Applied Mathematics & Information Sciences
An International Journal

@ 2012 NSP
Natural Sciences Publishing Cor.

MATLAB-Based Multi-Marker Data Analysis System for Early Detection of Ovarian Cancer

Yu-Seop Kim[1,2] , Hye-Jeong Song[1,2] , Seung-il Kim[3] , Jong-Dae Kim[1,2] , Chan-Young Park[1,2] , Erika Lee[4] and Jongwon Kim[5]

[1] Department of Ubiquitous Computing, Hallym University, Chuncheon, South Korea
[2] Bio-IT Research Center, Hallym University, Chuncheon, South Korea
[3] Department of Computer Engineering, Hallym University, Chuncheon, South Korea
[4] C&Tech, Ahngook Pharmaceutical Co., Ltd, Seoul, South Korea
[5] BioMedLab, Seoul, South Korea
Corresponding author: Email: hjsong@hallym.ac.kr

Received June 22, 2010; Revised March 21, 2011; Accepted 11 June 2011
Published online: 1 January 2012

Abstract: Recent research has focused on testing multiple biomarkers for the early detection of cancer to overcome the intrinsic limitations of individual markers. An analysis of multiple biomarker data requires data-mining tools and a statistical suite to investigate individual and combined characteristics. The existing general purpose systems lack compact functionality and an intuitive graphical user interface that allows the user to easily identify characteristics that can distinguish among target disease states. This paper presents an analytical system with statistical and data-mining functions for the multi-marker detection of ovarian cancer. The system was written in MATLAB, because this platform provided a high degree of functionality and facilitated the flow of information among the appropriate analyses. MATLAB also provided several visualization choices and an intuitive graphical user interface. The proposed system was tested with serum level data measured with a high-throughput, multiplex, bead-based, immunoassay platform (Luminex). The system was designed to offer the user the choice of manually selecting marker combinations or automatically selecting marker combinations. It was constructed of modular analyses that evaluate selected markers for classification efficiency based on the best performance. The results can be visualized with dot and box plots for individual markers and 2D and 3D scatter plots of the combined characteristics of multiple markers. This system offers a user-friendly method for evaluating multiple markers and facilitating clinical diagnoses.

Keywords: Biomarkers, Data-Mining Tools, Marker Selection, Multiplex Immunoassay

1 Introduction

Medical diagnosis and biological data analysis requires a data mining tool appropriate for statistical evaluations and graphical display [1-2]. The R project, SPSS, and GraphPad PRISM are commonly used, general data mining and data analysis systems, but they require statistical expertise. Thus, many clinical scientists leave the analysis to statisticians, because most general data analysis systems demand statistical knowledge and expertise with statistical tools. However, this practice may fail to take advantage of the clinical scientist's empirical knowledge, which may be crucial for the diagnosis. Consequently, a data mining system is needed that is intuitive to the clinical scientist, can produce useful knowledge, and provides data visualization to facilitate rapid, accurate interpretation [1-5].

Biomarkers are used for the early detection of disease and for monitoring the curative value of treatments. They provide an indirect screening method for detecting unique, biological-molecular processes that change according to the disease state. Recently, there has been an increase in the development of research methods and tools for identifying biomarkers [6].

Unfortunately, the use of individual markers has lacked sufficient sensitivity or specificity for detection of early-stage disease [8]. Recent studies have shown that a combination of serum tumor markers could outperform individual markers [7-9]. Consequently, various techniques are in development for combining biomarkers to detect ovarian cancer at an incipient stage and to investigate cancer selectivity [7-9].

This combined biomarker approach requires multiplex assays. Multiplex proteome analyses, like Luminex or ELISA, can provide high sensitivity in serum-based protein diagnosis and analysis. In particular, Luminex has been widely used in diagnoses due to rapid processing and high sensitivity. It also enables the concurrent analysis of one sample with many analytes when used with multi analyte profiling (MAP), bead-based technology [10].

This paper presents a statistical, data mining analysis system for identifying multi-markers for the early detection of ovarian cancer. The analysis program was written with MATLAB, which offers multiple statistic and graphical functions [3-5].

