International Journal of Bio-Science and Bio-Technology
Vol.6, No.2 (2014), pp.169-178
http://dx.doi.org/10.14257/ijbsbt.2014.6.2.17

Biomarker Discovery and Data Visualization Tool for Ovarian Cancer Screening

Ki-Seok Cheong[2,3], Hye-Jeong Song[1,3], Chan-Young Park[1,3], Jong-Dae Kim[1,3] and Yu-Seop Kim[1,3 *]

[1]*Dept. of Ubiquitous Computing, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 200-702 Korea*
[2]*Dept of Computer Engineering, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 200-702 Korea*
[3]*Bio-IT Research Center, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 200-702 Korea*
(vseominjungv, hjsong, cypark, kimjd, yskim01)@hallym.ac.kr

**Corresponding Author*: *yskim01@hallym.ac.kr*

Abstract

With the increase in various clinical applications of medical knowledge, a large amount of bio-data have been generated. In this paper, we report on an integrated software tool developed to enable the easy analysis of such bio-data for diagnostic medical testing without the deep use of statistics-related knowledge or tools. Specifically, this system provides an analysis tool for biomarker discovery by applying data mining techniques. It also provides a tool to visualize data, thus enabling a human analyst to easily analyze data, aided by the system. The biomarker data used in this system were generated by using the Luminex equipment, but data generated by other equipment can be used, too. The main modules include marker selection, data visualization and marker evaluation, which have been developed on the basis of the MATLAP. This system is tailored to the early diagnosis of ovarian cancer.

Keywords: *biomarker, Data-Mining Tools, Marker Selection, Multiplex Immunoassay, Logistic Regression, MATLAB*

1. Introduction

With the increase in clinical applications of medical knowledge, the analysis of various bio-data in the field of diagnostic medical testing is increasing in importance. Analysis of such bio-data involves both statistical processes and easy graphing, such as the R project, SPSS and GraphPad PRISM [1-3]. However, the statistical analysis systems require clinicians to have an extensive knowledge of statistics as well as the ability to use these tools.

Data-driven reasoning processes ensure diagnostic accuracy by applying appropriate algorithms, and should be applicable to a variety of cases. However, in most cases, clinicians

* Corresponding author

ISSN: 2233-7849 IJBSBT

have their data analysis performed by statistical professionals. This can ensure diagnostic accuracy, but a variety of empirical knowledge cannot be taken into account in the analysis. Thus, in order for clinicians to utilize their own empirical knowledge while conducting data analysis, it is necessary to provide them with technology for the visualization and analysis of the data, leading to the generation of useful knowledge [4].

In the biomarker data analysis system, a method to enhance diagnostic accuracy is sought by analyzing each performance of various combinations of multiple biomarkers by using statistical techniques. This analysis system converts the values of protein concentration obtained from a multi-immunoassay method into an easily analyzable format, visualizes the form, and finds patterns inherent to the data, using data mining techniques [4].

To detect disease in an early stage and monitor the effects of treatment, molecular changes in biomarkers that occur during the development of certain diseases are detected to quantify changes. By identifying the association between the disease and the changes in biomarkers on the basis of the analysis results, it is possible to detect the disease in its early stage. In this regard, research has been increasing, in both the methods and tools of biomarker discovery in the field of biotechnology [5].

In diagnostic studies of serum-based protein levels, the methods of multiplex proteome analysis, such as Luminex and ELISA, are emerging as useful platforms due to their high sensitivity [5]. In particular, Luminex can measure multiple markers at the same time and has high-speed throughput and high-sensitivity performance, and has been used in a variety of diagnostic testing in recent years [6-9].

In this paper, we propose a system that can conduct statistical analysis of multi-marker data required for disease diagnosis in the field of diagnostic medical testing. The system has been developed for clinicians that lack statistical knowledge or the ability to use statistical tools. The system includes visualization techniques that measure biomarker data needed to diagnose disease with the use of the Luminex equipment, converts the data into an easy-to-analyze format, and includes data mining techniques that find patterns inherent to the data and extract useful knowledge. The main modules include marker selection, data visualization and marker evaluation modules, developed using MATLAB that can enable uni- or multi-dimensional graphing and include various bioinformatics and statistical packages.

