An Efficient Functional Analysis Method for Micro-array Data Using Gene Ontology

Dong-wan Hong*, Jong-keun Lee*, Sung-soo Park*, Sang-kyoon Hong*, and Jee-hee Yoon*

Abstract: Microarray data includes tens of thousands of gene expressions simultaneously, so it can be effectively used in identifying the phenotypes of diseases. However, the retrieval of functional information from a large corpus of gene expression data is still a time-consuming task. In this paper, we propose an efficient method for identifying functional categories of differentially expressed genes from a micro-array experiment by using Gene Ontology (GO). Our method is as follows: (1) The expression data set is first filtered to include only genes with mean expression values that differ by at least 3-fold between the two groups. (2) The genes are then ranked based on the t-statistics. The 100 most highly ranked genes are selected as informative genes. (3) The t-value of each informative gene is imposed as a score on the associated GO terms. High-scoring GO terms are then listed with their associated genes and represent the functional category information of the micro-array experiment. A system called HMDA (Hallym Micro-array Data analysis) is implemented on publicly available micro-array data sets and validated. Our results were also compared with the original analysis.

Keywords: *Micro-array data, Functional analysis, Gene Ontology, Informative genes.*

1. Introduction

By April 2003, the sequences of the 3 billion base pairs that make up the human DNA had been completely determined. However, the derived information about genomic sequences is insufficient to investigate the human life phenomenon, which is exceptionally complicated. Functional genomics, which focuses on investigating the function of the analyzed genome, has become an important research topic. DNA micro-array is one of the main experimental/analyzing tools used in functional genomics. Micro-array experiments enable scientists to obtain a tremendous amount of gene expression data at one time, so they are effectively used in identifying the phenotypes of diseases. However, the retrieval of functional information from a huge body of gene expression data remains a time-consuming task.

In this paper, we propose an efficient functional analysis system, HMDA (Hallym Microarray Data analysis), for micro-array data. It is based on a statistical analysis model and identifies the functional categories of differentially expressed genes from a micro-array experiment by using Gene Ontology (GO) [1]. HMDA correlates the micro-array data as follows: it first extracts informative genes from the micro-array data. It uses a threefold filtering method and t-statistics, and selects the 100 highest-ranked genes. The threefold filtering method involves the assumption that the larger expression differences in micro-array experiments imply only the meaningful genes. GO is then used to identify the functional categories of selected informative genes. A matching process between gene symbols and Probe Set IDs is required in order to search the GO. We have imported the MAD_SCHEMA provided by the NBN (National Bioinformatics Network) [2].

Our system has been validated on publicly available micro-array data sets. A dataset from the micro-array experiments described in reference [3] is used, and our functionally analyzed results were compared with the original analysis.

2. Related Work

2.1 Gene Ontology

As data stored in each bio database is not independent but related to each other, a service which can show related data together is needed. For example, the proteins essential for maintaining life are not restricted to a single species but exist in multiple species. Gene Ontology (GO) was developed for consistent descriptions of gene products in different databases. GO is a large knowledge structure consisting of three branches: biological process, molecular function, and cellular component. Each branch is organized as a taxonomy of nodes (terms) representing different categories of genomic characteristics, connected by either is-a or has-part links. Recently, GO has become a typical

Manuscript received November 24, 2006; accepted March 13, 2007.
This work was supported by grant by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MOST)(No. R01-2006-000-11106-0).

Corresponding Author: Dongwan Hong
* Department of Computer Engineering, Hallym University, Okcheon-Dong, Chuncheon, 200-702, Korea
(dwhong, jeikei, sspark, kyoons, jhyoon)@hallym.ac.kr

tool for gene expression data analysis.

2.2 Previous Systems

Though the early systems were only used for simple analyses of expression data [4,5,6], some tools have recently become available for the functional analysis of genes in broader areas [7,8,9]. These systems usually use GO as the biologically interoperable profile, and "GO-Mapper [7]" and "Onto-Express [8]" are the representative systems. These systems, however, are mainly focused on the analysis of the sample data of each experiment and have a number of limitations [10].

3. Method

The HMDA (Hallym Microarray Data Analysis) system consists of four functional modules: the "threefold filtering module", the "t-test and ranking module", the "GO mapping module", and the "gene knowledge-base module". Figure 1 shows the data analysis process of our system.