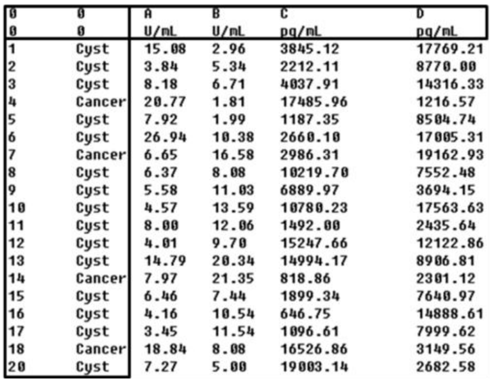
2 Data Set

Serum was collected from 27 patients with ovarian cancer, 38 individuals with cysts, and 31 healthy women. We measured 21 serum biomarkers related to ovarian cancer with the Luminex assay [8].

Figure 1 illustrates the data file format of the multi-marker analysis system. Only the first 25/96 samples are shown and six markers are listed (A-F).

3 Multi-Marker Data Analysis System

Figure 2 shows the structure of the multi-marker analysis system developed for biomarker data analysis. It consists of several modules that operate sequentially. The data is normalized, markers are selected, marker performance is evaluated, and the data is visualized.



0	0	A	B	C	D
0	0	U/mL	U/mL	pg/mL	pg/mL
1	Cyst	15.08	2.96	3845.12	17769.21
2	Cyst	3.84	5.34	2212.11	8770.00
3	Cyst	8.18	6.71	4037.91	14316.33
4	Cancer	20.77	1.81	17485.96	1216.57
5	Cyst	7.92	1.99	1187.35	8504.74
6	Cyst	26.94	10.38	2660.10	17005.31
7	Cancer	6.65	16.58	2986.31	19162.93
8	Cyst	6.37	8.08	10219.70	7552.48
9	Cyst	5.58	11.03	6889.97	3694.15
10	Cyst	4.57	13.59	10780.23	17563.63
11	Cyst	8.00	12.06	1492.00	2435.64
12	Cyst	4.01	9.70	15247.66	12122.86
13	Cyst	14.79	20.34	14994.17	8906.81
14	Cancer	7.97	21.35	818.86	2301.12
15	Cyst	6.46	7.44	1899.34	7640.97
16	Cyst	4.16	10.54	646.75	14888.61
17	Cyst	3.45	11.54	1096.61	7999.62
18	Cancer	18.84	8.08	16526.86	3149.56
20	Cyst	7.27	5.00	19003.14	2682.58

Figure 1: Data file format of serum sample measurements. Samples (1-25) are classified by disorder (cyst, cancer), and the value of each biomarker (A-F) is shown.

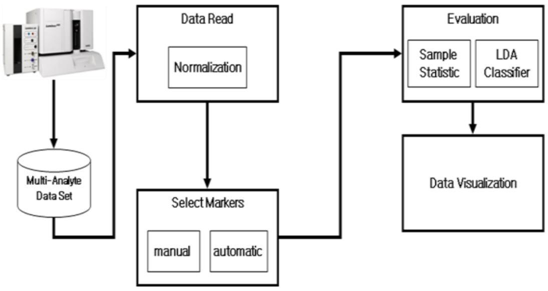


Figure 2: Structure of Multiple Marker Analysis System. Arrows show the sequence of analysis as data is passed to each module. The functions are indicated inside the modules, and the toolboxes are shown in grey below the corresponding modules.

The normalization module carries out the min-max normalization to fit the range of values spanned by the markers. The Select Markers module allows the user to choose manual selection or automatic selection. The Evaluation module calculates the sensitivity, specificity, and accuracy of the selected markers. The Data Visualization module includes plotting in various dimensions.

3.1 Select Marker Module

Figure 3 shows the structure of the Select Markers module. Marker selection allows the user to select markers manually, based on empirical knowledge, or to select markers automatically, according to an algorithm. The Rank Markers toolbox selects several potential markers by estimating the resolution achievable for distinguishing differences between the patients with cancer and the control group.