Chapter 2 discusses the input data format used in this system, and chapter 3 introduces the algorithms and the overall system implementation. Chapter 4 provides a list of assessment methods possible in the system, and Chapter 5 describes the data visualization part of the system. Finally, Chapter 6 draws conclusions regarding the system.

2. Sample Data Format

The input data for the program are obtained by reacting serum samples from cancer patients and normal subjects with the Luminex-bead equipped with biomarkers and by measuring the fluorescence values of antibodies from each bead by using Luminex. This procedure accesses concentration data for a large number of biomarkers that respond to a particular disease, aided by Luminex. The sample data format is an Excel file as shown in the example in Figure 1.

Test No.	Diagnosis	M1	M2	M3	M4	M5	M6	Menopause	Stage
68	Benign	1.7	121.4	24.6	5.18	60.9	30.9	Y	-
22	Benign	1.7	93.8	33.2	11.88	67	8.1	Y	-
36	Benign	2.8	112.3	26.1	12.23	58.3	11	Y	-
60	Benign	0.3	90.0	80.8	143.29	53.9	11.6	Y	-
148	Benign	0.6	99.6	5.1	2	66.1	11.7	Y	-
101	Benign	0.8	77.4	28.1	5.4	61.7	23	Y	-
96	Cancer	0.6	82.5	6.7	32.36	213.9	363.6	N	1
S6	Cancer	0.9	88.7	22.1	869.12	140.4	300.3	N	1
235	Cancer	6.6	76.1	78.1	3.24	43.7	1587.5	N	1
269	Cancer	5.8	53.7	39.4	9.18	290.6	2248.0	N	4
2	Cancer	17.2	89.8	17.9	2	684.4	5382.0	N	4
120	Cancer	0.3	73.5	97.3	4.68	46.4	116.2	N	1
99	Cancer	0.4	96.5	31.8	48.84	261.4	141.4	Y	3

Figure 1. Marker data

Figure 1 shows the format of the file used in the program for multi-marker visualization tools. The data consist of the identification number of the sample, diagnosis result (the presence of cancer), fluorescence value of individual biomarker (or normalized fluorescence value), menopausal status and stages of cancer progression.

3. Algorithms and Systems

In this system, the logistic regression algorithm was used as a method for assessing the accuracy of cancer diagnosis using multiple biomarkers. If an analysis target is divided into two or more groups (multivariate data), logistic regression [10] is a common statistical algorithm that is used in the development of a model for analyzing and predicting the group into which each assessed value can be classified. The logistic regression is similar to normal regression analysis for analysis purposes and procedures. The difference is that logistic regression is used when dependent variables are categorical ones measured on a nominal scale. Unlike discriminant analysis, the logistic regression has the advantage of being able to put the categorical variable into the predictive variable.

Figure 2 shows the structure of the system for multi-marker analysis developed for biomarker data analysis. The system consists of a module that can read and normalize data, the marker selection module, the module for marker performance verification, and the data visualization module. To compare the range of values and of various units obtained from the markers, the normalization module converts the values obtained into values between 0 and 1. The marker selection module has the function of selecting single or multiple markers. The single and multiple markers use the sample statistic and the logistic regression, respectively, as the module for the performance verification of the marker to confirm the performance. The data visualization module represents the distribution of the given data on a variety of graphs.

