Outlier Identification using Nonmetric Multidimensional Scaling of Yeast Cell Cycle Phase using Gene Expression Data

Julie Ann Salido, *Member, IAENG*

***Abstract*—Current researches focused on gene function classification and discovery are with the use of wet laboratory. This research focused on the identification of outlier yeast genes, *Saccharomyces cerevisiae* involved in a eukaryotic cell cycle using time series normalized gene expression data. A method for identifying outlier genes using Nonmetric Multidimensional Scaling (nMDS) with confidence intervals of 95% and confidence ellipse of 95% is used for the computing method for identifying the goodness of fit per group. This method shows a good identification of outlier genes based on the identified genes per cell cycle phases, using criteria identified for visualization associated with confidence interval. Visualization of the data set captures the group structure of genes based from the cell cycle. It shows the characteristics of the events of the genes and identified outliers are included at the adjacent groups. Based on this study, 25 outlier genes were identified, 6.51% of the gene set population.**

***Index Terms*—Gene expression, outlier, Nonmetric Multidimensional Scaling, Saccharomyces cerevisiae.**

I. 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 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. Analysis without biological significance is futile. Identification of gene functions is carried out by doing specific laboratory techniques which are often very tedious. Cell cycle is associated with numerous biological changes, making it an attractive model for the genome wide regulation of gene activity.

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 [1], [2]. The group of Cho[1] 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[3], [4], [5], [12], computational methods for estimating cell cycle distribution [6], functional analysis [7] and identification of cell cycle regulated genes by microarray hybridization [2] among others.

This work was supported in part by the Commission on Higher Education (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, 5600 e-mail: salidojulieann2@gmail.com

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) microarrays.

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 analysis results are highly dependent on basic informations about samples and not all available time series gene expression data include these information. Identification of biological functions of genes that can lead to gene and pattern discovery that will guide to new biomedical applications. Identification of biological functions in silico will minimize tedious wet laboratory experiments. Gene function discovery can lead to development of treatments and drugs for diseases and identify appropriate medical treatment to specific types of diseases. Gene expression analysis, leads to drug development, drug response, and therapy development [10]. This research endeavored to develop a method for identifying outlier genes of yeast cell cycle phases, of *Saccharomyces cerevisiae* [11] genes using scientific visualization of gene expression data.

A. Basic Definition and Notations

1) Cell cycle: The cell cycle [8] refers to the events that take place in a cell between its inception and subsequent replication as shown in Figure 1. The cell cycle is composed of 4 distinct phases: G1 phase, S phase, G2 phase, and M phase. Interphase is a collection of G1, S, and G2 phases.

The phases of the eukaryotic cell cycle:

a) M phase. The Mitosis (M) phase is relatively brief and consist of nuclear division, or mitosis, followed by division of the cytoplasm. Biosynthetic activities are largely halted during M phase. The 2 resulting daughter cells enter interphase once the M phase is complete.

b) G1 phase. The Growth 1 (G1) phase is the first phase in interphase. The G indicates "gap" or "growth". Cells in G1 are $2n$ and biosynthetic activities resume that were suspended during the M phase.

c) S phase. DNA Synthesis (S) phase begins when DNA synthesis starts. At the end of S phase, all chromosomes have been replicated, providing a pair of sister chromatids.

d) G2 phase. Growth 2 (G2) lasts until the cell enters mitosis. Normal activities of the cell including metabolism, growth, and differentiation occur during G2.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

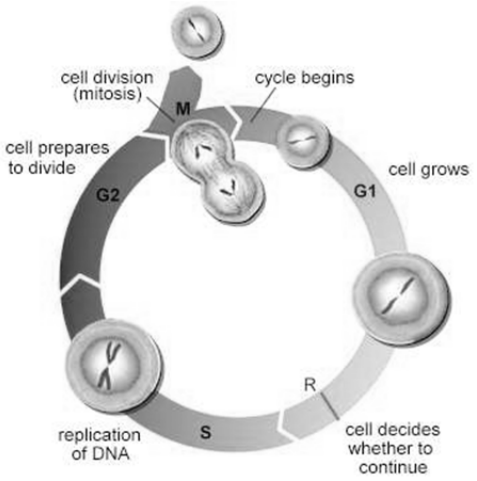


Fig. 1: A eukaryotic cell cycle.

