International Journal of Fuzzy Logic and Intelligent Systems, vol. 12, no. 2, June 2012, pp. 167-172
http://dx.doi.org/10.5391/IJFIS.2012.12.2.167 pISSN 1598-2645 eISSN 2093-744X

Pathway Retrieval for Transcriptome Analysis using Fuzzy Filtering Technique and Web Service

Kyung Mi Lee and Keon Myung Lee*

Department of Computer Science, Chungbuk National University, Cheongju, 361-763, Republic of Korea

Abstract

In biology the advent of the high-throughput technology for sequencing, probing, or screening has produced huge volume of data which could not be manually handled. Biologists have resorted to software tools in order to effectively handle them. This paper introduces a bioinformatics tool to help biologists find potentially interesting pathway maps from a transcriptome data set in which the expression levels of genes are described for both case and control samples. The tool accepts a transcriptome data set, and then selects and categorizes some of genes into four classes using a fuzzy filtering technique where classes are defined by membership functions. It collects and edits the pathway maps related to those selected genes without analyst' intervention. It invokes a sequence of web service functions from KEGG, which an online pathway database system, in order to retrieve related information, locate pathway maps, and manipulate them. It maintains all retrieved pathway maps in a local database and presents them to the analysts with graphical user interface. The tool has been successfully used in identifying target genes for further analysis in transcriptome study of human cytomegalovirous. The tool is very helpful in that it can considerably save analysts' time and efforts by collecting and presenting the pathway maps that contain some interesting genes, once a transcriptome data set is just given.

Key Words: bioinformatics, pathway, fuzzy filtering, web service, information retrieval

1. Introduction

For past decades the advances in biotechnology have changed the paradigm of biological studies. The high-throughput devices have enabled data-centered experiment design and exploratory analysis study instead of hypothesis-based experiment design-oriented study. Biological data such as sequence data, interaction network data, protein structure data and biology-related publication data have been being accumulated into databases which are accessible over the networks. Those well-organized, public data resources open up an environment in which researchers are facilitated to keep up with up-to-date findings.[1,2] Various tools have been developed and in use to help researchers easily search for relevant data from those resources over the communication networks.[1,3,4,5] It is sometimes burdensome to be familiar with those tools, and sometimes tedious work to have to repeat the similar operations for large number of data. It deserves the value to provide software tools which save researchers' efforts in bioinformatics works.

A transcriptome sequencing tool like RNA-Seq produces large volume of RNA expression data spanning up to tens of thousands genes. RNA-Seq is a revolutionary tool for transcriptomics and provides a precise and sensitive measurement of levels of transcripts using next-generation deep sequencing technology.[6] In transcriptome study, biologists first tries to identify some interesting genes which show different behaviors across case and control samples, where case sample indicates subject that have the condition like disease, and control indicates the sample without such condition. After that, they usually search pathway database to see which genes are participating in which biological processes.

In practice, the number of genes are huge and the analyst is in charge of choosing some genes for analysis from the given transcriptome data set. The pathway databases may have for genes the naming scheme different from the transcriptome data. In that case, she needs to find the corresponding names for the chosen genes prior to pathway database access. Once the gene names are determined, she search for the pathway maps one by one using them from the pathway database. Due to the complicated nature of pathway maps, it is usually not easy to locate where the genes are on the maps.

Manuscript received Apr. 30, 2012; revised Jun. 7, 2012; accepted Jun. 15, 2012.
*Corresponding author: Keon Myung Lee(kmlee@cbnu.ac.kr)
This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (grant No.2012-0000478).

This paper presents a bioinformatics tool developed to take care of all the burdensome tasks mentioned above for pathway retrieval. Once a transcriptome data set is given, the tool chooses some interesting genes using fuzzy filtering technique in which four classes of genes are described by membership functions. It converts the gene names into the ones used in the pathway database KEGG[3] from which pathway maps are retrieved. Using a series of web service invocations, it collects pathway maps related to the interesting genes, and edits those maps so that it is easy to spot which genes belong to which class on the maps. The collected and edited pathway maps are maintained in a local database and can be browsed by graphical user interface. All those tasks of the tool are carried out by the tool without human intervention.

2. Related Works

2.1 Pathway Databases and Pathway Tools

Living organisms grow and reproduce, maintain their structures, and respond to their environment in the course of their life. For all these processes, various chemical reactions occur in the cells of living organism. Metabolism indicates the enzyme-catalyzed chemical reactions that break down and synthesize the molecules needed for life.[8] Enzymes are proteins used to increase the rate of chemical reactions to break down and synthesize new biochemicals. For example, suppose that there is some enzyme e which catalyzes the reaction from m_1 and m_2 to m_3 as shown in Figure 1. In such reaction, m_1, m_2 and m_3 are called metabolites, especially m_1 and m_2 are called substrates and m_3 is called product.