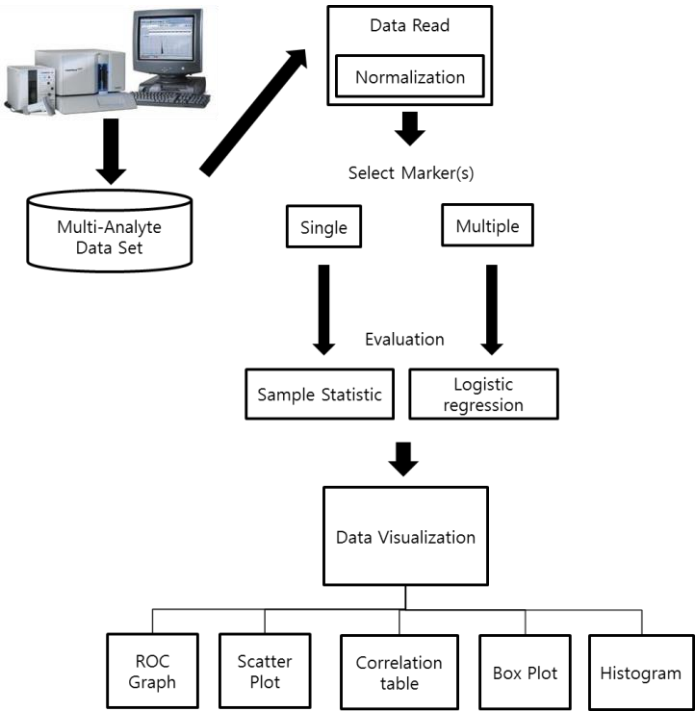

Figure 2. System architecture

4. Performance Evaluation

To evaluate the performance of single or multiple markers selected by the user, descriptive statistics and classification performance are provided. For single or multiple markers in cancer patients and controls, a 5-fold cross validation is used to calculate the values including the area under the curve (AUC), 95% confidence interval (CI), sensitivity, specificity, accuracy, the positive predictive value (PPV) and negative predictive value (NPV). In addition, the mean, standard deviation, median, and maximum and minimum values of single markers are also provided to the user. Analysts can evaluate combinations of multiple markers through the classification performance of the selected multi-marker.

5. Data Visualization

The concentrations of multiple analytes obtained from the Luminex are in the form of numerical data that makes it difficult for analysts to assess the concentration. To remedy this, the values should be visualized on graphs so that they can be analyzed as easily and conveniently as possible. The multi-marker analysis system provides a variety of graphs that enable easy and simple identification of information from the data on the combinations of single markers or multiple markers.

