Microarray Gene Selection Using Self-Organizing Map

SIRIRUT VANICHAYOBON, SIRIPHAN WICHAIDIT, WIPHADA WETTAYAPRASIT
Artificial Intelligence Research Laboratory, Department of Computer Science,
Prince of Songkla University, Songkhla, THAILAND

Abstract: - Accuracy, precision, and rapidity of disease prediction are important for disease evaluation in clinic and laboratory studies because different diseases would have different drugs and treatments. This study presents a new technique for cancer prediction from DNA microarray data. The prediction composes of two main steps that are the step of the important gene selection by using statistic methodology and the step of clustering cancer data by using self-organizing map. The experimental DNA microarray data sets are carcinoma, leukemia, and lung cancer. The experimental results are the rules of gene with 100% accuracy for cancer prediction.

Key-Words: - DNA Microarray, Self-Organizing Map, Bioinformatics, Gene Prediction

1 Introduction

The disease prediction using DNA micorarray is a well-known method [1,2,3] because of the advantage technique of DNA microarray that is different from the traditional biological method. With the DNA microarray, a large number of gene expressions will be observed at one time. The clustering technique is also included with the study that contributes to increasing efficiency and reducing cost of disease evaluation. The clustering techniques are hierarchical clustering [2,3], k-mean clustering [4], fuzzy neural network [5,6], neural network [7], support vector machine [8], and so on. However, the barrier of bringing DNA microarray to the analysis is the huge number of genes or attributes that are thousands up to ten thousand genes when other data have less attributes. Further more, the number of records of DNA microarray for disease analysis is quite small [9].

Since all huge numbers of existing genes do not have effect on clustering of cancer data. Then before bringing DNA microarray data to analysis by the step of clustering, there should be the step of dimensional reduction or reducing the number of genes that does not have effect on the disease prediction [13]. Dimensional reduction will contribute to reducing working time of the system because time used for the less number of genes will be less than the time used for the more number of genes [10]. The dimensional reduction of data will contribute to receiving a number of important genes. There are many well-known techniques used for the dimensional reduction such as principle component analysis [11], minimum entropy [12], support vector machine and so on.

This paper presents a new technique base on the statistical hypothesis testing using level of significance (p-value) of both unpaired value and paired value for dimensional reduction. Unpaired value is appropriate with data that do not have the characteristic of pairing when paired value is appropriate with data that have the characteristic of pairing. The self-organizing map is used for cancer clustering.

Section 2 of this paper will mention on the DNA microarray, self-organizing map, and level of significance (p-value). Section 3 presents steps of Microarray Gene Selection by using Self-Organizing Map (MGS_SOM) for cancer prediction. Section 4 presents the experimental results. And section 5 is the conclusion.

2 DNA Microarray, Self-Organizing Map, and Level of Significance

2.1 DNA Microarray

DNA microarray is a set of various types of DNA with differences of base orderly dropped or has been synthesized systematically on the surface of the chip. The DNA microarray can be called “DNA Chip” or “Gene Chip” [14].

The experiment of microarray composes of two samples that are controlled sample and tested sample. For examples, controlled sample is healthy cell and tested sample is sick cell. Complementart DNA or cDNA used for the experiment will be hybridized on a spot of slide surface. This rule of particular paired base are A pairs with T when C pairs with G. Normally, cDNA will be labeled with fluorescent tabs. The fluorescent tabs have two

colors which green color (Cy3) is the controlled sample and red color (Cy5) is the tested sample. Cell molecules that have been labeled with fluorescent tabs will be detected when being stimulated by laser beam [14].

The experimental result of microarray composes of many spots of colors as showed in Fig 1. Color of each spot occurs from the capturing of cDNA. Red spots will relate to high expression genes. Green spots will relate to low expression genes. Yellow color indicates that the components of both gene samples are equal. Black color indicates that positions have no mixing of both gene samples. The color positions received from array can be conversed to real values in the form of matrix. For example, matrix M which is called gene expression when M_{ij} is equal to $\log_2(Cy5/Cy3)$ where i is the number of gene and j is the number of sample data. Matrix M at line i shows the gene expression and column j shows the experimental value [14].