TABLE I: A summary of periodic, biologically classified and unclassified genes by Cho[1] and Yeung[3].

Phases	No. of Periodic Genes[3]	No. of Classified Genes[1]	No. of Unclassified Genes
Early G1	67	30	37
Late G1/G1	135	81	54
S	75	40	35
G2	52	24	28
M	55	30	25
Total	384	205	179

2) Data Set: Cell cycle dependent periodic genes were found in 416 of the 6218 monitored transcripts by [1]. 384 genes are Reduced Yeast Cell Cycle (RYCC) data set of [3]. Of the 384 genes identified by [3] to the different cell cycle phases, 205 were characterized or classified by [1] as summarized in Table I. The summarized genes were divided in the 4 cell cycle phases and early G1(resting phase). The total characterized genes are 205 out of 384, there are 11 genes characterized by Cho et. al. of which has peak in more than one phase of cell cycle and not included in the data set of Yeung[3] and 5 genes characterized by Cho and not on data set of Yeung[3].

The 384×17 (genes$\times$sample) data set, and each sample were taken in a 10 minutes interval starting at time 0 until 160 minutes after, the full and detailed data set are available in [11]. The table is a summarized number of genes from the data set of [3] and [11].

II. Methodology

In this study, the set of classified genes with their cell cycle phases were used, for the analysis specially in validating and assessing the quality of our visualizations for the outlier detection. This study focused on the 384 x 17 normalized data set of *Saccharomyces cerevisiae* from [3].

This steps in computing framework for the identification of outliers genes are:

1) Compute for the Non-Metric Multidimensional Scaling (nMDS) of the given data set.

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 objects 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 [13].

2) Visualize the result of 384 x 2 data matrix using a scatterplot graph for each cell cycle phase G_p, where G is a set of genes and p = (1, 2, ..., 5), with genes based on Table I, periodic genes classified by [3]. Graph using 2D scatterplot S_p, where S_p = (S_1, S_2,...,S_5).

3) Build a confidence ellipse E_p, where E_p = (E_1, E_2,...,E_5) with 95% confidence interval[16]. 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

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

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 bound. Set ellipse to 95% confidence coefficient. Scale the graph to the equal maximum extent and step increment for the $nMDS\ x$ and $nMDS\ y$ coordinates.

4) Identify all genes which are and not bounded by confidence ellipse based on the criteria per phase. And filter the set of genes per phase G'_p, without the outliers identified.

Potential outliers are points found near or at the periphery of a region occupied by a cluster in the 2-dimensional visualization [14], [15]. The potential outliers are classified into a) absolute potential outliers; b) valid potential outliers; and c) ambiguous potential outliers through the use of confidence ellipses.

a) Absolute potential outliers. An absolute potential outlier is a point lying outside the confidence curve and confidence ellipse. This point is no longer bounded by the confidence ellipse and is not represented by fitted curve.

b) Valid potential outliers. A valid potential outlier is a point lying outside the confidence ellipse but is still within the confidence curve. This point is no longer bounded by the confidence ellipse but is still represented by fitted curve.

c) Ambiguous potential outliers. An ambiguous potential outlier is a point that is bounded by two different confidence ellipses or two different confidence curve, or a point that

