Molecular Marker Discovery for Ovarian Maturation Level of the Black Tiger Shrimp from Microarray Data Using Genetic Algorithm

Supoj Hengpraprohm[1], Suwimol Jungjit[2], Kairung Hengpraprohm[3], and Dech Thammasiri[4]

[1, 3]Program in Data Science, Faculty of Science and Technology,
[4]Program in Business Computer, Faculty of Management Science,
Nakhon Pathom Rajabhat University, Thailand
[2]Department of Computer and Information Technology,
Faculty of Science,
Thaksin University, Thailand
[1]supojn@yahoo.com
[2]suwimol@tsu.ac.th
[3]kairung2011.heng@gmail.com
[4]dechit@msn.com

***Abstract* - This paper proposes an embedded-based gene selection method in the black tiger shrimp microarray dataset using Genetic Algorithm (GA). Moreover, we analyze the effect of selected genes for ovarian maturation level in the term of biological perspective. We run experiment on 6 black tiger shrimp microarray datasets which were publically available. First, we run GA to find the optimized individual length. Next, the GA-based predictive model is run to predict the Gonadosomatic Index (GSI) of ovarian maturation level in black tiger shrimp. Finally, the most frequently selected genes from GA are observed and analyzed. The experimental results show that the optimized individual length for GA is 4 and the selected informative genes are from heart and gill tissues of black tiger shrimp.**

***Keywords* - Gene Selection, Microarray Data, Genetic Algorithms, Black Tiger Shrimp, Ovarian Maturation Level**

I. INTRODUCTION

Black tiger shrimp was the major economic product in Thailand. However, the shrimp industry has experienced many problems such as slow growth, epidemics and incomplete breeding of shrimp brood stock from ponds. The shrimp's industry is greatly reduced [1]. Especially the non-reproductive problems of the black tiger shrimp are necessary to capture. The results of using too much of the wild shrimp are that cannot control the quality of the shrimp. This problem obstructs progress in the development of animal species and livestock industry.

Eyestalk Ablation is a popular method to stimulate the development of ovaries in black tiger shrimp. Although the cuttings from shrimp ponds can stimulate the development of ovaries, the shrimp is still low quality for the industry. Therefore, understanding the effect of eye shear on the reproduction of black tiger prawn is very important. In [2] authors studies on the expression of genes that result from eye removal comparing of black tiger prawn from pond before and after cutting at different time.

Microarray is a chip-based technology used to study the gene expression levels of tens of thousands of gene simultaneously. The downside of microarray datasets is the large number of features (attributes/genes) while the number of samples is very small. The large number of features is the great challenge for data mining and machine learning. There are many researches related to feature selection method on microarray data for classification [3-4]. But the study about the feature selection method on black tiger shrimp microarray dataset using the biological knowledge is very small because we still cannot decode the genes and find all the functions of the genes like human genomes.

In this paper we use microarray data of the different expression genes between before and after eye removal from black tiger shrimp tissues which may help us to find the informative genes related to black tiger shrimp reproduction. Also, those genes may relate to ovarian development in black tiger shrimp. Moreover, in this study, we investigated all selected genes that are important for the level of ovary development of black tiger shrimp. The goal of our experiment is that if we can find genes which are important, we will have gene marker to indicate the level of ovary development of black tiger shrimp. This will lead to a study on how to improve the good tiger prawn culture.

The rest of this paper is organized as follows. Section II gives a background of Feature selection, Genetic algorithm and related works. Section III the proposed method is described. Section IV reports the experimental results and the conclusion and future works are described in Section V.

II. BACKGROUNDS

A. Feature Selection

Feature selection is a process which selects a relevant feature subset according to an evaluation criterion [5]. The main goals of feature selections are to avoid model overfitting, improve the predictive performance of the model, eliminate irrelevant features and reduce the computational time [6]. In the context of microarray datasets, normally it has a very high dimensionality (the number of features/genes) while the number of instances or sample is very small. Clearly, in this case, feature selection is very important step to decrease the risk of model overfitting and the complexity of predictive model [6-7].

