Estimating Biological Function Distribution of Yeast Using Gene Expression Data

Julie Ann A. Salido
Aklan State University, College of Industrial Technology, Kalibo, Aklan, Philippines
Email: salidojulieann2@gmail.com

Stephanie S. Pimentel
Capiz State University-Burias Campus, Burias, Mambusao, Capiz, Philippines
Email: lectin9@yahoo.com

***Abstract*—Microarray technologies that monitor the level of expression of a large number of genes have emerged. And given the technology in deoxyribonucleic acid (DNA) - microarray data for a set of cells characterized by a phenotype an important problem is to identify "patterns" of gene expression that can be used to predict cell phenotype. The potential number of such patterns is exponential in the number of genes. Detection of genes biological function *in silico* which has not yet been discovered through other means aside from wet laboratories is of practical significance. In this research, biological function distribution of budding yeast cell *Saccharomyces cerevisiae*, using 170 classified gene expression data of yeast is used for visualization and analysis, and evaluated with the reference time distribution using FACS and budding index analysis. We define the criteria using edit distances, a good scientific visualization with 83.78% prediction on time series distribution on the first peak and 86.49% on the second peak.**

***Index Terms*—DNA microarrays, gene expression, computational biology, saccharomyces cerevisiae, FACS, budding index analysis**

I. Introduction

In an attempt to understand complicated biological systems, large amounts of gene expression data have been generated by researchers. 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. Detection of biological functions distribution of gene using gene expression data have not been studied much in the field of computational biology. There are two known cell cycle phase distribution in budding yeast cell population, one is nonparametric method, fluorescence-activated cell sorting (FACS) and the other is budding index analysis [1].

Identification of estimated cell cycle phase distribution on yeast using spectral clustering and kernel k-means were presented by [2]. Yeast gene expression has been investigated for gene expression analysis in [3]-[5], computational methods for estimating cell cycle [1], protein-protein interaction mapping and synthetic genetic interaction analysis in [6].

In literature about fifty percent of the yeast genes have unclassified biological functions. In this research the biological phase distribution of budding yeast cell Saccharomyces cerevisiae is studied, and analyzed patterns using gene expression data with the biological functions identified as cell cycle regulation, directional growth, DNA replication, mating pathway, glycolysis replication, biosynthesis, chromosome segregation, repair and 2 recombination and transcriptional factors [3] for pattern analysis, compare and analyzed the relationship across function of patterns with the reference time distribution.

The study focus on classified (identified) budding yeast cell Saccharomyces cerevisiae involved in cell cycle regulation using its gene expression data as discussed in Section A. The biological function used this research is based on known biological functions presented in [3] as follows:

- 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)

A. Reduced Yeast Cell Cycle

The comprehensive catalog of yeast genes whose transcript level varies periodically within the cell cycle was created by the group of Spellman *et al*. [6]. Table I shows the 170 genes that were identified from 211 genes to their 5 cell cycle phases and 9 biological functions, and meet the minimum criterion for cell cycle regulation of budding yeast cell Saccharomyces cerevisiae. The data set used did not include the miscellaneous biological function genes, since this group of genes are known and

Manuscript received May 22, 2014; revised November 16, 2014.

doi: 10.12720/joig.2.2.158-163

classified but exhibit multiple type of biological function. These sets of genes are functionally active at a specific phase of cell cycle. The 5 groupings of genes based from the 5 phases of cell cycle are: (1) Phase 1: the post mitotic phase, G1/M, Early G1; (2) Phase 2 where DNA replication and cell growth take place, G1[1]; (3) Phase 3 DNA and protein synthesis phase, S[2]; (4) Phase 4 cell growth and preparation for mitosis stage, G2[3] and (5) Phase 5 cell growth stops and will now start to divide, M as shown in Table I. Table II shows, 35 of the 205 identified genes were induced in two different cell cycle phases and did not display a predominant peak and the miscellaneous biological function were eliminated for analysis. Table II shows the summary of number of genes per biological functions for all phases, this is the set of data that is used for analysis.

TABLE I. SUMMARY OF THE NUMBER OF RYCC GENES PER PHASE.

Phases	No. of Genes (Data Set)	No. of Classified Genes
Phase 1, EG1	25	30
Phase 2, G1	67	81
Phase 3, S	35	40
Phase 4, G2	17	24
Phase 5, M	26	30
Total	170	205

TABLE II. SUMMARY OF THE NUMBER OF GENES PER BIOLOGICAL FUNCTION FOR ALL PHASES.