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

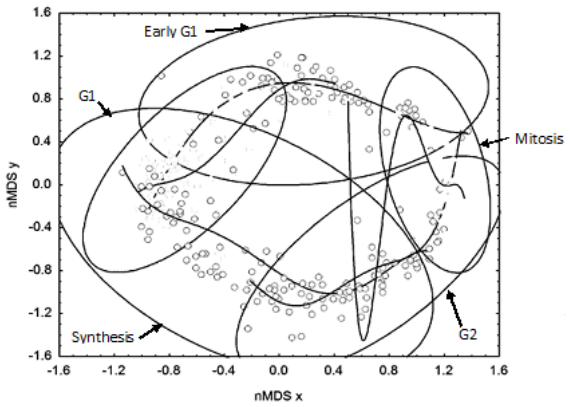


Fig. 2: nMDS scatterplot S_p of cell cycle phases of RYCC data set with 95% confidence ellipse per phase E_p and confidence curve per phase C_f.

is within the confidence ellipse but outside the confidence band. It is unclear as to which cluster should this point be identified with.

5) Visualize the result of G'_p based on Table II, using a scatterplot for each cell cycle phase without identified outliers S'_p, where S'_p = (S'_1, S'_2,...,S'_5, graph using 2D scatterplot.

6) Build a confidence ellipse E'_p, where E'_p = (E'_1, E'_2,...,E'_5) with 95% level of confidence per phase. By setting the normal ellipse to 95% confidence coefficient, scale the graph to the equal maximum extent and step increment for the $nMDS\ x$ and $nMDS\ y$ coordinates.

7) Identify the genes that are potential outliers and classify accroding to absolute potential outliers G_{ab}, valid potential outliers G_v and ambiguous potential outliers G_{am} [14].

III. Major Findings

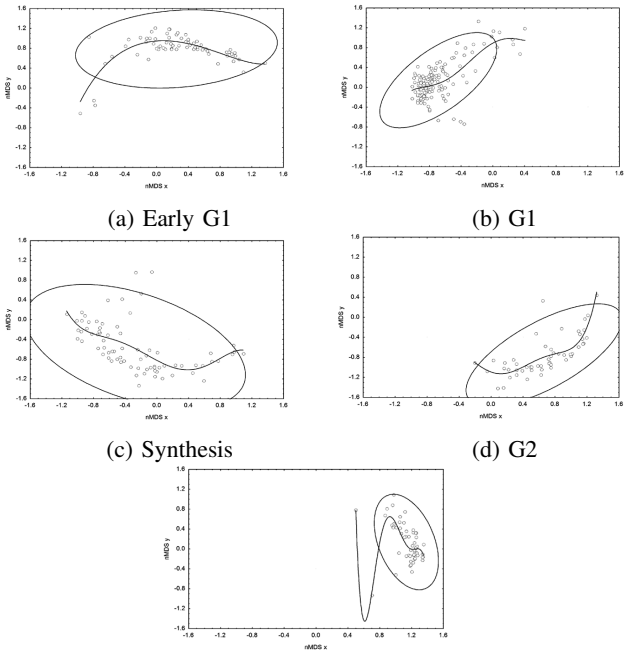
A. Closeness of Co-members

The visualization of the computed nMDS of RYCC as seen in Figure 2 shows the confidence curve and ellipses E_p, of 384 periodic genes as enumerated by [3] in his website [11]. As seen in Figure 2, nMDS visualization showed a significant clustering of genes with respect to its 5 groups. Genes belonging to a group are projected closer to one. The reason for this behavior lies on the fact that nMDS compares each gene to one another by virtue of its dissimilarity matrix. It is clearly seen that genes belonging to a group exhibits a common expression level through time and nMDS visualization captures that property.

The visualization of all phases on the set of parameters as defined in the methods are shown in Figure 3. In the computed curve of all phases, on the set of genes in all phases the confidence curves are polynomial and quintic. The ellipses per phase shows a temporal pattern based on the order of cell cycle of a budding yeast.

B. Relationship accross phases

The nMDS visualization captures the sequence of the phases in the cell cycle, Early G1 is followed by G1, G1 is followed by S, S is followed by G2 and G2 is next to M.