Feature selection methods can be separated into 3 approaches: 1) the filter approach, 2) the wrapper approach, and 3) the embedded approach [6]. The filter approach evaluates the candidate feature or candidate features subset independent from the classifier. There are two types of methods following the filter approach: 1) feature ranking-based methods which evaluate each feature and rank all features according to their relevance. Then, top k features were selected for the classifier and 2) search-based methods which considers the relationship between features in feature subset. Candidate feature subset was created using search method. Then, each candidate feature subset was evaluated using the merit function according to the target problem [8]. On the other hand, the wrapper approach, the best feature subset is selected by a classifier's performance level. In the wrapper approach, the classification algorithm used to measure the merit of each candidate feature subset. Then the best feature subset is selected. The last approach, called the embedded approach, the process of searching a good feature subset is embedded into the classifier construction process. An example of embedded feature selection is decision tree algorithm, where during the tree construction process, a feature is selected at each internal node of the tree.

Note that, in this paper our approach follows the embedded fashion while the algorithm predicts the GSI level of ovarian maturation level in black tiger shrimp using GA, the algorithm also selects the genes (candidate feature subset) which suitable for the predictor. After that we select genes based on how often they are found in individual representation of GA. Finally, the most frequently selected genes are studied in the term of biological perspective.

B. Genetic Algorithms

Evolutionary Algorithms (EAs) are stochastic search methods inspired by the process of natural selection, which introduced in Darwin's' evolutionary theory [9]. There are several types of EAs – e.g. Genetic Algorithm (GA), Genetic Programming (GP), and Evolutionary Programming (EP). In general, GA is the most popular search method used for feature selection [10-11] and also shows outperformed predictive accuracy comparing with other classification algorithms. In this paper, we focus on GA for embedded-based feature selection.

In GA, each individual (chromosome or candidate solution) is evaluated by a fitness function according to the target problem. For feature selection task, the individual is typically represented selected features in the different ways such as a bit string representation, list of feature indexes; and a two-part of bit string. A bit string is the simplest used in many publications on feature selection area [8] where each bit takes the value 1 or 0. Another individual representation is a list of feature indexes which used in [12-13]. List of feature indexes can be separated into two ways: 1) a variable-length representation where each individual consists of the k features (when k <= the total number of features and k is a pre-defined number) and 2) a fixed-length individual representation where the individual length is equal to k and each gene in individual represent index of feature. The last individual representation is named "a two-part bit string" where each individual can be separated into two parts: the selected features and the weight of features [8, 14].

Moreover, GA in the context of feature selections can be separated into two approaches: 1) filter-based GA for feature selection and 2) wrapper-based GA for feature selection [11]. In wrapper-based approach, the fitness function uses the accuracy of a classification model built with the features selected by the individual, while the filter approach uses a simpler fitness function that is independent from the classification algorithm to evaluate the quality of the feature subset represented by the individual.

In general, the first step of GA for feature selection starts with a population of individuals or a population of candidate feature subsets. Second, for each iteration, the GA operations are performed for creating new "child" individuals and each child individual (candidate solution) was evaluated using fitness function. This step is iteratively repeated until a stopping criterion is satisfied. The child individuals tend to inherit characteristics (feature subsets) of good parents and the population in GA tends to evolve to a near-optimal candidate solution (feature subset). Many literatures show that GAs obtain good predictive accuracy results in classification comparing with other traditional search methods on feature selection for classification problem [15-16, 10-11].

C. Related Works

There are many publications which employed GAs as a feature subset selection method in classification problems [17-19].

In [17] proposed the Hybrid approach which including the filter-based IG (Information Gain) and wrapper-based GA for feature selection. The proposed method is divided into two steps. The first step is a filtering feature selection which the discriminative power of each individual feature is calculated using IG. Then, the most informative features are selected. The second step, all selected features from the first step is used in GA. In this step, GA is used as a wrapper method to select the most relevant features out of all features.