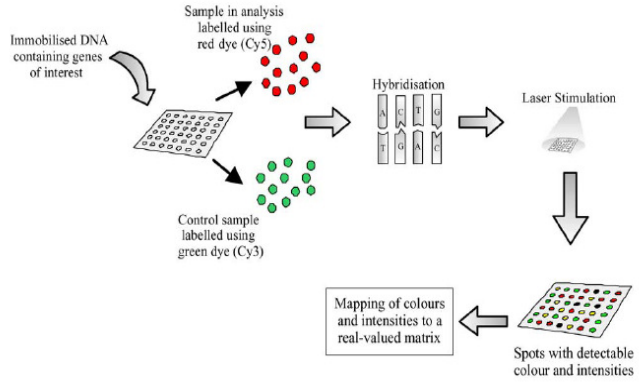

Fig 1. Process of Producing DNA Microarray.

2.2 Self-Organizing Map

Self-organizing map is an unsupervised learning technique of artificial neural network, which is used for clustering group of data. The benefit of self-organizing map is reducing the dimensional data to 1 dimension or 2 dimensions. Self-organizing map composes of neurons which represent each or tested data that show on the map table [15,16,17]. Learning steps of Self-Organizing Map composes of 4 main steps [16] are 1) Synaptic weight initialization: randomly selects neuron samples on the map table or sometime calls sample vector from tested data and specify weight value for neuron, 2) Competitive process: calculates distance between randomly selected samples and other neurons on the map table. The neuron with the least distance will be the best matching unit (BMU) or has the most characteristic similar to the most randomly selected neuron sample, 3) Corporative process: calculate the radius of neighbor neurons using Gaussian function after receiving the winning neurons, and 4) Adaptive process: nodes in the radius calculated on the previous step will be adjusted to receive the most weight value close to the selected nodes.

2.3 Level of Significance

Level of significance is an analytical statistic method using probability for main data management. The p-value received from the testing hypothesis will be used to deny the experimental hypothesis. If p-value is small, this means that there will be more weight to deny the experimental hypothesis, or this means that the group of data has much difference in probability value. If p-value is large, this means that there will be less weight to deny the experimental hypothesis, or this means that the group of data has little difference in probability value. Then the selection of small p-value will represent the group of data that the representing groups of genes received will have significant difference from all other groups. The p-value can be calculated from t-test. There are two types of formulas to calculate t-test [19] that are paired t-test as equation (1) to (3).

$$t-test=(\bar{x}-\bar{y})\sqrt{\frac{n(n-1)}{\sum_{i=1}^{n}(\hat{x}_i-\hat{y}_i)^2}} \quad (1)$$

$$\hat{x}_i=(x_i-\bar{x}) \quad (2)$$

$$\hat{y}_i=(y_i-\bar{y}) \quad (3)$$

Let x be the average value of the first data set, y be the average value of the second data set, n be the number of samples. Unpaired t-test shows as equation (4) and (5).

$$t-test=\frac{\bar{x}-\bar{y}}{\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}} \quad (4)$$

$$s^2=\frac{\sum_{i=1}^{n_1}(x_i-\bar{x})^2-\sum_{i=1}^{n_2}(y_i-\bar{y})^2}{n_1+n_2-2} \quad (5)$$

Let n1, n2 be the numbers of sample 1 and sample 2, respectively [18,19].

The p-value can be calculated by using equation (6) and (7) as follows.

If $|t_{value}|>t_{\alpha,n-1}$, then $p_{value}<2\alpha$ (6)

If $|t_{value}|<t_{\alpha,n-1}$, then $p_{value}>2\alpha$ (7)

When t_{value} is a value received from equation (1) or (4) and $t_{\alpha,n-1}$ is the value obtained from the statistic table. There is a notice that p-value is invert with t-test (t_{value}). The p-value is the area under graph on the rightmost side or leftmost side. From the hypothesis, if p-value is small (t-test is large), this means that there is a few area of representing value on the rightmost side or leftmost side that will contribute to high probability value of denying hypothesis. However, if p-value is large (t-test is small), this means that there is a large area of representing value on the rightmost side or leftmost side that will contribute to low probability value of denying hypothesis [19].

3 Microarray Gene Selection Using Self-Organizing Map

The Microarray Gene Selection Using Self-Organizing Map (MGS_SOM) algorithm for cancer prediction composes of 4 parts, which are part 1. Gene Selection Process, part 2. Self-Organizing Map Clustering Process, part 3. Rule Creation Process, and part 4 Rule Evaluation Process. Details of MGS_SOM algorithm are shown in Fig 2.