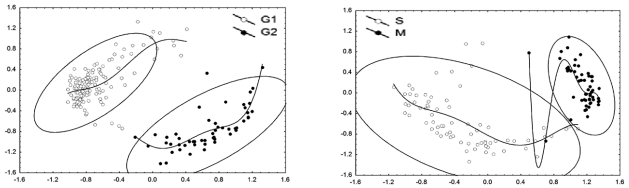


(a) Early G1 (b) G1

(c) Synthesis (d) G2

(e) Mitosis

Fig. 3: nMDS visualization of RYCC data set with 95% confidence intervals of (a) Early G1 (b) G1 (c) Synthesis (d) G2 (e) Mitosis, with the computed $Confidence\ Ellipse\ and\ Band$.



(a) G1 and G2 (b) Synthesis and Mitosis

Fig. 4: nMDS scatter plot of (a) both growth phases of RYCC data set with 95% confidence intervals per phases and goodness of fit and (b) synthesis and mitosis phases of RYCC data set with 95% confidence intervals per phases and goodness of fit.

The transition was captured in the x and y axes of the scatter plot. As seen in Figure 3, the projection of each group using nMDS, the position of each group in the euclidean space with respect to some gene function, genes involved in cell growth is located in the left hand side of the x axis and in the lower part of y axis. As observed, the 2 dimensional projection of genes capture its functional property not just its temporal behavior. And the visualization of both growth phases as seen in Figure 4 with outliers and even without outliers in Figure 6 are linearly separable. The visualization in synthesis and mitosis for phases Figure 4 and Figure 6 are also linearly separable in both the confidence ellipses. It is very clearly seen that all genes in synthesis and mitosis are linearly separable without the outliers.

C. Outlier Identification

Outlier as we defined earlier are those genes projected outside the ellipse for each group. The ellipse was based on a confidence interval having a 95% confidence level. This

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE II: Summary of periodic, biologically classified and unclassified genes without outliers.

Phases	No. of Periodic Genes[3]	No. of Outlier Genes	No. of Filtered Genes
Early G1	67	4	63
Late G1/G1	135	12	123
S	75	4	71
G2	52	3	49
M	55	2	53
Total	384	25	359

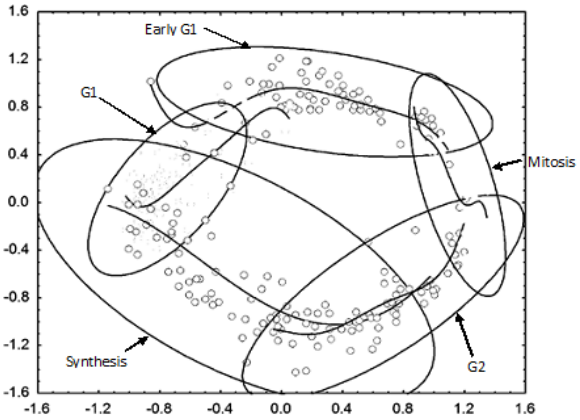
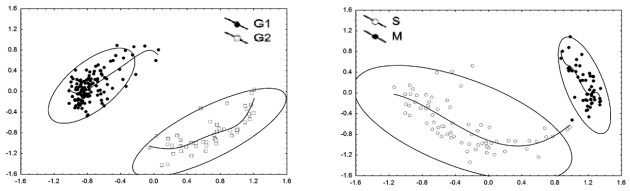


Fig. 5: nMDS scatter plot of all biological phases of RYCC data set with 95% confidence intervals per phases and goodness of fit without outliers E'_p.



(a) G1 and G2 (b) Synthesis and Mitosis

Fig. 6: nMDS scatter plot of (a) both growth phases (G1 and G2) of RYCC data set with 95% confidence intervals per phases and goodness of fit without outliers and (b) synthesis (S) and mitosis (M) phases of RYCC data set with 95% confidence intervals per phases and goodness of fit without outliers.

means that we are 95% confident that the genes included inside the ellipse are members of the group. In Table II is the summary of the number of genes identified as outliers, and fitted an ellipse for each group and identify a set of outliers as shown in Figure 5 and identified as shown in Table IV. Figure 5 shows a better visualization than Figure 2 for the synthesis and the second growth phase. These genes are subjected for further analysis for cross validation. The enumerated outliers per phase are shown in Table III with its cluster membership as outlier. All outliers identified have candidate cluster classification for further analysis by biologist and domain experts.