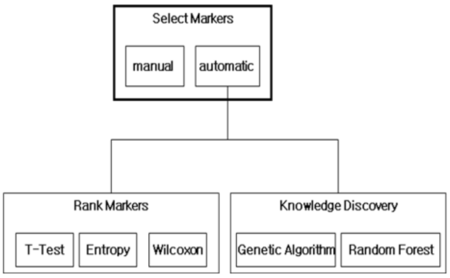


Figure 3: Structure of Select Markers Module.

The user can choose to rank markers with the t-test, the Wilcoxon rank-sum test, or the entropy test. The t-test analyzes the average values for two independent sample groups, and determines whether there is a significant difference. In other words, it tests whether the marker levels observed in patients with cancer is significantly different from that observed in the control group. The Wilcoxon rank-sum test is a non-parametric test used when it is uncertain whether the data is distributed normally. The entropy test, a statistical value of uncertainty, is a method for deducing the resolution of a specific marker by combining the probabilities of all individual markers.

The Knowledge Discovery selects multi marker set that can classify cancer based on multiple characteristics in the data. This is a data mining method that finds the optimal subset of randomly combined markers by measuring the discernment of each subset [1].

The Genetic Algorithm [11] is a random search algorithm that imitates the process of natural evolution to find solutions that optimize identification. It starts with randomly combined markers and evolves these markers in each cycle by transposition and mutation until the best combination is achieved. The genetic algorithm uses information gained during the previous search to guide the subsequent iterations by adjusting the standard of decision. In this work, for example, the size of the population, number of generations, crossover rate, and mutation ratio was 50, 25, 80%, and 0.1%, respectively. The parameters were chosen because the former researchers varied such parameters of the GA to determine the best parameter value, but no significant results were obtained. The population was created by randomly combining the markers with the same number of inhabitants from the top of the rank of the whole population.

The Random Forest [12] is a classification method that creates several decision trees with sets of randomly extracted samples and determines the optimum class by weighted voting. Random Forest can make rather precise classifications with relatively few clinical samples, because it uses the bootstrap method to generate training data for the learning algorithm. Also, it shows high accuracy when new data (separate from the training data) is entered, because it has the capacity to embrace a variety of patterns. The latter feature arises from the fact that many decision trees are generated with randomly extracted training data during the training process.

3.2 Evaluation

The Evaluation module evaluates the performance of manually or automatically selected multi-markers by calculating the technical statistics and deriving a classification efficiency. It calculates the average, standard deviation, median, maximum, and minimum values for the individual markers in the patients with cancer and the control group. A linear discriminant analysis (LDA) is used to test the efficiency of the multi-marker for correctly classifying the patients with cancer. A 5-fold cross validation is used to calculate the accuracy rate, the error rate, sensitivity, and specificity of the multi-marker.

From the classification efficiencies of the selected markers, the analyzer is able to estimate the best combination of multiple markers.

3.3 Data Visualization

As data understanding requires taking a closer look at the data, visualization techniques will be great helpful for the analysis of individual and combined attributes.

The appropriate plots of the data can provide valuable information [1].

The proposed multi-marker analysis system provides various graphs to facilitate examination of individual marker data or combined multi-marker data. Figure 4 shows the structure of the Data Visualization module.

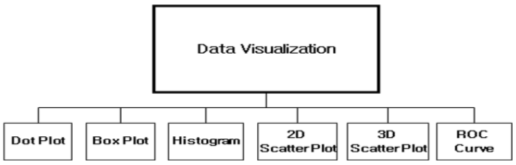
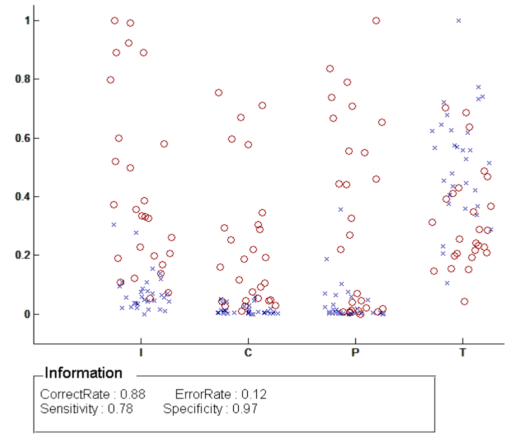