Part 1. Gene Selection Process: This part has steps as follows. Step 1.1 is a randomly selection of data and then separate data into 2 sets. These two sets are trained data set and tested data set. Each data set has to compose of samples that have both cancer cells and normal cells. Step 1.2 is t-test value calculation for each of gene samples. If data have relationship in term of pairs, this means that there is the cancer data sample from only one person with disease cell data and normal cell data are in pairs, then calculate a paired t-test type of t-test value calculation from equation (1) to (3). If the data are not in pairs, the data have no relationships. This means that cancer data are from general samples that do not come from the same person. If disease cell data and normal cell data are from different people, then calculate an unpaired t-test type of t-test value from equation (4) and (5). Step 1.3 is p-value calculation from equation (6) and (7). Step 1.4 is Gene Selection. This step is gene selection from genes with low p-value.

Part 2. Self-Organizing Map Clustering Process: This part is taking the selected genes from step 1.4 clustering by using Self-Organizing Map. After that, the accuracy of trained data set will be calculated.

Part 3. Rule Creation Process: This part is a rule creation by taking the genes that have accuracy value calculated from part 2 to create rules in the form of if-then rule.

Part 4. Rule Evaluation Process: This part is taking the tested data set to be tested with the rules from part 3 to select rules that have high accuracy.

Part 1 : Gene Selection Process

1.1) Separate the data into trained data set and tested data set.

1.2) Calculate t-test value of each gene (attribute).

- If data have characteristic of pair relationship, the paired t-test will be calculated from equation (1), (2), and (3).
- If data do not have characteristic of pair relationship, the unpaired t-test will be calculated from equation (4) and (5).

1.3) Calculate p-value by using equation (6) and (7).

1.4) Select the lowest p-value.

↓

Part 2 : Self-Organizing Map Clustering Process

2.1) Take the selected genes on step 1.4 for clustering using Self-Organizing Map.

2.2) Find the accuracy from the trained data set.

↓

Part 3 : Rule Creation Process

3.1) Select genes that have high accuracy from step 2.2 to create rules in the form of if-then rules for cancer prediction.

↓

Part 4 : Rule Evaluation Process

4.1) Take tested data set to test with rules created from step 3.1.

4.2) Calculate accuracy value obtained from the testing data set.

4.3) Select rules that have high accuracy from both trained data set and tested data set to create rules for cancer prediction.

Fig 2. MSG_SOM Algorithm.

4 Experimental Results

The experimental results by using DNA microarray databases of carcinoma, leukemia, and lung cancer are the followings.

4.1 DNA Microarray Data of Carcinoma

Data compose of total 7,457 genes for a total of 36 samples. There were 18 cancer samples and 18 normal samples [20]. Samples have paired characteristics. The experiment results indicated the number of genes and p-value in each range as showed in Fig 3. The experiment found that most of

the genes (2,727 genes) were in the range of low p-value (0 to 0.1). The average accuracy of each range of p-value was shown in Fig 4. The experimental result is shown that the accuracy (79.5%) with low p-value would have more accuracy than the accuracy (69.3%) with high p-value. The average accuracy in each range of t-test is shown in Fig 5. The experimental result is shown that the accuracy (79.5%) with high t-test value would have more accuracy (69.3%) than accuracy with low t-test value. This means that p-value and t-test inverse in their values.

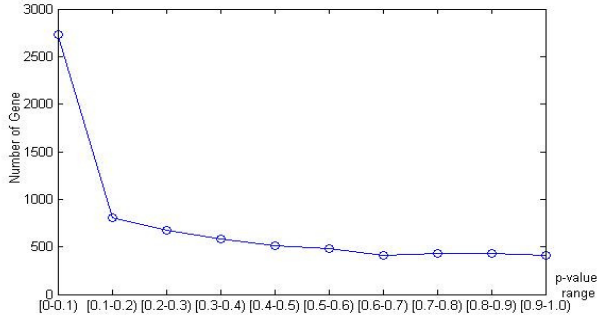

Fig 3. Shows the number of genes from each range of p-value.

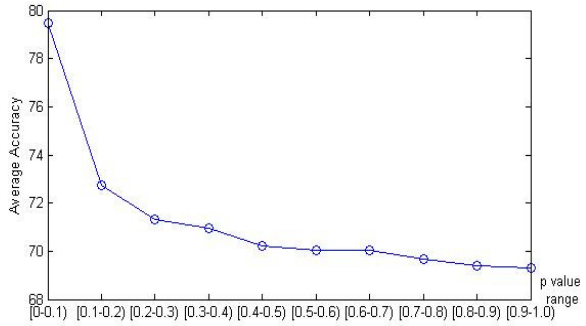

Fig 4. Shows the average accuracy from each range of p-value.