Biological Function	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Total
CCR	3	7	0	1	4	15
DG	1	8	1	6	3	19
DNAR	6	17	4	0	1	28
MP	3	5	0	0	2	10
GR	9	1	1	0	2	13
BIO	3	3	8	5	2	21
CS	0	10	12	3	6	31
RR	0	13	1	1	1	16
TF	0	3	8	1	5	17
Total	25	67	35	17	26	170

II. REVIEW OF RELATED LITERATURE

A. Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization

The catalog of yeast genes whose transcript level vary periodically within the cell cycle was created by the group of Spellman *et al*. [6]. They were able to shows the 800 genes that were identified to their 5 biological phases and meet an objective minimum criterion for cell cycle regulation[4] with 57 wet laboratory experiments, and not all of them were characterized and classified based on their biological functions and processes as summarized in Table III. More than 35.125% are unclassified in function and 49.5% are unclassified in processes.

TABLE III. SUMMARY OF IDENTIFIED REGULATED YEAST GENES PER CELL CYCLE PHASE BY SPELLMAN.

Phases	No. of Genes	No. of Unclassified Function	No. of Unclassified Process
M/G1	113	35	61
G1	300	111	150
S	71	19	26
S/G2	121	43	58
G2/M	195	73	101
Total	800	281	396

B. A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle

The study through filtering the initial set from [6], data from [3] and [4] eliminate genes that are associated to more than one phase of the cell cycle and those genes that have negative gene expression values resulting to a 384 × 17 (genes × sample) data set as summarized in Table IV.

The characterized genes are further classified according to their biological functions such as: cell cycle regulation, directional growth, DNA replication, mating pathway, glycolysis replication, biosynthesis, chromosome segregation, repair and recombination, transcriptional factors and miscellaneous functions which represent others functions different from the nine enumerated Section I.

TABLE IV. SUMMARY OF PERIODIC, BIOLOGICALLY CLASSIFIED AND UNCLASSIFIED GENES BY CHO [3] AND YEUNG [4].

Phases	No. of Periodic Genes[4]	No. of Classified Genes[3]	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

C. Computational Methods for Estimation of Cell Cycle Phase Distributions of Yeast Cells

This study uses computational methods for estimating the cell cycle phase distribution of a budding yeast Saccharomyces cerevisiae cell population [1]. Nonparametric method used that is based on the analysis of DNA content in the individual cells of the population, and DNA content is measured with a fluorescence-activated cell sorter (FACS). Budding index analysis uses automated image analysis method is presented for the task of detecting the cells and buds. The study uses quantitative information on the cell cycle phase distribution of a budding yeast S.cerevisiae population. They therefore provide a solid basis for obtaining the complementary information needed in deconvolution of gene expression data. Fig. 1 shows the reference time series distribution of *Saccharomyces cerevisiae*; it shows the 17 time points of the sample genes using FACS and budding index analysis.

[1] First growth phase [6].
[2] Synthesis [6].
[3] Second growth phase [6].
[4] Gene regulation is the process of turning genes on and off and ensures that the appropriate genes are expressed at the proper times[7]

Figure 1. Reference time series distribution of Saccharomyces cerevisiae.

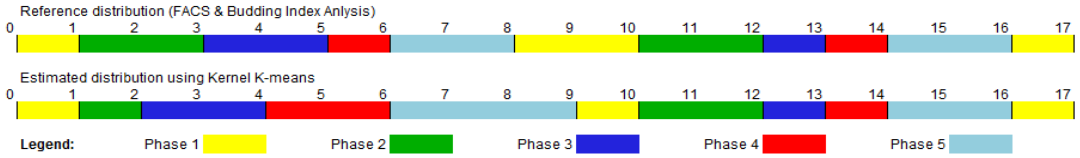

Figure 2. Reference time series distribution and Spectral clustering of Saccharomyces cerevisiae

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

The time domain data for each group were graph and align the cell cycle phase distribution from FACS and budding index analysis. The results were observed per groups M/G1, G1 and M have expression levels that peak in their corresponding phases identified by the reference cell cycle distribution. The kernel k-means and spectral clustering is used to estimate the cell cycle phase distribution and the reference distribution based from using FACS and budding index analysis. The resulting estimate of the edit distances obtained for each cell cycle phase is 82.35% using the kernel k-means algorithms [2]. Fig. 2 shows the reference distribution and the spectral clustering distribution from the study of [2].