Figure 4: Structure of Data Visualization Module

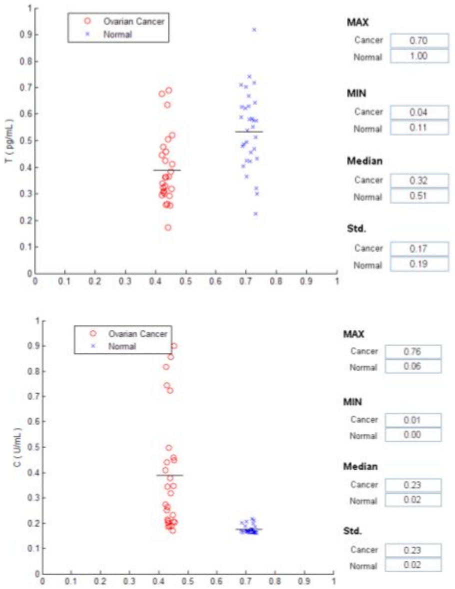
Figure 5 shows a dot plot of the distribution of selected markers in the patients with cancer and the control group.



CorrectRate : 0.88 ErrorRate : 0.12
Sensitivity : 0.78 Specificity : 0.97

Figure 5: Dot plots of selected markers. This example used the t-test ranking method.

The dot plot (Figure 6) shows the comparison of the distribution of two groups (cancer and control) and horizontal lines indicate mean values.

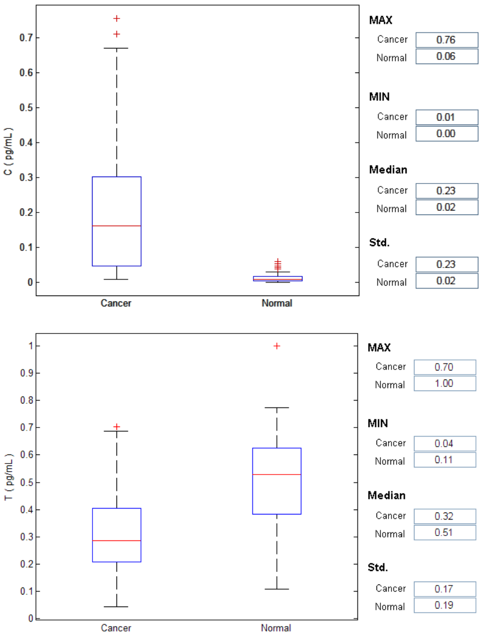


MAX
Cancer 0.70
Normal 1.00
MIN
Cancer 0.04
Normal 0.11
Median
Cancer 0.32
Normal 0.51
Std.
Cancer 0.17
Normal 0.19

MAX
Cancer 0.76
Normal 0.06
MIN
Cancer 0.01
Normal 0.00
Median
Cancer 0.23
Normal 0.02
Std.
Cancer 0.23
Normal 0.02

Figure 6: Two dot plots for the serum values of marker (T, C)

The box plot (Figure 7) shows the minimum, maximum, quartile, and median values of the individual markers in patients with cancer and controls.



MAX
Cancer 0.76
Normal 0.06
MIN
Cancer 0.01
Normal 0.00
Median
Cancer 0.23
Normal 0.02
Std.
Cancer 0.23
Normal 0.02

MAX
Cancer 0.70
Normal 1.00
MIN
Cancer 0.04
Normal 0.11
Median
Cancer 0.32
Normal 0.51
Std.
Cancer 0.17
Normal 0.19

Figure 7: Two box plots for the serum values of marker (C, T)

The histogram (Figure 8) shows the frequency distribution for the serum values of the individual markers.