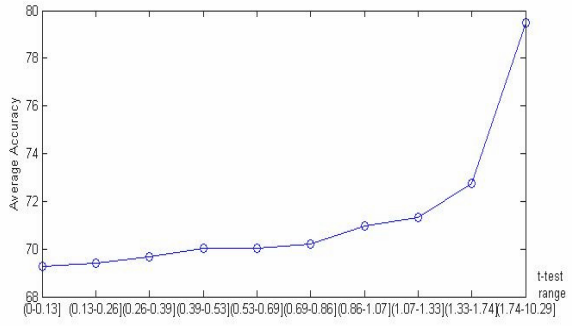

Fig 5. Shows the average accuracy from each range t-test.

The experiment used MGS_SOM algorithm to select the first 100 genes of lowest p-value less than 4.55E-06. The genes in this range have high accuracy equal to 93.5%. Then the selected genes are clustering by SOM, which will receive 25 genes that have high accuracy equal to 100%. After that, 25 genes are used to create rules, and finally rules are tested by using tested data set to receive 11 rules with 100% accuracy. The gene numbers are T64297, M97469, T96548, J02854, Z49629, U37019, M63391, M76378, J03037, U17899, and M77836 as shows in Table 1. For example, if gene number T64297 has value less than 703.76, then the person is classified as a carcinoma patient and so on.

Table 1: Rules of Carcinoma genes.

Gene No	Rule	Accuracy
T64297	**If T64297 < 703.76 then cancer**	**100 %**
M97469	If M97469 < 158.98 then cancer	100 %
T96548	If T96548 < 251.27 then cancer	100 %
J02854	If J02854 < 137.75 then cancer	100 %
Z49629	If Z49629 < 54.599 then cancer	100 %
U37019	If U37019 < 107.01 then cancer	100 %
M63391	If M63391 < 208.75 then cancer	100 %
M76378	If M76378 < 218.89 then cancer	100 %
J03037	If J03037 < 7.4663 then cancer	100 %
U17899	If U17899 > 29.822 then cancer	100 %
M77836	If M77836 > 46.619 then cancer	100 %

4.2 DNA Microarray Data of Leukemia

Data compose of 7,129 genes for a total of 72 samples. There were 47 Actual Lymphoblastic Leukemia (ALL) samples and 25 Actual Myeloid Leukemia (AML) samples [21]. Data are independent from each other (unpaired data).

The experimental result indicated the number of genes and p-value of each range, which is shown in Fig 6. Most of the genes (2,654) are in the range of low p-value (0 to 0.1). The accuracy from each range of p-value is shown in Fig 7. The experimental result found that the accuracy (83.6%) with low p-value would have more accuracy than the accuracy (80.6%) with high p-value. The average values of accuracy in each range of t-test value is shown in Fig 8. The experimental result indicated that the accuracy (83.6%) at high t-test value would have more accuracy than the accuracy (80.6%) with low t-test value. This means that p-value and t-test inverse in their values.

The experiment used MGS_SOM algorithm to select the first 100 genes of lowest p-value less than 2.15E-06. The genes in this range have average accuracy equal to 89.7%. Then the selected genes are clustering by SOM, which will receive 4 genes that have accuracy equal to 100%. After that, the 4 genes are used to create rules. The rules are tested by using tested data set to receive 2 rules with 100% accuracy. The gene numbers are G6855 and G2288 as shows in Table 2. For example, if gene number

G6855 has value less than 475.78, then the person is classified as an AML patient. However, if gene number G6855 has values more than or equal to 475.78, this means that the person is classified as an ALL patient, and so on.

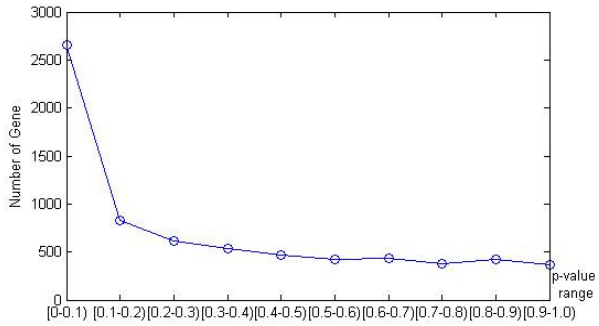

Fig 6. Shows the number of genes from each range of p-value.