Each module works as follows:

(1) The threefold filtering module: After importing the micro-array data stored in a Microsoft Excel format, it filters only those genes with mean expression values that differ by at least threefold between the two groups (refer to formula 1).

$$\log_{10}\frac{average_{\text{control group}}}{average_{\text{test group}}} \geq \log_{10}3 \qquad \text{(formula 1)}$$

(2) The t-test and ranking module: The t-test is used to verify whether the means of two groups are statistically different from each other. To consider the distribution of data in each group, the t-value is calculated by the formula 2. Next, genes were ranked based on the t-value.' In this system, the top 100-ranked genes were selected as the informative genes.

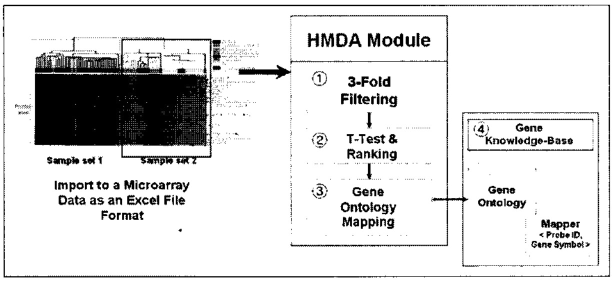


Fig. 1. Data analysis process

$$t-value = \frac{|average_{\text{control group}} - average_{\text{test group}}|}{\sqrt{\frac{\text{var}_{\text{control group}}}{n_{\text{control group}}} + \frac{\text{var}_{\text{test group}}}{n_{\text{test group}}}}} \qquad \text{(formula 2)}$$

(3) The gene ontology mapping module: Informative genes are linked to GO terms and their functional information is retrieved as follows. The t-value of each informative gene is imposed as a score on the associated GO terms and all of its parental terms. After linking all genes to the GO terms, we calculate the significant value (SV) which represents a normalized score of a GO term. SV is calculated by formula 3.

$$sv = \frac{\dfrac{\sum_{j=1}^{m} t-value(gene_j)}{m}}{\dfrac{\sum_{i=1}^{n} t-value(gene_i)}{n}} \qquad \text{(formula 3)}$$

Here, n and m denote the total number of informative genes and the number of informative genes related to the specific GO term, respectively.

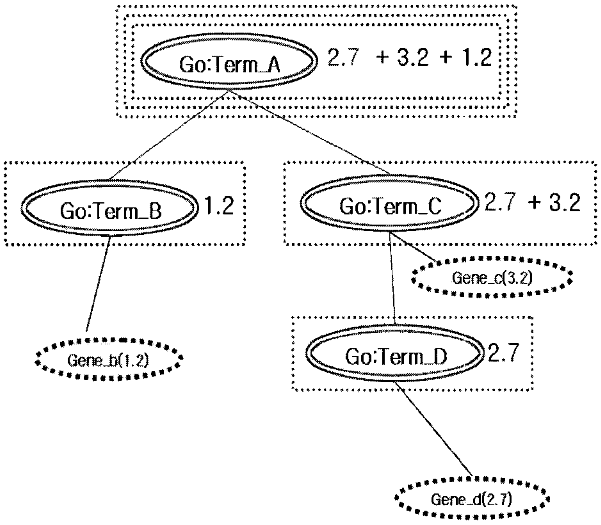


Fig. 2. Example of GO term scoring

Table 1. Significant values of GO terms

	Total Score	SV
GO: Term_A	7.1	1
GO: Term_B	1.2	0.508
GO: Term_C	5.9	1.25
GO: Term_D	2.7	1.144

Figure 2 shows an example. We assume that the three informative genes of Gene_b, Gene_c, and Gene_d are selected and have t-values of 1.2, 3.2, and 2.7 respectively. Each of these t-values is added as a score to the associated GO terms and all of its parental terms. The figure shows the total scores of each GO term. Each SV of the GO terms is given in Table 1. In this case, "GO:Term_C" with an SV of 1.25 and "GO:Term_D" with an SV of 1.144 can be listed as the representative functional categories.