Another hybrid trend for feature selection using GA proposes in [18]. In the first step, the features are evaluated using a correlation weight and select the relevant features based on that criterion. Next, the population for the GA is generated using obtained features from the first step. Then, each individual is evaluated. In this case, the fitness function is the predictive accuracy of k-NN.

In [19] proposes a multiple criteria feature subset selection approach. Their method uses

multiple criteria such as t-score, entropy-based and SVM recursive feature elimination. Those criteria are used to select a good feature subset in the feature space. Then the best feature subset according to all criteria is added in a collection of candidate features for GA. In the next stage, a GA is used to evaluate each individual (candidate fitness subset) using a fitness function based on the classification accuracy and number of selected features.

III. PROPOSED METHOD

Our proposed method can be separated into 3 steps: 1) finding the optimized individual length using GA, 2) running GA to predict GSI value using the optimized individual length obtained from step 1, and 3) investigating the top selected genes according to criteria. The detail of each step was described below.

Step 1: Finding the optimized individual length. For this step, we run GA to find the optimized individual length for our prediction model. The candidate individual length are 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, and 50, respectively.

Step 2: Running GA and predict the GSL value. For this step we set the individual length equal to the optimized individual length obtained from previous step and run GA. The best solution was selected from this step and send to Step 3.

Step 3: Investigating genes according their frequency selected from the best solution from step 2. In our experiment, we select genes which selected over 30 times across 1,000 runs.

Note that, for all parameters of the GA such as individual length (l), population size (p), the number of generations (g), gene crossover probability (CrossP) and gene mutation probability (MutP) using in GA are shown in Table I.

TABLE I
PARAMETERS OF GA

GA Parameters	Running GA for	
	Finding the optimized individual length	Predicting GSI
l	2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, and 50.	4 (optimized individual length)
p	100,000	100,000
g	100	100
CrossP	90%	90%
MutP	5%	5%

A. Individual Representation and GA Operations

We use the list of feature index individual representation [12-13] where each individual consists of at most l genes and each gene represents the index of a feature or "0". Note that in this case "0" represent no selected gene. In our experiment, each gene in individual represents the gene id (from 1 to 5,568). The list of feature index individual representation shows in Fig. 1.

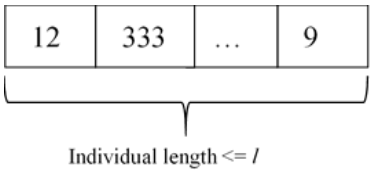

Fig. 1 The List of Feature Index Individual Representation

For GA operations, one-point crossover and bit-flip with random number mutation are used in our experiment. The one-point crossover is the simplest crossover in GA. In this case, a crossover position is randomly selected on parent individuals. Then, the gene values between two parents are swapped producing two new child individuals. The one-point crossover was used in many literatures [18, 20, 12]. Moreover, in our experiment we use bit-flip with random number mutation. The mutation point is randomly selected, then the number in selected gene will be replace with the new random number.

B. Fitness Function

Each individual (candidate feature subset) in the population is evaluated using equation (1).

$$Fitness\ Function = \sum_{i=1}^{r} \left| GSI_i - \sum_{j=1}^{l} \exp(G_{x_{i,j}}) \right| \quad (1)$$

where r is the number of samples, GSI_i is the ovarian maturation level in black tiger shrimp in sample i, l is the individual length; and $exp(G_{x(i,j)})$ is the gene expression data point of $x_{i,j}$.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Datasets and Experiment Setup

In this experiment we use six gene microarray datasets accessed from http://www.ncbi.nlm.nih.gov/geo/, with GEO accession number GSE 29025. Each dataset contains 5,568 genes. For each missing value, we replace missing value with '999'. The dataset characteristic shows in Table II. The first column is a dataset name (or sample) that shows the data after cutting the eyestalk at different time (e.g. D41 means the data after cutting the eyestalk for 4 days of the first dataset). G1 to G5568 are gene index which show the expression level of gene extracting from difference black tiger shrimp tissues. The last column shows the GSI value of each sample.