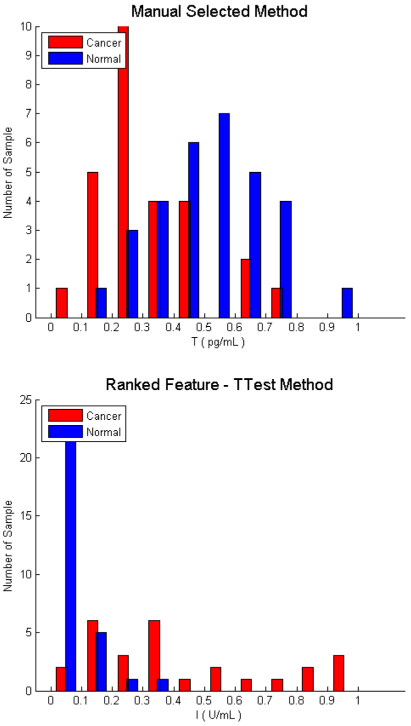


Figure 8: Histogram shows the frequency distribution for the serum values of marker (T, I)

2D scatter plot (Figure 9) displays where two potential biomarkers are plotted against each other. 2D scatter plot compares the distribution of serum values of two markers.

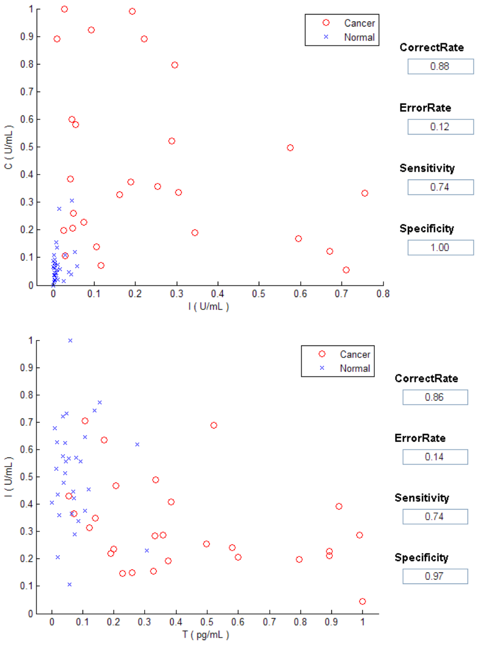


Figure 9: 2D Scatter Plot compares the distribution and performance of two potential biomarkers (I-C, T-I)

Figure 10 shows a 3D scatter plot. It compares the distribution of serum values of three potential biomarkers.

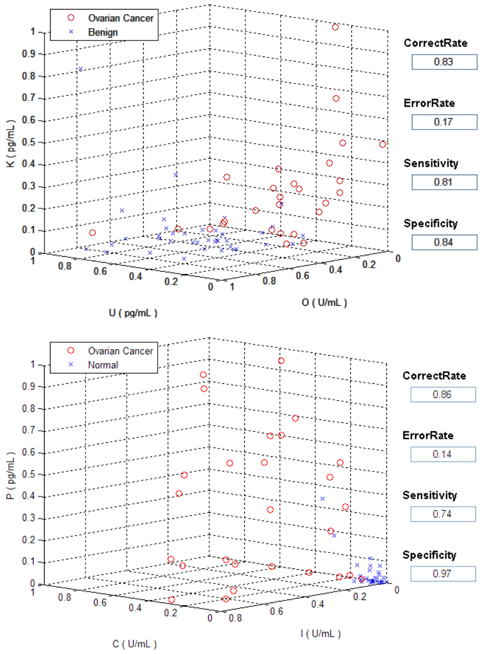


Figure 10: 3D Scatter plot compares the distribution and performance of three potential biomarkers (K-U-O, P-C-I).

Figure 11 shows the receiver operating characteristic (ROC) curves for different markers.

The ROC curve is created by first calculating the sensitivity and specificity from the standard dot plots for patients with cancer and control individuals. The x-axis of the ROC plot represents the false-positive rate (1-specificity), and the y-axis represents the sensitivity. The area under the curve (AUC) indicates the accuracy of the disease diagnosis method. As the sensitivity increases, the false-negative rate also increases. Therefore, the optimum is the smallest value on the x-axis combined with the largest value on the y-axis. Therefore, a high AUC represents a high accuracy in the diagnosis. These features make the ROC curve a useful method for comparing the performance of several markers.