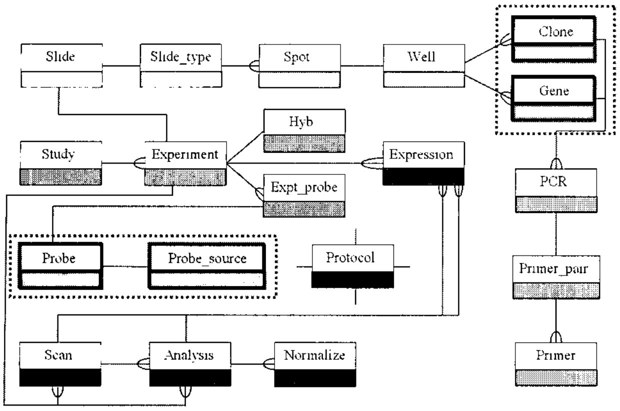


Fig. 3. MAD (Micro-Array Data in NBN) schema

(4) The gene knowledge-base module: For the purpose of matching between the gene symbol of GO and the "Probe Set ID" of micro-array data, we imported the MAD schema offered by the NBN of the TIGR (The Institute for Genomic Research) and established a Matching Information Knowledge-Base. Figure 3 shows the conceptual level of the MAD schema. In this system, we imported the "Clone", "Gene", "Prove", and "Probe_source" tables.

4. Results and Analysis

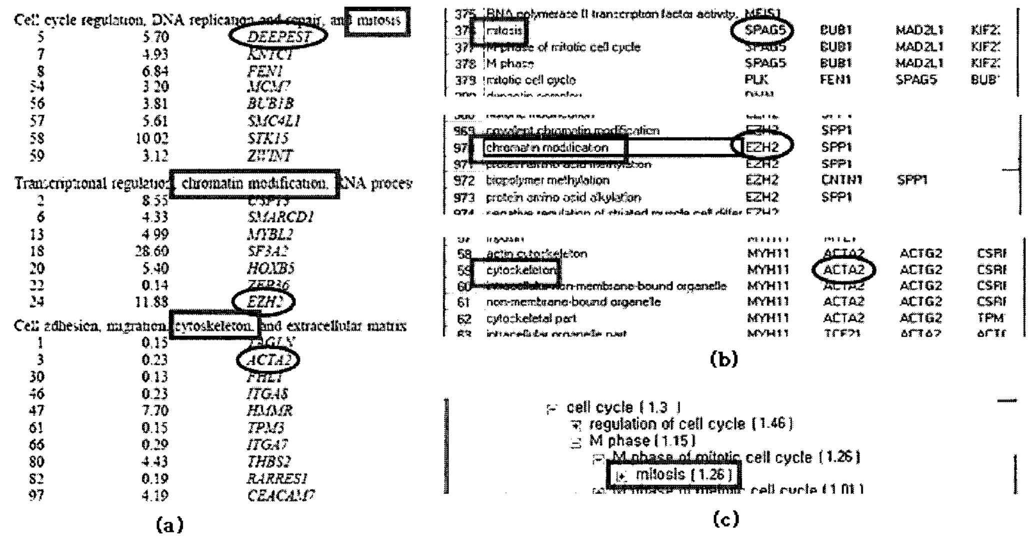
In order to show the validity of the proposed method, our results were cross-checked with the analysis results described in reference [3]. The micro-array experiments in reference [3] were carried out using Affymetrix U95 human gene arrays with 12,559 probe sets. They performed a comprehensive gene expression analysis of prostate cancer to identify genes with strong and uniform differential expression between non-recurrent primary prostate cancers and metastatic prostate cancers. 14 non-recurrent primary prostate samples and 9 metastatic prostate cancers were evaluated.

Figure 4 shows a part of the comparison results. Figure 4-(a) represents the experimental results of reference [3]. The functional classification of this paper was conducted by the expert's manual work. Figures 4-(b) and 4-(c) represent our results. Figures 4-(a) and 4-(b) show the extracted functional categories based on the differentially expressed genes. Here, circles and squares represent the gene symbols and functional categories, respectively. According to the comparison results, we see that our method extracts most of the functional categories which appeared in figure 4-(a). Here, 'DEEPEST' of figure 4-(a) and 'SPAG5' of figure 4-(b) represent the identical gene symbol. Also, the first column of each line in figure 4-(b) shows the rank of the extracted term, which can be applied in further analysis. Figure 4-(c) shows the terms with SV as a GO tree structure, enabling scientists to understand the hierarchical structure of the extracted terms.

5. System Interface

This system was developed using a tool of Visual Studio.net 2003. An importing module of micro-array data was implemented using a Microsoft JET OLEDB 4.0 library, and a Gene Ontology Repository and Gene knowledge-base were built in with MySQL 5.0.