TABLE II
DATASET CHARACTERISTIC

Dataset	G1	G2	G3	…	G5568	GSI
D11	null	0.20519	0.5639	…	null	1.2
D12	null	Null	1.9663	…	null	1.2
D41	null	null	-0.264	…	null	3.8
D42	null	-0.1114	-0.016	…	null	4.2
D71	null	0.5561	-1.0875	…	null	6.8
D72	null	0.4059	-0.9468	…	null	7.4

Note that, in our experiment, we extract gene from 8 part of black tiger shrimp tissues. The detail of tissues used in experiment described in Table III.

TABLE III
TISSUE DESCRIPTION

No.	Tissue	Meaning
1	GL	Gill
2	HC	Hemocyte
3	Hp	Hapatopancreas
4	HT	Heart
5	IN	Intestine
6	LP	Lymphoid organ
7	OV	Ovary
8	TT	Testis

B. Experimental Results and Discussion

1) Results for Individual Length Optimization

Recall that, first, we run experiment to find the optimal individual length. In this step, we set up GA using all parameters mention in Table I. The candidate individual length used in this experiment are 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, and 50. Table IV shows the predicting error of GSI values from GA using different individual lengths. The first column is a number of run time of GA in our experiment. Then, the rest columns show titled "l =" following with the number denote the individual length equal to corresponding number in each column. For examples, "l=2" means the individual length equal to 2 and "l=3" mean the individual length equal to 3 and so on.

From the experiment results, clearly, the average error of GSI is high when individual length equal to 2 and 3 (GSI = 1.3103 and 0.6601, respectively). Then, we set individual length equal 4, GA slightly obtains the smaller error of GSI (0.4191 which is dropped from 0.6601). After that, when we set the individual length equal to 5,6,7 and so on, the error of GSI values obtained from GA are not much difference (the error values vary between 0.3304 - 0.4536). We decided to choose the optimal individual length equal to 4 because it is the smallest size of feature subset for GA to avoid the redundant and the simplicity of classification model [5].

TABLE IV
THE VALUES OF GSI OBTAINED FROM GA WHEN USING THE DIFFERENT INDIVIDUAL LENGTHS

No	l=2	l=3	l=4	l=5	l=6	l=7
1	1.265122	0.741688	0.576741	0.34296	0.358532	0.519221
2	1.265122	0.676949	0.437845	0.448267	0.471919	0.350948
3	1.265122	0.706455	0.453466	0.488395	0.346854	0.40246
4	1.265122	0.744141	0.451261	0.453014	0.577341	0.437585
5	1.491261	0.537691	0.391484	0.333832	0.339524	0.331197
6	1.265122	0.668659	0.472601	0.49711	0.473793	0.280273
7	1.265122	0.367188	0.420061	0.452824	0.532735	0.296295
8	1.265122	0.69578	0.255767	0.241359	0.527239	0.455802
9	1.491261	0.741688	0.374588	0.421774	0.447496	0.523789
10	1.265122	0.720341	0.357012	0.467658	0.460757	0.418991
Avg	**1.3103**	**0.6601**	**0.4191**	**0.4147**	**0.4536**	**0.4017**
SD	**0.0953**	**0.1195**	**0.084**	**0.0822**	**0.0825**	**0.086**
No	**l=8**	**l=9**	**l=10**	**l=20**	**l=30**	**l=40**
1	0.251309	0.294303	0.379189	0.255258	0.441716	0.440747
2	0.505122	0.337122	0.377943	0.460374	0.33275	0.408749
3	0.234799	0.35198	0.368435	0.4107	0.390279	0.408965
4	0.336497	0.499555	0.38177	0.456588	0.324182	0.411063
5	0.210634	0.550188	0.43767	0.341277	0.418412	0.405594
6	0.356542	0.314332	0.505618	0.323045	0.353449	0.528155
7	0.415517	0.465434	0.440822	0.364502	0.34061	0.419977
8	0.278554	0.268779	0.26788	0.467206	0.516215	0.295507
9	0.241581	0.338287	0.303279	0.328961	0.445483	0.468789
10	0.47343	0.509189	0.220039	0.350817	0.347602	0.417776
Avg	**0.3304**	**0.3929**	**0.3683**	**0.3759**	**0.3911**	**0.4205**
SD	**0.105**	**0.1022**	**0.0854**	**0.0705**	**0.0629**	**0.0583**

2) Results for Gene Selection

After the previous step, we set the individual length of candidate feature subsets equal to four and run GA to predict the GSI value.