III. METHODOLOGY

The methods involve on this study are:

1. Preprocess RYCC, filter genes according to their biological function *(f)* from [3] data set, and used the normalized data set for each phase *(p)*, as seen in Table I.

2. Identify genes per biological functions (G_{f_p}), where f = {1, 2, 3... 9} for each phase p, where p = {1, 2, 3, 4, 5} as seen in Table II.

3. Compute for the mean normalized gene expression of all identified genes per biological function per phases (1) $\overline{M_{f_p}}$,

$$\overline{M_{f_p}} = \frac{\sum_1^n M_{n_p}}{N_{f_p}} \quad (1)$$

where:

M_{f_p} = mean normalized gene expression of identified genes for biological function f of phase p.

M_{n_p} = normalized gene expression of nth biological function of phase p.

N_{f_p} = number of genes in each biological function f of phase p.

4. Compute for the mean normalized gene expression of all identified genes per biological function per phases $\overline{M_{f_p}}$, from time point t_1 to t_{17}.

5. Visualize the output of $\overline{M_{f_p}}$, for each time points t_n, using a line graph for each time domain distribution of G_{f_p} . Align the visualization per biological function, based on the time domain graph per phase V_{f_p} with the reference distribution.

6. Group per cell cycle phase p and align the visualization to the reference time distribution.

7. Identify the peak per time points of the reference time distribution from time point t_1 to t_{17}.

8. Identify the peak per time points of the V_{f_p} from time point t_1 to t_{17}.

9. Visualization analysis for cross-reference of the results using edit distance.

IV. RESULTS AND DISCUSSION

The time domain graph of the RYCC data as seen in all figures shows the gene expression level fluctuation of each gene per biological function. The graphs showed that there are almost 2 peaks of their expression level, possibly because of the 2 cell cycles captured by the 17 time points.

The visualization of the normalized RYCC data set using its mean normalized values of all identified genes per biological function for all phases that includes all identified genes of all phase at every subsection in A, B, C, D and E. The edit distance is used to measure the consistency of the estimates with respect to the reference distribution, comparing the peak of the time domain distribution with the peak of reference. As much as possible we want our edit distance to be a minimum. The total edit distance in the first peak is 6 out of 37 biological functions for all phase, with an 83.78% approximation on the first peak for genes to peak on the computed time points. And 5 out of 37 biological function for all phase with an 86.49% approximation on the second peak for genes to peak on the computed time points. The summary of edit distances are shown in Table V.

TABLE V. SUMMARY OF THE EDIT DISTANCES FOR ALL PHASES.

Phases	Edit Distance		Reference	
	1st Peak	2nd Peak	1st Peak	2nd Peak
Phase 1	0	0	6	6
Phase 2	0	0	9	9
Phase 3	1	1	7	7
Phase 4	3	2	6	6
Phase 5	2	2	9	9
Total	6	5	37	37

A. Time Domain Graph of Early G1

TABLE VI. SUMMARY OF THE TIME-POINT PEAK PER BIOLOGICAL FUNCTION FOR ALL PHASE 1.

Biological Function	Estimated Distribution		Reference	
	1st Peak	2nd Peak	1st Peak	2nd Peak
CCR	10	17	8,9,10	16,17
DG	8	17	8,9,10	16,17
DNAR	10	17	8,9,10	16,17
MP	10	17	8,9,10	16,17
GR	10	17	8,9,10	16,17
BIO	10	17	8,9,10	16,17
Total	0	0	6	6

The visualization of the normalized RYCC data set using its mean normalized values of all identified genes per biological function for phase 1 as shown in Fig. 3 that includes all identified genes. The edit distance to measure the consistency of the estimates with respect to the reference distribution is shown in Table VI. As much as possible we want our edit distance to be a minimum. The edit distances between the reference and the computed time series distribution for phase 1 is 0, which means there is no difference in the reference distribution.

B. Time Domain Graph of Phase 2, G1

The visualization of the normalized RYCC data set using its mean normalized values of all identified genes per biological function for phase 2 as shown in Fig. 4 that includes all identified genes. The edit distance to measure the consistency of the estimates with respect to the reference distribution is shown in Table VII. As much as possible we want our edit distance to be a minimum. The edit distances between the reference and the computed time series distribution for phase 2 is 0, which means there is no difference in the reference.