TABLE III: The table shows the outlier genes identified in Early G1, G1, Synthesis, G2, M in cell cycle using 95% confidence interval.

Phase	Gene Name	Biological Function[1]	Membership
Early G1	YLR015w	Unclassified	5
Early G1	YML109w	Unclassified	2 & 3
Early G1	YGL055w	Miscellaneous	2 & 3
Early G1	YNL016c	Biosynthesis	2 & 3
G1	YJR043c	Unclassified	3
G1	YHR039c	Unclassified	3 & 1
G1	YDL124w	Unclassified	1
G1	YDL119c	Unclassified	1
G1	YDR493w	Unclassified	1
G1	YLL021w	Mating Pathway	3
G1	YPL127c	Transcription factors	3
G1	YDR297w	Miscellaneous	3
G1	YNL173c	Mating Pathway	1
G1	YHR038w	Repair and recombination	1
G1	YOR317w	Biosynthesis	1
G1	YOR316c	Miscellaneous	1
S	YCRX04w	Unclassified	4
S	YNL073w	Biosynthesis	2 & 1
S	YER017c	Miscellaneous	2 & 1
S	YMR198w	Chromosome segregation	4
G2	YKL053w	unclassified	3
G2	YLR014c	Biosynthesis	3 & 1
G2	YOR274w	Biosynthesis	5
M	YAL040c	Cell cycle regulation	1
M	YPR167c	Micellaneous	4 & 3

TABLE IV: The table shows the number of outlier genes identified on all phases of cell cycle, 95% confidence ellipse.

Phases	No. of Potential Outlier	No. of Unclassified
Early G1, phase 1	4	2
G1, phase 2	12	5
S, phase 3	4	1
G2, phase 4	3	1
M, phase 5	2	0
Total	25	9

IV. Conclusion

Based from the defined criteria Section III-C , nMDS visualization is a good tool for gene expression analysis. From the methods and tools used in this study, the following were achieved:

1) The visualization captures the group structure of genes based from the biologically defined groups.

2) It follows the temporal pattern of gene expression based from the events of cell cycle.

3) There is a significant transition in x and y axes of the nMDS space with respect to the cell growth function. Groups of gene involved in that function are on the leftmost and lower part of the graph based on per phase.

4) Genes that are identified outliers are identified to be included in the confidence interval of adjacent groups.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

5) There visualization of 2 growth phases are identified to be linearly separable. Also the synthesis and mitosis are shown to be linearly separable.

6) nMDS compare each gene to one another by virtue of its dissimilarity matrix. It is clearly seen that genes belonging to a group exhibits a common expression level through time and nMDS visualization captures that property.

7) The ellipses per phase shows a temporal pattern based on the order of cell cycle of a budding yeast.

V. Recommendations

All outliers identified have candidate cluster classification for further analysis by biologist and domain experts.Further analysis of domain experts on the set of outlier genes detected, with proteins of unknown functions from [1] and MIPS database. Consider also visualizing another gene expression data in time series using nMDS visualization.

Acknowledgment

The author would like to thank the Commission on Higher Education, for funding support through the Commission on Higher Education Science and Engineering Graduate Scholarship (CHED SEGS) Program. I also wishes to express gratitude to Aklan State University, headed Dr. Danilo E. Abayon, particularly the College of Industrial Technology, headed by Dr. Ersyl T. Biray.

References

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

[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] 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.

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

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

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

[7] Oliver, Stephen G., Winson, Michael K., Kell, Douglas B. and Baganz, Frank , "Systematic Functional Analysis of the Yeast Genome", Trends Biotechnol,p373-378, 1998.

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

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

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

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

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

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

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

[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.,*"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), (2010), 1-6, 10.1109/HUMANCOM.2010.5563344.

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

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.