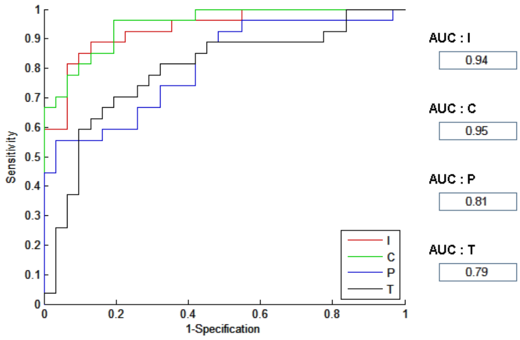


Figure 11: ROC Curves compare the accuracy of four potential serum markers (I, C, P, T). Markers with a high area under the curve (AUC) provide a highly accurate diagnosis.

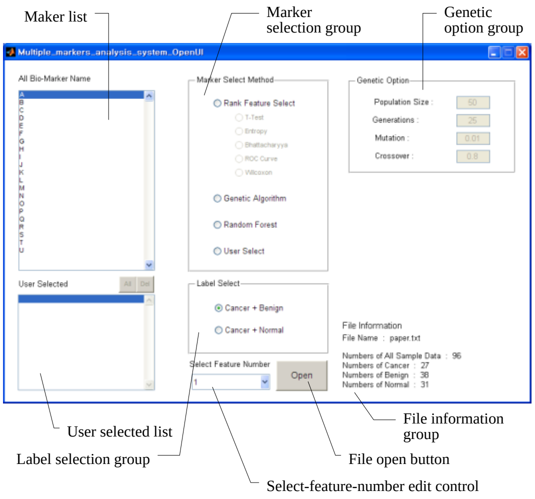
4 User Interface Issue

Logic coverage criteria are mainly used for formal specification-based testing [3]. It generates test cases by analyzing the predicates and literal truth value relationship. The formal specification is

The user interface to the whole functions is another important issue to be resolved for the intuitive usage of the clinical scientists. It should be better if the user interface is graphical and is also integrated into a single widow-based application. MATLAB has built-in functions useful for building customized graphical user interfaces. This is one of the main reasons why MATLAB has been selected in this work.

Figure 12 shows the graphical user interface (GUI) that has been developed in this study. In this GUI, some widgets are grouped by using group-controls. For examples, marker selection methods are implemented with a group of radio-button controls because they are independent of each other. The widgets of the genetic option group are activated only when the genetic algorithm is selected in the marker selection group. The overall design has been continuously modified to accommodate the domain user requirement. For example, the user selected list, which lists the already selected markers, is added by the clinical scientists after they have experienced this system. They want to check the markers that they have chosen. The functionalities of the widgets or the group of widgets are follows:

- Marker list: The list of markers in the lot data that is opened with 'File open button'
- Marker selection group: A group of radio buttons which initiate a selection method. The 'Rank feature select' button has a subgroup of radio buttons which choose one of the rank feature selection methods.
- Genetic option group: This group of edit boxes is activated only when the 'Genetic algorithm' button in the marker selection group is selected to input the genetic algorithm parameters.
- User selected list: The list shows the already chosen markers.
- Label selection group: The group of ratio buttons to select the pair of the patient groups to be classified.
- Select-feature-number edit control: The drop-down list box to define the number of the markers to be investigated.
- File open button: To open the lot data file to be analyzed.
- File information group: The group of read-only edit boxes to summarize the opened lot data.



Population Size : 50
Generations : 25
Mutation : 0.01
Crossover : 0.8

● Cancer + Benign
○ Cancer + Normal

Select Feature Number: 1

File Information
File Name : paper.txt
Numbers of All Sample Data : 96
Numbers of Cancer : 27
Numbers of Benign : 38
Numbers of Normal : 31

Figure 12: Example of the integrated user interface

5 Conclusion

Currently, researchers are actively developing methods for early detection of specific diseases in the field of medical diagnosis. Approaches for the early detection of disease and for monitoring the curative value of a treatment must be able to determine the disease status based on the values of multiple markers.