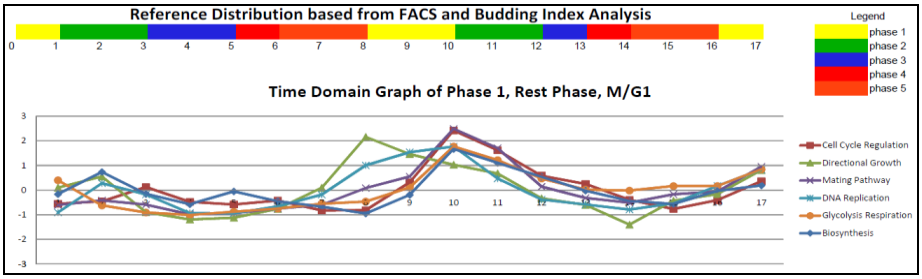

Figure 3. Phase 1 Visualization time series distribution.

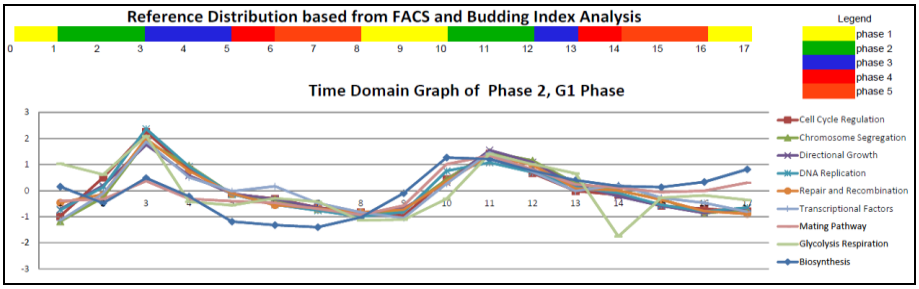

Figure 4. Phase 2 Visualization time series distribution.

TABLE VII. SUMMARY OF THE TIME-POINT PEAK PER BIOLOGICAL FUNCTION FOR PHASE 2.

Biological Function	Estimated Distribution		Reference	
	1st Peak	2nd Peak	1st Peak	2nd Peak
CCR	3	11	1,2,3	10,11,12
DG	3	11	1,2,3	10,11,12
DNAR	3	11	1,2,3	10,11,12
MP	3	11	1,2,3	10,11,12
GR	3	11	1,2,3	10,11,12
BIO	3	10	1,2,3	10,11,12
CS	3	11	1,2,3	10,11,12
RR	3	11	1,2,3	10,11,12
TF	3	11	1,2,3	10,11,12
Total	0	0	9	9

C. Time Domain Graph of Phase 3, S

The visualization of the normalized RYCC data set using its mean normalized values of all identified genes per biological function for phase 2 as shown in Fig. 5 that includes all identified genes. The edit distance to measure the consistency of the estimates with respect to the reference distribution is shown in Table VIII. As much as possible we want our edit distance to be a minimum. The edit distances between the reference and the computed time series distribution for phase 3 is 1 in the first peak and 1 in the second peak.

TABLE VIII. SUMMARY OF THE TIME-POINT PEAK PER BIOLOGICAL FUNCTION FOR PHASE 3.

Biological Function	Estimated Distribution		Reference	
	1st Peak	2nd Peak	1st Peak	2nd Peak
DG	3	12	3,4,5	12,13
DNAR	5	12	3,4,5	12,13
GR	4	11	3,4,5	12,13
BIO	4	12	3,4,5	12,13
CS	5	13	3,4,5	12,13
RR	6	13	3,4,5	12,13
TF	3	13	3,4,5	12,13
Total	1	1	7	7

D. Time Domain Graph of Phase 4

The visualization of the normalized RYCC data set using its mean normalized values of all identified genes

per biological function for phase 4 as shown in Fig. 6 that includes all identified genes. The edit distance to measure the consistency of the estimates with respect to the reference distribution is shown in Table IX. As much as possible we want our edit distance to be a minimum. The edit distances between the reference and the computed time series distribution for phase 4 is 3 in the first peak and 2 in the second peak.

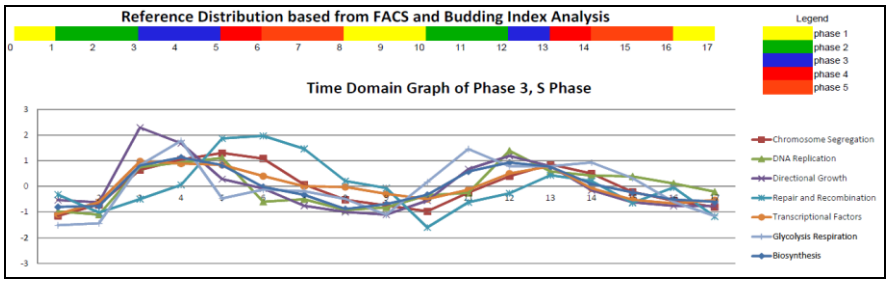

Figure 5. Phase 3 Visualization time series distribution.