Figure 5 shows a sample screen shot of the user interface of our system. The user can import a micro-array data set by clicking the "Open" button. Additionally, our system has the following function buttons and operates as follows. With the "3-Fold" button, it filters out the genes which have mean differences of more than threefold. Figure 5-(a) shows the Probe Set IDs of the extracted genes. With the "T-test" button, it then calculates the t-values of the current genes and ranks them. The results are shown in figure 5-(b).



Cell cycle regulation, DNA replication and repair, and mitosis		
5	5.70	*DEEPEST*
7	4.93	*KNTC1*
8	6.84	*FEN1*
54	3.20	*MCM7*
56	3.81	*BUB1B*
57	5.61	*SMC4L1*
58	10.02	*STK15*
59	3.12	*ZWINT*
Transcriptional regulation, chromatin modification, RNA proces		
2	8.55	*CSPT5*
6	4.33	*SMARCD1*
13	4.99	*MYBL2*
18	28.60	*SF3A2*
20	5.40	*HOXB5*
22	0.14	*ZFP36*
24	11.88	*EZH2*
Cell adhesion, migration, cytoskeleton, and extracellular matrix		
1	0.15	*TAGLN*
3	0.23	*ACTA2*
30	0.13	*FHL1*
46	0.23	*ITGA8*
47	7.70	*HMMR*
61	0.15	*TPM3*
66	0.29	*ITGA7*
80	4.43	*THBS2*
82	0.19	*RARRES1*
97	4.19	*CEACAM7*

(a)

375	RNA polymerase II transcription factor activity,	MEIS1			
376	mitosis	SPAG5	BUB1	MAD2L1	KIF2
377	M phase of mitotic cell cycle	SPAG5	BUB1	MAD2L1	KIF2
378	M phase	SPAG5	BUB1	MAD2L1	KIF2
379	mitotic cell cycle	PLK	FEN1	SPAG5	BUB

969	covalent chromatin modification	EZH2	SPP1	
970	chromatin modification	EZH2	SPP1	
971	protein amino acid methylation	EZH2	SPP1	
972	biopolymer methylation	EZH2	CNTN1	SPP1
973	protein amino acid alkylation	EZH2	SPP1	

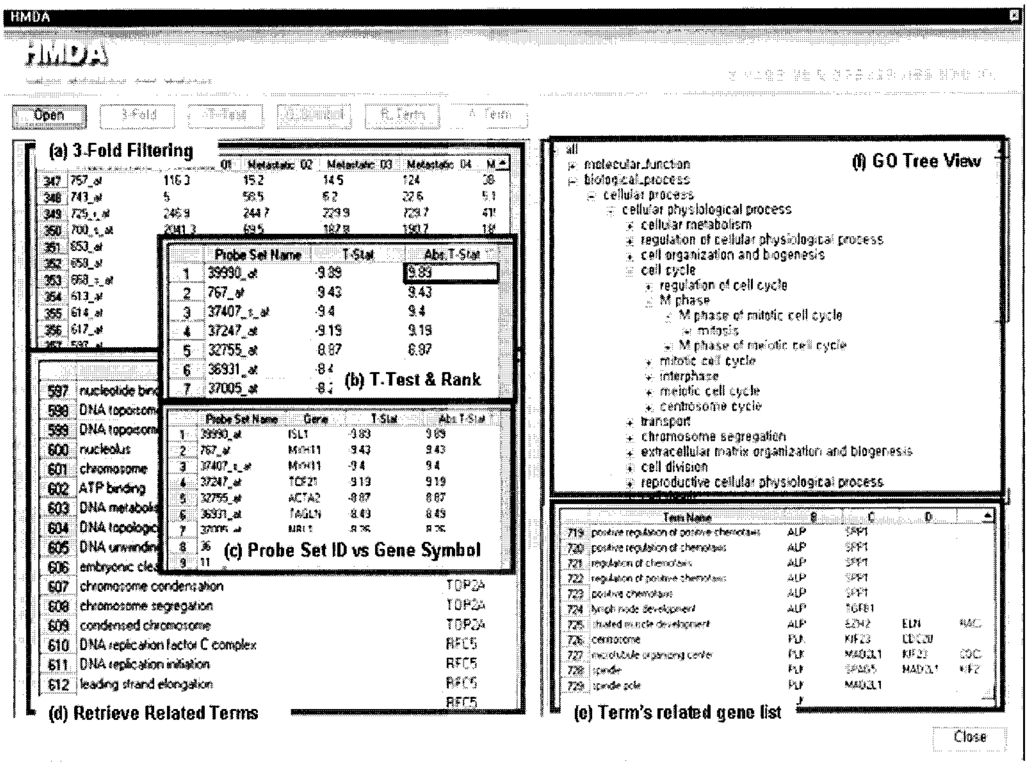
58	actin cytoskeleton	MYH11	ACTA2	ACTG2	CSRP
59	cytoskeleton	MYH11	ACTA2	ACTG2	CSRP
60	intracellular non-membrane-bound organelle	MYH11	ACTA2	ACTG2	CSRP
61	non-membrane-bound organelle	MYH11	ACTA2	ACTG2	CSRP
62	cytoskeletal part	MYH11	ACTA2	ACTG2	TPM