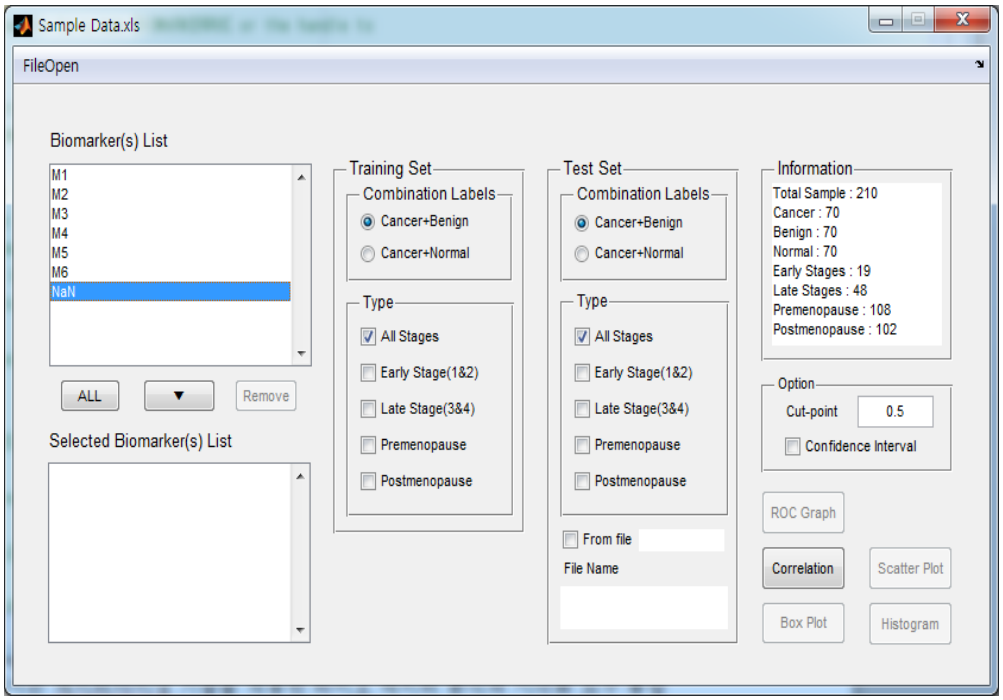

Figure 3. Main user interface

Figure 3 shows the main screen that reads the sample datasets. When the Excel file described in Figure 1 is loaded via the "FileOpen" button, the "Biomarker(s) List" shows the list of markers, and the "Information" window shows basic statistics of samples at the same time. The "Training Set" and the "Test Set" consist of all options for classes targeted by the respective classification models and the information related to other diseases. Cancer-benign and cancer-normal groups are classification options in this system. In addition, the information related to the disease includes stages of cancer progression and menopausal status. Unlike the Training Set window, the Test Set window can load the additional data that can be used as the Test Set separately.

Once you use the Option window, you can set the cut-off point and choose whether you want to include the confidence interval (CI) into the performance result in the ROC graph. It is time consuming to calculate the CI, which makes it an option that is chosen only when absolutely necessary. All five buttons below the Option are for data visualization. In this system, the ROC Graph, Correlation, Scatter Plot, Box Plot and Histogram are provided as visualization options.

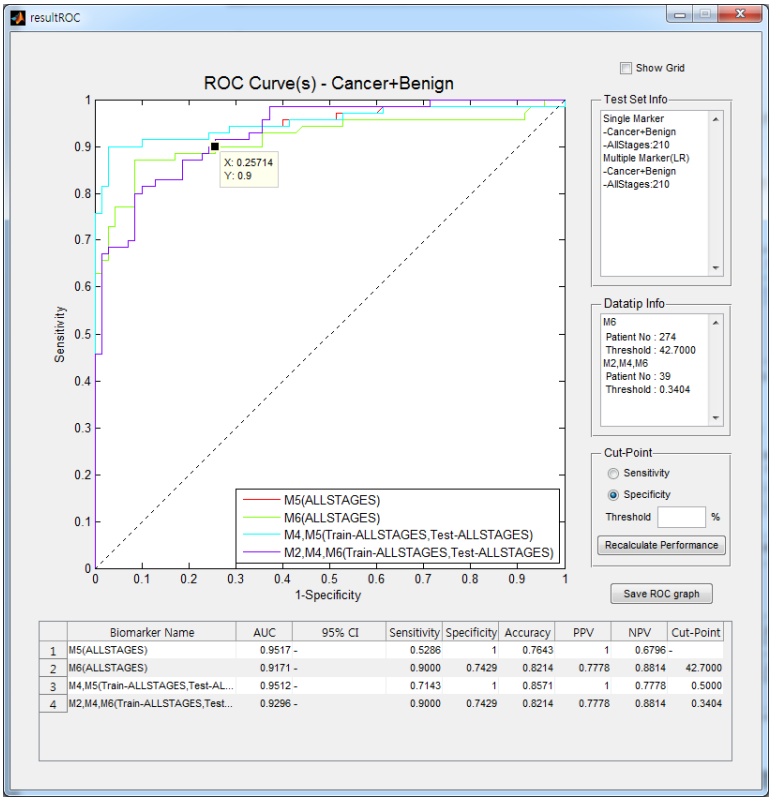

Figure 4. Roc graph & performance table

Figure 4 shows the ROC Graph that expresses the distribution of cancer patients and controls for single or multiple markers. We set several reference points in the distribution of cancer patients and controls, calculated the sensitivity and specificity and then set the x-axis and the y-axis as the false-positive rate (1 - specificity) and as the sensitivity, respectively, to make the ROC graphs. The AUC, the area under the ROC Graph, is a measure that represents the accuracy of disease test. The greater the AUC value, the higher the sensitivity and the lower the false positive rate, which enables more accurate diagnosis. Thus, the ROC Graph is useful for comparing the performance of the selected markers.