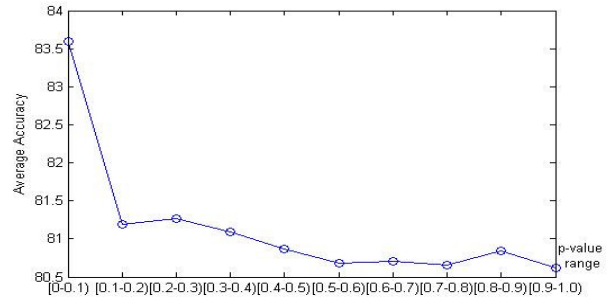

Fig 7. Shows average accuracy from each range of p-value.

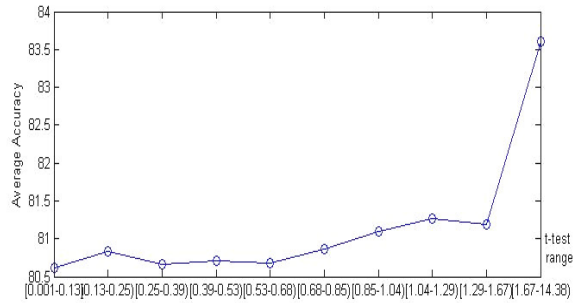

Fig 8. Shows the average accuracy from each range of t-test.

Table 2: Rules of leukemia genes.

Gene No	Rule	Accuracy
G6855	**If G6855 < 475.78 then AML else ALL**	**100 %**
G2288	If G2288 < 222.62 then ALL else AML	100 %

4.3 DNA Microarray of Lung Cancer

Data compose of 12,533 genes for a total of 181 samples. There were 31 Mesothelioma samples and 150 ADCA samples [21]. Data are independent from each other (unpaired data).

The experimental result indicated the number of genes and p-value in each range as shows in Fig 9. Most of the genes (5,890) are in the range of low p-value (0 to 0.1). The average accuracy from each range of p-value is shown in Fig 10. The experiment found that the accuracy (91.8%) at low p-value would have more accuracy than accuracy (88.7%) at high p-value.

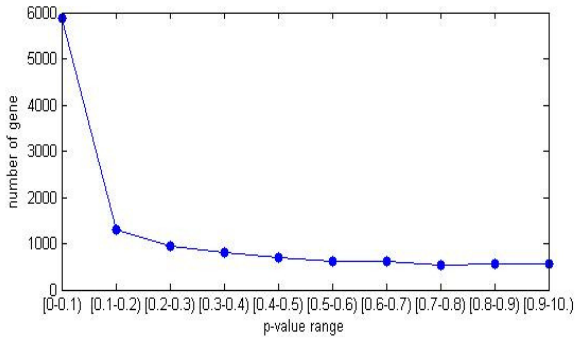

Fig 9. Shows the number of genes from each range of p-value.

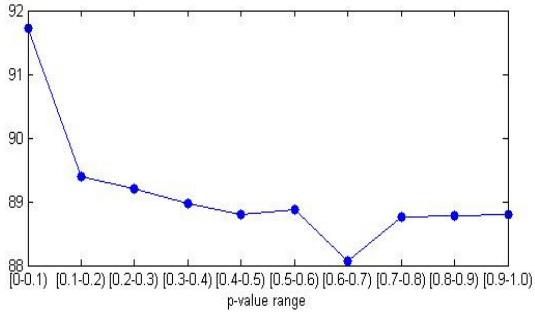

Fig 10. Shows average accuracy from each range of p-value.

The experiment used MGS_SOM algorithm to select the first 100 genes of lowest p-value less than 5.72E-06. The genes in this range have average accuracy equal to 96.8%. Then the selected genes are grouped by SOM, which will receive 44 genes that have accuracy more than 98%. After that, the 44 genes are used to create rules. The rules are tested by using tested data set to receive 2 rules with 100% accuracy. The gene numbers are G12114 and G5301 as shows in Table 3. For example, if gene number G12114 has value less than 48.65, then the person is classified as a Mesothelioma patient. However, if gene number G12114 has values more than or equal to 48.65, this means that the person is classified as an ADCA patient, and so on.

Table 3: Rules of lung cancer genes.