(b)

- cell cycle (1.3)
  - regulation of cell cycle (1.46)
  - M phase (1.15)
    - M phase of mitotic cell cycle (1.26)
      - mitosis (1.26)

(c)

Fig. 4. Functional analysis results



		01	Metastatic 02	Metastatic 03	Metastatic 04	M
347	757_at	116.3	15.2	14.5	124	38
348	743_at	5	58.5	6.2	22.6	5.1
349	725_s_at	246.9	244.7	229.9	729.7	41
350	700_s_at	2041.3	69.5	187.8	190.7	18
351	653_at					
352	658_at					
353	668_s_at					
354	613_at					
355	614_at					
356	617_at					

	Probe Set Name	T-Stat	Abs.T-Stat
1	39990_at	-9.89	9.89
2	767_at	-9.43	9.43
3	37407_s_at	-9.4	9.4
4	37247_at	-9.19	9.19
5	32755_at	-8.87	8.87
6	36931_at	-8.4	
7	37005_at	-8.3	

	Probe Set Name	Gene	T-Stat	Abs.T-Stat
1	39990_at	ISL1	-9.89	9.89
2	767_at	MYH11	-9.43	9.43
3	37407_s_at	MYH11	-9.4	9.4
4	37247_at	TCF21	-9.19	9.19
5	32755_at	ACTA2	-8.87	8.87
6	36931_at	TAGLN	-8.49	8.49
7	37005_at	NBL1	-8.36	8.36
8	36			
9	11			

597	nucleotide bind	
598	DNA topoisom	
599	DNA topoisom	
600	nucleolus	
601	chromosome	
602	ATP binding	
603	DNA metaboli	
604	DNA topologic	
605	DNA unwindin	
606	embryonic cle	
607	chromosome condensation	TOP2A
608	chromosome segregation	TOP2A
609	condensed chromosome	TOP2A
610	DNA replication factor C complex	RFC5
611	DNA replication initiation	RFC5
612	leading strand elongation	RFC5
		RFC5

- all
  - molecular_function
  - biological_process
    - cellular process
      - cellular physiological process
        - cellular metabolism
        - regulation of cellular physiological process
        - cell organization and biogenesis
        - cell cycle
          - regulation of cell cycle
          - M phase
            - M phase of mitotic cell cycle
              - mitosis
            - M phase of meiotic cell cycle
          - mitotic cell cycle
          - interphase
          - meiotic cell cycle
          - centrosome cycle
        - transport
        - chromosome segregation
        - extracellular matrix organization and biogenesis
        - cell division
        - reproductive cellular physiological process

	Term Name	B	C	D	
719	positive regulation of positive chemotaxis	ALP	SPP1		
720	positive regulation of chemotaxis	ALP	SPP1		
721	regulation of chemotaxis	ALP	SPP1		
722	regulation of positive chemotaxis	ALP	SPP1		
723	positive chemotaxis	ALP	SPP1		
724	lymph node development	ALP	TGFB1		
725	striated muscle development	ALP	EZH2	ELN	RAC
726	centrosome	PLK	KIF23	CDC20	
727	microtubule organizing center	PLK	MAD2L1	KIF23	CDC
728	spindle	PLK	SPAG5	MAD2L1	KIF2
729	spindle pole	PLK	MAD2L1		

Fig. 5. Sample screen shot of the HMDA GUI

With the "G_Symbol" button, it returns the gene symbols which correspond to the Probe Set IDs of the current genes. Figure 5-(c) shows the results. With the "R_term" button, it retrieves the terms which represent the functional categories of the related genes. The result is shown in figure 5-(d). With the "A-term" button, it shows the related gene list of each term, and the result is shown in figure 5-(e). Also, the GO tree which represents the hierarchical structure of related terms is given in figure 5-(f).