The lower part of the window shows the performance information for each marker and all of its combinations. The button for showing the grid in a graph appears on the top right corner. Underneath, there are three windows: Test Set Info that generates information on separate test set, sample number that corresponds to the selected location in the graph, Datatip Info that shows the value or score value, and Cut-Point that recalculates the performance after the selection and input of the cut point. Finally, there is the Save ROC Graph button that can save the ROC Graph as an image file. Clicking on the performance information at the bottom leads to the Raw Data Table that can check the test results of all the samples and save them. Via the "Save as .xls" button of the Raw Data Table in Figure 5, the data are saved as an Excel file.

Figure 6 is a correlation table that analyzes the correlation between the markers. Coefficients are shown using the Pearson`s correlation coefficients [11]. The table can be saved as an Excel file by clicking on the "Save as .xls" button.

The scatter plot in Figure 7 shows the distribution of 3 selected markers. Clicking on "Show DataTip" in the upper right-hand corner leads to the patient information, and "Log" in the "Y Scale" converts the data in a log scale. Unclicking on "Log" displays the data as a

Linear Scale. The graph is stretched upward and downward using the "Upper Bound" and the “Lower Bound", respectively. The selected markers will appear in the "Marker List" part.

The 2D scatter plot in Figure 8 is activated when two or more single markers are selected. Clicking on the scatter plot yields a 2D scatter plot. Figure 9 shows a box plot that indicates the status of cancer patients and the controls, respectively. It shows the minimum, maximum, 2nd quartile, 4th quartile and median values. Figure 10 shows a histogram of the distribution of a marker between the cancer patients and the controls, expressed in the bar and curve formats.

