gene Expression Data, Published in the International Proceedings of Computer science and Information Technology, IPCSIT vol.6 (2011), Singapore, Presented in the 2011 International Conference on Information and Electronics Engineering, May 28-29, 2011, Bangkok Thailand, Published in Engineering & Technology Digital Library.

Prof. Salido is a member of International Association of Computer Science and Information Technology (IACSIT), SCIEI and Philippine Society of Information Technology Educators WV. 1st Runner-up in the 24th WESVARRDEC Research Symposium in Iloilo City, Philippines. Best Paper and Presenter for both Research proposal and Completed Research category in R & D In-House Review 2014 in Aklan, Philippines. Best Paper for Research Proposal category and Best Presenter for Research Proposal, Presented in the R & D In-House Review, October 23, 2013, Weather Analysis through Data Mining.