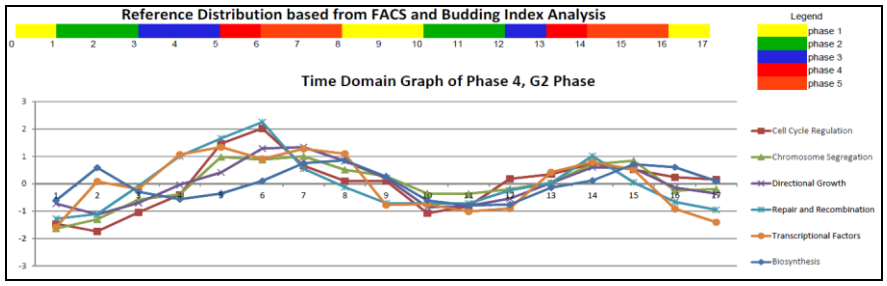

Figure 6. Phase 4 Visualization time series distribution.

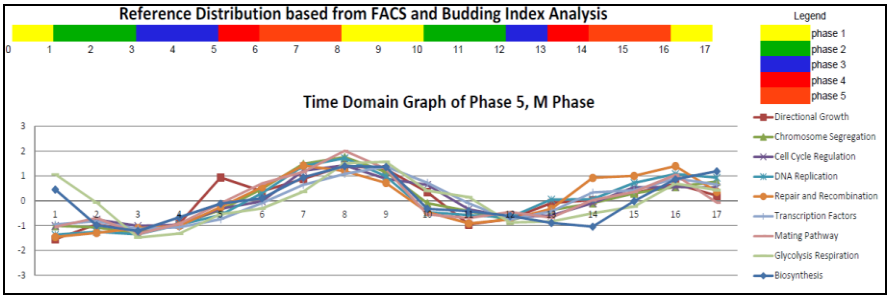

Figure 7. Phase 5 Visualization time series distribution.

TABLE IX. SUMMARY OF THE TIME-POINT PEAK PER BIOLOGICAL FUNCTION FOR ALL PHASE 4.

Biological Function	Estimated Distribution		Reference	
	1st Peak	2nd Peak	1st Peak	2nd Peak
CCR	6	14	5,6	13,14
DG	7	14	5,6	13,14
BIO	8	15	5,6	13,14
CS	7	15	5,6	13,14
RR	6	14	5,6	13,14
TF	5	14	5,6	13,14
Total	3	2	7	7

E. Time Domain Graph of Phase 5

The visualization of the normalized RYCC data set using its mean normalized values of all identified genes per biological function for phase 5 as shown in Fig. 7 that includes all identified genes. The edit distance to measure the consistency of the estimates with respect to the reference distribution is shown in Table X. As much as possible we want our edit distance to be a minimum. The edit distances between the reference and the computed time series distribution for phase 5 is 2 in the first peak and 2 in the second peak.

TABLE X. SUMMARY OF THE TIME-POINT PEAK PER BIOLOGICAL FUNCTION FOR ALL PHASE 5.

Biological Function	Estimated Distribution		Reference	
	1st Peak	2nd Peak	1st Peak	2nd Peak
CCR	8	16	6,7,8	14,15,16
DG	8	16	6,7,8	14,15,16
DNAR	8	16	6,7,8	14,15,16
MP	8	16	6,7,8	14,15,16
GR	9	16	6,7,8	14,15,16
BIO	8	17	6,7,8	14,15,16
CS	8	17	6,7,8	14,15,16
RR	7	16	6,7,8	14,15,16
TF	9	16	6,7,8	14,15,16
Total	2	2	9	9

V. CONCLUSIONS

The synchronized population of classified gene expression data of RYCC, obtained the estimated biological function distribution that approximates the result with reference distribution peaks for each phases, by 85.14%. With the defined criteria using the edit distances, with an 83.78% prediction on time series distribution on the first peak and 86.49% on the second peak, it already gives a good candidate time distribution.

The visualization also captures the characteristics of the data set, which is almost two cell cycles And the difference usually on the time series distribution of the peak that vary from the reference peak is just 1 time points, which can be attributed to the time interval of samples.