In this step GA is run for 1,000 times. Each run the best candidate solution was selected by GA. Then, from 1,000 selected solutions, we sort all selected genes according their frequency. Next, we select genes which selected by GA over 30 times across 1,000 runs. The list of selected genes shows in Table V.

TABLE V
THE LIST OF 18 GENES WERE SELECTED MORE THAN 30 TIMES FROM GA

Freq.	UNIQID	Gene Description	Tissue
131	PMYT10G05	Unknown	HT
118	PMYT04B05	Unknown	HT
86	PMYT30G03	Conserved hypothetical protein	GL
70	PMYT31G05	Guanine nucleotide binding3-like	GL
66	PMYT31E05	Mgc80370 protein	GL
64	PMYT32D12	Unknown	GL
63	PMYT04D11	Cytochrome c oxidase subunit i	HT
61	PMYT32H08	ATP synthase f0 subunit 6	GL
56	PMYT30E04	Fact complex subunit spt16 (facilitates chromatin transcription complex subunit spt16) (Fact 140 kda subunit)(chromatin-specific transcription elongation factor 140 kda subunit)	GL
52	PMYT05F03	tRNA-Ile	HT
50	PMYT05C10	Hypothetical protein	HT
50	PMYT05F07	Hypothetical protein	HT
49	PMYT32G12	Unknown	GL
46	PMYT09A10	Unknown	HT
35	PMYT04C09	ENSANGP00000011689	HT
32	PMYT30H05	Ubiquinol-cytochrome c reductase complex ubiquinone-binding protein qp-c	GL
31	PMYT30A05	Protease m1 zinc metalloprotease	GL
30	PMYT30B09	60S ribosomal protein L44,	GL

Eighteen genes which more than 30 times from 1,000 iterations are selected to investigate in details. The most frequently selected are gene extract from heart and gill tissues of black tiger shrimp. Table VI shows the list of selected genes and literatures which supports those genes in biological perspective.

The first place of frequently selected gene is unknown gene (PMYT10G05 with 131 selected times) and the second place is unknown gene (PMYT04B05 with 118 selected times), both of them are genes extracted from heart tissue of black tiger shrimp. For these two genes there are no evident to support how related with the Gonadosomatic Index of ovarian maturation level in black tiger shrimp. The interesting gene is "Guanine nucleotide binding3-like: GNL3" gene which selected 70 times out of 1,000 run of GA. GNL3 produce protein in nucleus cell which related to pre-rRNA processing, cell division and genomic integrity [21-22].

Next gene is named "Fact complex subunit spt16: FACT complex" which is a chromatin factor. FACT complex acts to reorganize nucleosomes and involves in multiple processes such as mRNA elongation, DNA replication and DNA repair [23-24].

From the experimental results, our algorithm also founds "tRNA-Ile" gene which related to ATP-building, tRNA processing and tRNA modification in the cell.

"Protease m1 zinc metalloprotease" is found 31 times from our experiment. This gene shows a big role in cell division (Mitosis and Meiosis) and also involves with hormones or minerals in maintaining balance of the body of plants and animals [25].