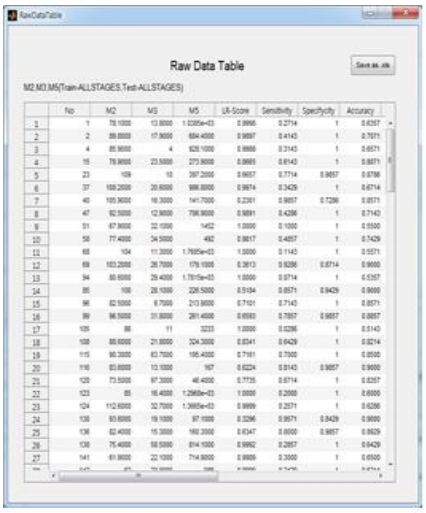

Figure 5. ROC Graph

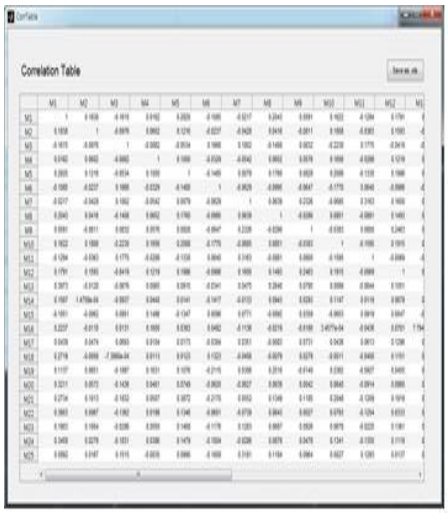

Figure 6. Correlation Table

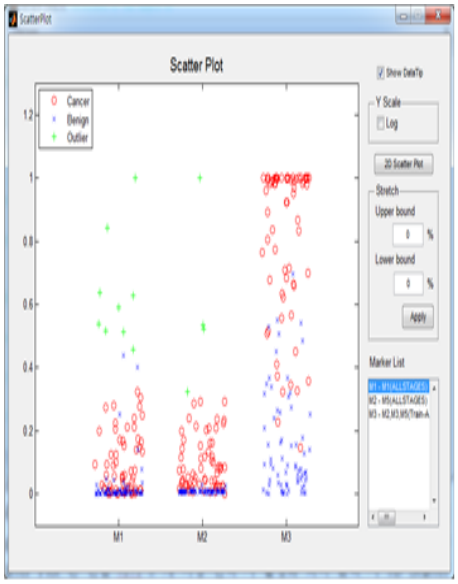

Figure 7. Scatter Plot

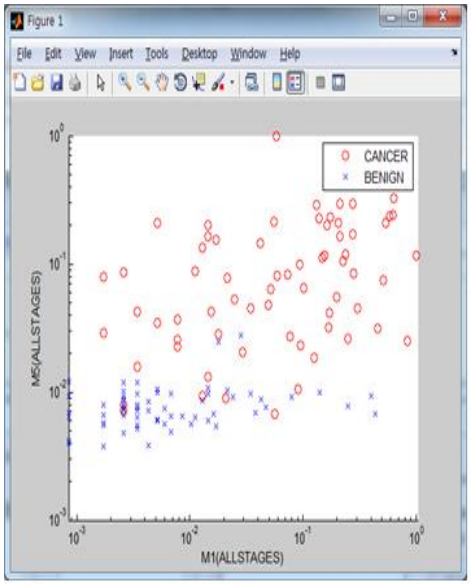

Figure 8. 2D Scatter Plot

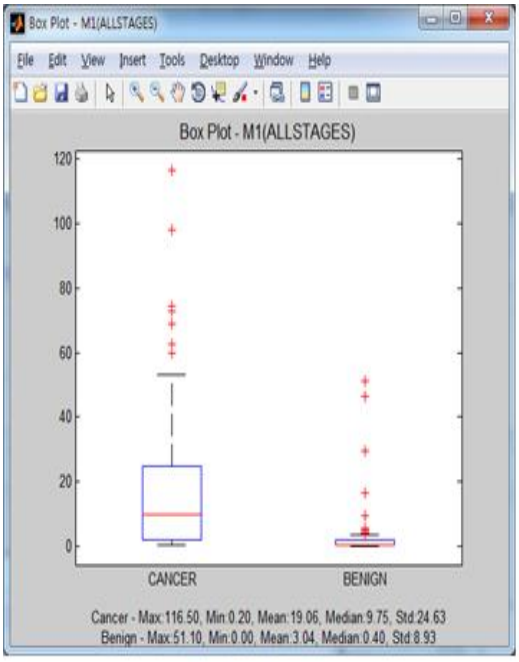

Figure 9. Box Plot

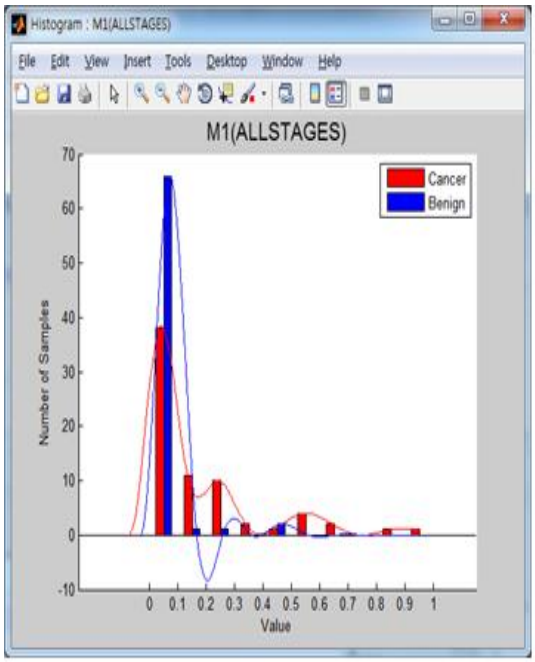

Figure 10. Histogram

6. Conclusion

Recently, a variety of methods has been investigated that enable the early detection of cancer or other specific diseases in the field of diagnostic medical testing. To monitor the effectiveness of early detection and the treatment of diseases, biomarkers are used. Biomarkers should be able to discern between the normal condition and the specific disease through the analysis of data obtained from multiple markers.