This paper describes an analysis system with an intuitive user interface for interpreting multi-marker data. It incorporates data mining technology that detects implicit patterns in the data and generates useful knowledge. As an example, we used data acquired with Luminex for the early detection of ovarian cancer. This system simplified analysis with side-by-side visualizations of the data and the analysis results at each step in the evaluation process.

This multi-marker analysis system was designed to facilitate data analysis for clinical scientists, who may have less statistical knowledge, but more interpretative capabilities than the statistician.

Thus, this analysis system is likely to lead to more accurate diagnoses of diseases.

Acknowledgements

This work was supported by the Gangwon Economics Leading Industry Development Project, which was supported by the Industry Researcher with financing from the Government (Ministry of Knowledge Economy), 2011. (No. 2009-G-1-A-Y0-C-14)

References

[1] M.R. Berthold, C. Borgelt, F. Hoppner and F. Klawonn, Guide to Intelligent Data Analysis (Springer Verlag, London, 2010).

[2] X. Chen,Y. Ye, G. Williams and X. Xu, Lecture Notes in Computer Science. 4819, 3 (2007).

[3] B. S. Hendriks and C. W. Espelin, Bioinformatics. 26, 4 (2010).

[4] perClass software site : http://perclass.com

[5] R. J. Saez, Bioinformatics. 24, 840 (2008).

[6] PubMed site : http://www.ncbi.nlm.nih.gov/pubmed

[7] Z. Zhang and D. W. Chan, Cancer Epidemiol. Biomarkers Prev. 19, 2995 (2010).

[8] Z. Yurkovetsky, S. Skates, A. Lomakin, B. Nolen, T. Pulsipher, F. Modugno, J. Marks, A. Godwin, E. Gorelik, I. Jacobs, U. Menon, K. Lu, D. Badgwell, R. C. Bast, Jr and A. E. Lokshin, Journal of Clinical Oncology. 28,

2159 (2010).

[9] S. D. Amonkar, G. P. Bertenshaw, T. H. Chen, K. J. Bergstrom, J. Zhao, P. Seshaiah, P. Yip and B. C. Mansfield, PLoS ONE. 4, e4599 (2009).

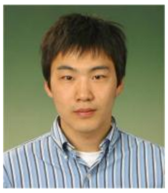
[10] M. Hartmann, J. Roeraade, D. Stoll, M. F. Templin and T. O. Joos, Anal Bioanal Chem. 393, 5 (2009) .

[11] D.E. Goldberg, Genetic Algorithms in Search Optimization & Machine Learning (Addison-Wesley, 1989).

[12] L. Breiman, Random Forests, Machine Learning. 45, 5 (2004).

Yu-Seop Kim received the PhD 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.

Hye-Jeong Song received the PhD degree in degree in Computer Engineering from Hallym University. He is a Professor in Department of Ubiquitous Computing, Hallym University. His recent interests focus on biomedical system and bioinformatics



Seung-il Kim received the BS degree in Computer Engineering from Hallym University. He study for a master's degree in Hallym University. His recent interests focus on biomedical system.

Jong-Dae Kim received the PhD 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.

Chan-Young Park received a B.S and M.S. from Seoul National Univ. and a 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.

Erika Lee received the PhD degree in Biotechnology from Yonsei university in Korea and took the executive MBA program of Helsinki School of Economics in Finland. With 10 years of experience in the technology analysis and overseas tech-transfer, she is currently a Deputy Director at New Business Department at Ahngook pharmaceutical company in Korea. Her research and business interests are in the areas of biomarker discovery and cancer diagnostic products.

Jongwon Kim received the BS and MS degree in Physics from Seoul National University in 1985 and 1987 respectively, and the PhD degrees in Biomedical Engineering from Seoul National University, Seoul, Korea, in 1992. During 1986-1996, he stayed in Seoul National University Hospital as research member. He is now the president of Biomedlab Co. from 1994 in Korea. His research interests are in the areas of biotechnology and biomedical system.