TABLE VI
THE LIST OF SELECTED GENES AND THE LIST OF LITERATURES WHICH SUPPORTS THOSE GENES IN BIOLOGICAL PERSPECTIVE

Freq.	Gene Description	Details
131	Unknown	No report
118	Unknown	No report
86	Conserved hypothetical protein	No report
70	Guanine nucleotide binding3-like	Tsai and Meg, 2009 Tsai,2014
66	Mgc80370 protein	No report
64	Unknown	No report
63	Cytochrome c oxidase subunit i	Energy metabolism related
61	ATP synthase f0 subunit 6	
56	Fact complex subunit spt16 (facilitates chromatin transcription complex subunit spt16) (Fact 140 kda subunit)(chromatin-specific transcription elongation factor 140 kda subunit)	Orphanide et al., 1998 LeRoy et al., 1998
52	tRNA-Ile	ATP-building tRNA processing tRNA modification
50	Hypothetical protein	No report
50	Hypothetical protein	
49	Unknown	
46	Unknown	
35	ENSANGP00000011689	
32	Ubiquinol-cytochrome c reductase complex ubiquinone-binding protein qp-c	Creating protein
31	Protease m1 zinc metalloprotease	Peer,2011
30	60S ribosomal protein L44,	GL

Moreover, we analyze the trend of gene expression level of 18 selected genes in datasets. Fig. 2 shows the trend of all selected 18 genes. Clearly, the gene expression level is growing when the day of eye removal takes longer. For example, when remove eye on day one (D11 and D12 datasets), the level of gene expression is less than 1 (between -1 to 1)

while remove eye on day four (D41 and D42 datasets), the gene expression level are vary from 1.0 to 3.0. Lastly, when remove eye on day seven (D71 and D72), the level of gene express is from 1.5 to 5.0, in approximately.

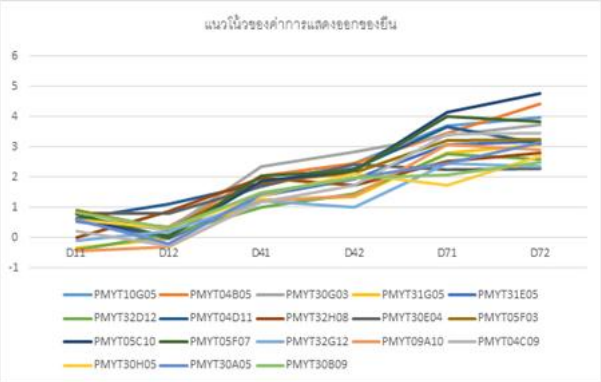

Fig. 2 The Trend of Gene Expression of Selected Genes after Eye Removal

V. CONCLUSION AND FUTURE WORKS

In this work, an embedded-based gene selection method in the black tiger shrimp microarray dataset using Genetic Algorithm is proposed. Moreover, we analyze the effect of selected genes for ovarian maturation level in the term of biological perspective. We separate our approach into three steps: 1) finding the optimized individual length, 2) running GA and Predict the GSI value using the optimized individual length obtained from previous step, and 3) investigating the selected genes according their frequency from GA run. Six black tiger shrimp microarray datasets publically available are used in our experiment. The experimental results show that the optimized individual length for GA is 4 and the 18 selected genes are from heart and gill tissues of black tiger shrimp. For future work, we will study on the effect of the unknown selected genes from our experiment and take the relationship between selected genes into account.

VI. ACKNOWLEDGMENTS

This work was supported by Nakhon Pathom Rajabhat University (NPRU).

REFERENCES

(Arranged in the order of citation in the same fashion as the case of Footnotes.)

[1] Tanticharoen, M., Flegel, T.W., Meerod, W., Grudloyma, U., and Pisamai, N. (2008). "Aquacultural biotechnology in Thailand: the case of the shrimp industry". International Journal of Biotechnology, Vol. 10, pp. 588.

[2] Uawisetwathana, U., Leelatanawit, R., Klanchui, A., Prommoon, J., Klinbunga, S., and Karoonuthaisiri, N. (2011). "Insights into eyestalk ablation mechanism to induce ovarian maturation in the black tiger shrimp". PLoS ONE 6, e24427.

[3] Hengpraprohm, S. (2013). "GA-Based Classifier with SNR Weighted Features for Cancer Microarray Data Classification". International Journal of Signal Processing Systems (IJSPS), Vol. 1(1), pp. 29-33.

[4] Joong, K.K. and Cho, S.B. (2014). "Meta-classifiers for high-dimensional, small sample classification for gene expression analysis". Pattern Analysis and Applications, DOI 10.1007/s10044-014-0369-7.