VI. Recommendation

For further studies, we recommend using this method in the uncharacterized yeast genes for possible identification of its biological function based on its gene expression distribution through time. We also recommend extending this study to other time series gene expression, asynchronous and prokaryotic data and other eukaryotic data set.

Acknowledgment

Ms. Salido acknowledges the support of the Commission on Higher Education Science and Engineering Graduate Scholarship (CHED-SEGS) Program. She also wishes to express gratitude to Dr. Ka Yee Yeung of University of Washington for granting the use of the data set in Saccharomyces cerevisiae in this study. And Ms. Jasmine A. Malinao and Jhoirene Clemente for the valuable inputs in data mining.

References

[1] A. Niemisto, N. Matti, *et al*., "Computational methods for estimation of cell cycle phase distributions of yeast cells," *EURASIP Journal of Bioinformatics and System Biology*, vol. 2007, pp. 1-9. 2007.

[2] J. A. Salido, J. Clemente, *et al*., "Estimating cell cycle phase distribution of yeast from time series gene expression data," in *Proc. 2011 International Conference on Information and Electronics Engineering IPCSIT*, Singapore, vol. 6, 2011, pp. 105-109.

[3] R. Cho, M. Campbell, *et al*., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65-73, 1998.

[4] K. Y. Yeung, "Cluster analysis of gene expression data," Department of Computer Science and Engineering, Ph.D. Dissertation: Computer Science Department at University of Washington, 2001.

[5] E. Domany, "Cluster analysis of gene expression data," *Journal of Statistical Physics*, vol. 110, no. 3-6, 2003.

[6] P. Spellman, G. Sherlock, *et al*., "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.

[7] National Center for Biotechnology Information. (October 2011). [Online]. Available: http://www.ncbi.nlm.nih.gov

Julie Ann A. Salido. 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, 2014, 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, Instructor in Aklan State University, and 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. Ms Salido is a member of International Association of Computer Science and Information Technology (IACSIT), SCIEI and Philippine Society of Information Technology Educators WV. 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.

Stephanie S. Pimentel. Born in Philippines on March 9, 1971. Ph.D. in Applied Marine Bioscience, Tokyo University of Marine Science and Technology, 2008 M.S. Biology, University of the Philippines, Diliman, 2003, B.S. in Biology, Far Eastern University, 1992.She is currently teaching in, Capiz State University-Burias Campus Burias, Mambusao, Capiz Philippines, June 2014 - present. Instructor (June 2013 to present) College of Fisheries and Marine Sciences, Aklan State University (New Washington campus), Program Coordinator of Marine Biology course Instructor (June 2012 to may 2013) School of Arts,and Sciences, Aklan State University (Banga campus) SAS Research Coordinator, Substitute Instructor (November 2011 to March 2012 College of Education, Arts,and Sciences, Capiz State University (Pontevedra campus), Assistant Professor of Biology (November 2009 to October 2011) Department of Biology, School of Science and Engineering, Ateneo de Manila University, Katipunan Ave., Quezon City, Philippines. Science Lecturer (June 2003-October 2005. Biology Department, Far Eastern University, Morayta St., Sampaloc, Manila, Philippines. Research assistant (November 1993 to March 1999) University Research Associate- Natural Sciences Research Institute, University of the Philippines, Diliman. Published researches: Probiotic Effects of Four *Lactobacillus* Species on *Edwardsiella tarda* Challenge Nile tilapia (*Oreochromis niloticus*). The Book of Abstracts 29th Annual PAASE Meeting and Symposium: "Linking Science and Engineering to Development". July 13-15, 2009. Suplementation of Probiotics on *Edwardsiella tarda* challenge Nile tilapia (*Oreochromis niloticus*). The 5th Asean Conference on lactic Acid Bacteria: Microbes in Diseases Prevention & Treatment. July 1-3, 2009. Differences of Probiotic Effects on *Edwardsiella tarda* Challenged Nile Tilapia (*Oreochromis niloticus*) Fed with Four *Lactobacillus* Species. *Aquaculture Sci* 56(3). Ms Pimentel is a member of National Academy of Science and Technology (NAST- Philippines), Philippine Society for Microbiology Japanese Society of Fish Pathology (JSFP) Science and Technology Advisory Council- Japan Chapter (STAC-J) Association of Filipino Students in Japan (AFSJ) Philippine Association of the Japanese Ministry of Education Scholar (PHILAJAMES) and Malacological Society of the Philippines.