6. Conclusion and Future works

The proposed method is based on a statistical analysis model and identifies the functional categories of differentially expressed genes by using GO. HMDA can perform a functional analysis of micro-array data which consists of two group samples. As a further study, we are investigating sophisticated techniques to analyze the multiple micro-array data results of the heterogeneous platform, data format, and normalization methods.

References

[1] http://www.geneontology.org/

[2] http://www.nbn.ac.za/Educatoin/14-microarray-2004

[3] E. LaTulippe, J. Satagopan, A. Smith, H. Scher, P. Scardino, V. Reuter and W. L. Gerald, "Comprehensive Gene Expression Analysis of Prostate Cancer Reveals Distinct Transcriptional Programs Associated with Metastatic Disease," Cancer Research, vol. 62, pp. 4499-4509, 2002.

[4] F. Al-Shahrour, R. Diaz-Uriarte and J. Dopazo, "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes," Bioinformatics, Vol. 20, no. 4, pp. 578-580, 2004.

[5] B. Adryan and R. Schuh, "Gene-Ontology-based clustering of gene expression data," Bioinformatics, Vol. 20, no. 16, pp. 2851-2852, 2004.

[6] P. N. Robinson, A. Wollstein, U. Bohme and B. Beattie, "Ontologizing gene-expression micro-array data: characterizing clusters with Gene Ontology," Bioinformatics, Vol. 20, no. 6, pp. 979-981, 2004.

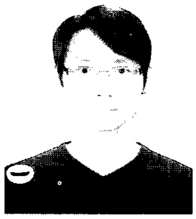
[7] M. Smid and L. C. J. Dorssers, "GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms," Bioinformatics, Vol. 20, no. 16, pp. 2618-2625, 2004.

[8] P. Khatri, S. Draghici, G. C. Ostermeier and S. A. Krawetz, "Profiling Gene Expression Using Onto-Express," Genomics, Vol. 79, no. 2, pp. 266-270, 2002.

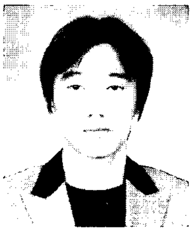
[9] A. Young, N. Whitehouse, J. Cho and C. Shaw, "Ontology Traverser: and R package for GO analysis," Bioinformatics, Vol. 21, no. 2, pp 275-276, 2005.

[10] P. Khatri and S.Draghici, "Ontological analysis of gene expression data: current tools, limitations and open problems," Bioinformatics, Vol. 21, no. 18, pp. 3587-3595, 2005.

Dong-wan Hong
Hong received BS and MS degrees in Computer Science and Computer Engineering from Hallym Univ. in 1996 and 1998, respectively. During 2003-2006, he worked as an assistant professor in the Dept. of Media Contents, Songgok College. He is now undertaking a doctorate course as a member of the Database Laboratory at Hallym Univ. His research interests include gene expression analysis, micro-array data integration, and gene detection using Gene Ontology.



Jong-keun Lee
Lee received BS and MS degrees in Computer Engineering from Hallym Univ. in 2005 and 2007, respectively. He is now studying a doctorate course as a member of the Database Lab. at Hallym Univ. His research interests include gene expression anaiysis, micro-array data integration, and gene detection using Gene Ontology.



Sung-soo Park
Park received BS and MS degrees in Computer Engineering from Hallym Univ. in 2005 and 2007, respectively. He is currently staying with SysGATE Inc. to develop the solutions with SI, ITSM. His research interests include gene expression analysis, micro-array data integration, gene detection using Gene Ontology, web service, semantic web, ontology and ITSM.

Sang-kyoon Hong
Hong received BS and MS degrees in Computer Science from Hallym Univ. in 2005 and 2007, respectively. He is now undertaking a doctorate course as a member of the Database lab at Hallym Univ. His research interests include sequence database and bioinformatics, XML security systems, and database systems.

Jee-hee Yoon
Yoon received a Ph.D. degree in Information Engineering from Kyushu Univ. in 1988. She has been a professor at Hallym Univ. since 1988. Her research interests are in the areas of DNA sequence search, shape-based retrieval in time-series databases, and micro-array data analysis.