[5] Lui, H. and Motoda, H. (1998). "Feature Selection for Knowledge Discovery and Data Mining". Massachusetts: Kluwer Academic.

[6] Saeys, Y., Inza, I., and Larranaga, P. (2007). "A review of Feature Selection Technique in Bioinformatics". Bioinformatics, Vol. 23(19), pp. 2507-2517.

[7] George, G.V.S. and Raj, V.C. (2011). "Review on Feature Selection Techniques and the Impact of SVM for Cancer Classification Using Gene Expression Profile". Computer Science & Engineering Survey, August, Vol. 2(3), pp. 16-26.

[8] Jungjit, S. (2016). "New Multi-Label Correlation-Based Feature Selection Methods for Multi-Label Classification and Application in Bioinformatics". (Doctoral dissertation, University of Kent).

[9] Eiben, A.E. and Smith, J.E. (2003). "Introduction to evolutionary computing". Berlin: springer, Vol. 53.

[10] Freitas, A.A. (2002). "Data Mining and Knowledge Discovery with Evolutionary Algorithms". Springer Science & Business Media.

[11] Freitas, A.A. (2005). "Evolutionary algorithms for data mining". In Data mining and knowledge discovery handbook, Springer, Boston, MA, pp. 435-467.

[12] Hong, J.H. and Cho, S.B. (2006). "Efficient huge-scale feature selection with speciated genetic algorithm". Pattern Recognition Letters, Vol. 27(2), pp. 143-150.

[13] Lee, C.P. and Leu, Y. (2011). "A novel hybrid feature selection method for microarray data analysis". Applied Soft Computing, Vol. 11(1), pp. 208-213.

[14] Hernandez, J.C.H., Duval, B., and Hao, J.K. (2007). "A genetic embedded approach for gene selection and classification of microarray data". In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, Springer, Berlin, Heidelberg, pp. 90-101.

[15] Sharpe, P.K. and Glover, R.P. (1999). "Efficient GA based techniques for classification". Applied Intelligence, Vol. 11(3), pp. 277-284.

[16] Kudo, M. and Sklansky, J. (2000). "Comparison of algorithms that select features for pattern classifiers". Pattern recognition, Vol. 33(1), pp. 25-41.

[17] Yang, C.H., Chuang, L.Y., and Yang, C. (2009). "IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data". Medical and Biological Engineering, Vol. 30(1), pp. 23-28.

[18] Chuang, L.Y., Yang, C.H., Wu, K.C., and Yang, C.H. (2011). "A hybrid feature selection method for DNA microarray data". Computer in Biology and Medicine, Vol. 41(4), pp. 228-237.

[19] Tan, F., Fu, X., and Zhang, Y. (2008). "A genetic algorithm-based method for feature subset selection". Soft Comput, Vol. 12(2), pp. 111-120.

[20] Dessì, N. and Pes, B. (2009). "An evolutionary method for combining different feature selection criteria in microarray data classification". Journal of Artificial Evolution and applications.

[21] Tsai, R.Y.L. and Meng, L. (2009). "Nucleostemin: A latecomer with new tricks". The International Journal of Biochemistry & Cell Biology, Vol. 41, pp. 2122-2124.

[22] Tsai, R.Y.L. (2014). "Turning a new page on nucleostemin and self-renewal". Journal of Cell Science, Vol. 127, pp. 3885-3891.

[23] Orphanides, G., Leroy, G., Chang, C.H., Luse, D.S., and Reinberg, D. (1998). "FACT, a Factor that Facilitates Transcript Elongation through Nucleosomes Cell". Vol. 92, pp. 105-116.

[24] Leroy, G., Orphanides, G., Lane, W.S., and Reinberg, D. (1998). "Requirement of RSF and FACT for Transcription of Chromatin Templates in Vitro". Science, Vol. 282, pp. 1900-1904.

[25] Peer, W.A. (2011). "The role of multifunctional M1 metallopeptidases in cell cycle progression". Annals of botany, Vol. 107(7), pp. 1171-1181.