In this paper, we developed an integrated system for bio-data analysis and a visualization tool by applying data mining technology. The proposed system provides clinicians with user-friendly tools for analyzing data compared to existing systems. In this system, data extracted by multiple biomarkers tailored to the diagnosis of specific diseases are measured using Luminex equipment and converted into easily analyzable formats to be used for visualization, which facilitates the identification of data patterns, thus enabling simple automatic or visual discernment between normal and diseased conditions. Clinicians will be able to make more accurate judgments in the diagnosis of disease by using various simple tools utilizing the visualization, marker selection, and performance confirmation functions embedded in the system.

Acknowledgements

The research was supported by the Ministry of Trade, Industry, and Energy (MOTIE), Korea Institute for Advancement of Technology (KIAT) and Gangwon Institute for Regional Program Evaluation (GWIRPE) through the Leading Industry Development Project for Economic Region.

References

[1] R Project, http://www.r-project.org.
[2] GraphPad Prism, http://www.graphpad.com.

[3] SPSS, http://www.spss.com.
[4] Hendriks, S. Bart and C. W. Espelin, "DataPflex: a MATLAB-based tool for the manipulation and visualization of multidimensional datasets", Bioinformatics, vol. 26, no. 3, **(2010)**, pp. 432-433.
[5] PubMed, www.ncbi.nlm.nih.gov/pubmed.
[6] Amonkar, D. Suraj, *et al.*, "Development and preliminary evaluation of a multivariate index assay for ovarian cancer", PLoS One, vol. 4, no. 2, **(2009)**, pp. e4599.
[7] Y. Kim, M. Jang, C. Park, H. Song and J. Kim, "Exploring Multiple Biomarker Combination by Logistic Regression for Early Screening of Ovarian Cancer", International Journal of Bio-Science and Bio-Technology, vol. 5, no. 2, (**2013**), pp. 67-75.
[8] M. Jang, Y. Kim, C. Park, H. Song and J. Kim, "Integration of Menopausal Information into the Multiple Biomarker Diagnosis for Early Diagnosis of Ovarian Cancer", International Journal of Bio-Science and Bio-Technology, vol. 5, no. 4, **(2013)**, pp. 215-222.
[9] H. Song, S. Ko, J. Kim, C. Park and Y. Kim, "Looking for the Optimal Machine Learning Algorithm for the Ovarian Cancer Screening", International Journal of Bio-Science and Bio-Technology, vol. 5, no. 2, (**2013**), pp. 41-48.
[10] J. P. Hoffmann, "Generalized linear models", MA: Allyn & Bacon, Boston, **(2003)**.
[11] J. Benesty, J. Chen, Y. Huang and I. Cohen, "In Noise reduction in speech processing", Springer Berlin Heidelberg, **(2009),** pp. 37-39.

Authors

Ki-Seok Cheong

He is now the B.E. student in Computer Engineering of Hallym University. His recent interests focus on biomedical system and bioinformatics.

Hye-Jeong Song

She received the Ph.D. degree in Computer Engineering from Hallym University. She is a Professor in Department of Ubiquitous Computing, Hallym University. Her recent interests focus on biomedical system and bioinformatics

Chan-Young Park

He received the B.S. and the M.S. from Seoul National University and the Ph.D. degree from Korea Advanced Institute of Science and Technology in 1995. From 1991 to 1999, he worked at Samsung Electronics. He is currently a Professor in the Department of Ubiquitous Computing of Hallym University, Korea. His research interests are in Bio-IT convergence, Intelligent Transportation System and sensor networks.

Jong-Dae Kim

He received the M.S. and the Ph.D. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology, Seoul, Korea, in 1984 and 1990, respectively. He worked for Samsung Electronics from 1988 to 2000 as an electrical engineer. He is a Professor in Department of Ubiquitous Computing, Hallym University. His recent interests focus on biomedical system and bioinformatics.

Yu-Seop Kim

He received the Ph.D. degree in Computer Engineering from Seoul National University. He is currently a Professor in the Department of Ubiquitous Computing at Hallym University, South Korea. His research interests are in the areas of bioinformatics, computational intelligence and natural language processing.