Gene No	Rule	Accuracy
G12114	If G12114 < 48.65 then Mesothelioma else ADCA	**100 %**
G5301	If G5301 <-159.68 then Mesothelioma else ADCA	100 %

5 Discussion and Conclusion

The experiment discusses the 4 parts of MGS_SOM algorithm from the selection of genes by using p-value and taking those genes clustering by SOM. The experiment can extract knowledge from the number of more than 7,000 genes in the form of if-then rules that easy to understandable by human. The experiment also received a small numbers of genes with 11 and 2 genes, respectively.

From the experimental results statistic method and Self-Organizing Map are methods of neural network that is used for gene selection. These techniques show the result of gene selection with the simple and small numbers of rule. The rules can predict cancer with 100% accuracy. The experiment indicates that MGS_SOM algorithm can be used to predict cancer, and can be applied for some other diseases with DNA microarray.

References:

[1] N. Manfred, C. Laiwan, Informative Gene Discovery for Cancer Classification from Microarray Expression Data, IEEE, (2005) pp: 393 - 398

[2] J. Herrero, A. Valencia, J Dopazo, A hierarchical Unsupervised Growing Neural Network for Clustering Gene Expression Patterns, Bioinformatics 17, (2001) pp:126-136.

[3] A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: a Review, ACM Computation. Surveys 31 (3) (1999), pp: 264-323.

[4] Y. Naoki and K. Manabu, Clustering Gene Expression Data Using Self-Organizing Maps and K-Mean Clustering, (2003), pp: 3211-3215.

[5] C. Feng , X. Wei , and W. Lipo, Gene Selection and Cancer Classification Using a Fuzzy Neural Network, IEEE (2004), pp: 555-559.

[6] W. Lipo, M. Senior, C. Feng and X. Wei, Accurate Cancer Classification Using Expressions of Very Few Genes, IEEE, (2007), pp: 40 -53

[7] Khan J., et al.: Classification and Diagnostic Prediction of Cancer using Gene Expression Profiling and Artificial Neural Network, Nature Medicine 2001, 7(6), pp: 637-679.

[8] Guyon I., Weston J., Barnhill S., Vapnik V., Gene Selection for Cancer Classification using Support Vector Machines, Maching Learning, (2002), pp: 389-422.

[9] T. Robert, H. Trevor, Balasubramanian N. and C. Gilbert, Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression, PNAS, (2002), pp: 6567-6572.

[10L. Xiaoxing, K. Arun, and M. Adrian, An Entropy-based Gene Selection method for Cancer Classification using Microarray Data, BMC, (2005).

[11] R. Soumya, S. M. Stuart and A. B. Russ, Principle Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series, Pacific Symposium on Biocompution (2000), pp: 452-463.

[12] L. Haifeng, Z. Keshu, J. Tao, Minimum Entropy Clustering and Applications to Gene Expression Analysis.

[13] R. Mansoor, G. Iqbla, G. David and C. L. Ross, Feature Selection and Classification of Gene Expression Profile in Hereditary Breast Cancer,IEEE,(2005), pp: 315-320.

[14] Lalinka de compos Teixeira Gomes, Fernando J. Von Zuben, Pablo Moscato, A Proposal for Direct-Ordering Gene Expression Data by Self-Organising Maps, Elsevier, (2004), pp: 11-21.

[15] A. Sugivamat, M. Kotani, Analysis of Gene Expression Data by using Self-Organizing Maps and K- means Clustering, IEEE, 2002, pp: 1342-1345.

[16] W. Wiphada and N. Putthiporn, Knowledge Extraction from Self-Organizing Map Using Minimization Entropy Principle Algorithm, ISCIT (IEEE), 2006.

[17] T. Petri, K. Mikko, W. Garry, C. Eero, Analysis of Gene Expression Data using Self-Organizing Maps, FEBS, 1999.

[18] J. Vesanto, J. Himberg, E, Alhoniemi, J. Parhankangas, SOM Toolbox for Matlab5, Libella Oy Espoo, 2000.

[19] J. Dovore and R. Peck, Statistics: The Exploration and Analysis of Data, 3rd ed. Pacific Grove, CA: Duxbury Press, 1997.

[20] Notterman, et al, Cancer Research vol. 61: 2001. [On-line] Available:http://microarray1. princeton.edu/oncology/carcinoma.html, August, 2006.

[21] Kent Ridge Bio-medical Data Set. [On-line] Available:http://sdmc.lit.org.sh/GEDatasets/Data sets.html.