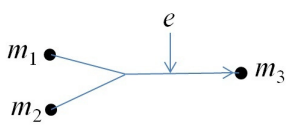

Figure 1: An enzyme-catalyzed chemical reaction

The product metabolites are further used in the synthesis of cellular components and more complex chemical molecules. A metabolic pathway is such a series of reactions that results in the transformation of a particular molecule into a different molecule.[8] Experimental discoveries about those chemical reactions are organized into pathway databases which consist of the sequence of biochemical reactions that produce a set of metabolites from a set of precursor metabolites in association with enzymes. The length of a pathway is the number of biochemical reactions from the precursor metabolite to the final product of the pathway, and the definition of a pathway is not unique. Hence, the number and length of pathways are different in pathway databases.[9]

Figure 2 shows a part of pathway for glycolysis which is a metabolic pathway that converts glucose into pyruvate along with releasing free energy used to form the high-energy compound.[8] In the figure, the rectangles on the edges indicate enzymes and the four-digit number in them represents the standardized enzyme identifier called Enzyme Commission number. The pathway networks assign metabolites to nodes and connect with edges the corresponding nodes to interconverting metabolites.

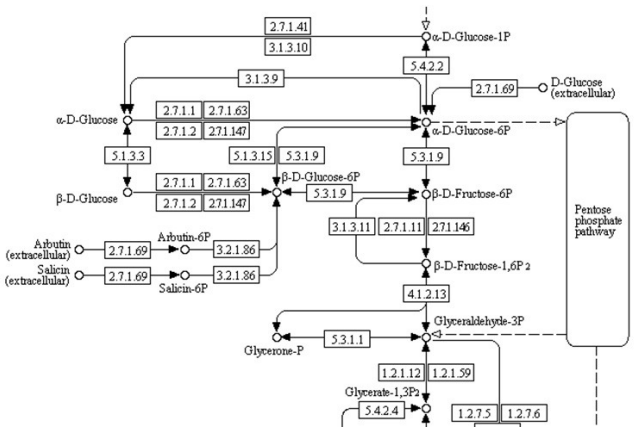

Figure 2: A part of pathway *Glycolysis* from KEGG

The pathway networks themselves have been analyzed to extract some meaning information. One of such analysis is based on graph theoretic methods which examine the connectivity and diameter of graphs.[11] The connectivity of a node (i.e., metabolite) means the number of edges (i.e., reactions) adjacent to thee node. The network diameter is the averaged path length of the shortest paths over all pairs of nodes. Ma and Zeng[11] examined the connectivities and the network diameters for the pathway networks for 80 organisms. They found that the pathway networks for all the organisms are a scale-free network in which the degree of connectivity of nodes follows a power law.

In the metabolic flux analysis, the metabolic pathways are used to simulate the fluxes (rates) of the reactions. The simulation is based on the the basic principle of mass balance which requires that at steady state the sum of fluxes of the rations producing a metabolite must be the sum of the fluxes of the reactions consuming this metabolite.[8] By the constrained convex optimization problem solving, the elementary flux modes and extreme pathways could be identified. An elementary flux mode is a subset of the biochemical reactions connecting a subset of metabolites in the network, satisfying the mass balance, from which a removal of a chemical reaction will disrupt the synthesis of the product. Extreme pathways are a unique and minimal set of convex basis vectors that consist of steady state functions of a metabolic network.[12] The information about extreme pathways is used to study the regulatory mechanisms of a metabolic network.[12] All these analyses are conducted based on the constructed pathway networks.

Various pathway databases have been developed and

available over the Internet. MetaCyc, KEGG, Reactome, Model SEED, and BiGG are some of the widely used metabolic pathway databases.[2] These databases provide various analysis and browsing tools that are available through Web-interface. In addition, there have been developed other tools which facilitate easy manipulation of the available databases. DAVID is a web-based integrated service for functional annotation which allows to systematically extract biological meaning from large gene or protein lists.[13] For a submitted list of genes, it performs some analysis using text and pathway mining tools such as gene functional classification, functional annotation chart or clustering, and functional annotation table. The Pathway Tools version 13.0 is an integrated software for pathway and genome informatics and systems biology for creating a type of model-organism database.[14] It has inference capabilities for prediction of metabolic pathways, prediction of metabolic pathway hole filters, and so on. It provides interactive editing tools for pathways. In this study, we are concerned with retrieving all pathway maps which contains some genes potentially of interest from transcriptome data. To our best knowledge, we are unaware of any systems to provide the pathway retrieval service in the manner this study works on.

2.2 Web Service

The developed pathway retrieval service has been implemented by using the web services provided by KEGG, which is a collection of online databases that have been developed by the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo, Japan.[3] A Web service is a communication method for an application to invoke services available over the web. The World Wide Web Consortium, the main international standards organization for the World Wide Web, has standardized the way to present the available services with the web service description language WSDL, to publish them using the directory service UDDI, and to describe messages with the protocol SOAP.[7] Each Web service has an interface described in a machine-readable format, WSDL. Other systems interact with the Web service using SOAP-messages. SOAP is a protocol that governs the format and processing rules of a SOAP message. The SOAP messages are typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.[7] As various services are made available from Web service providers, Web-connected programs with powerful capabilities, which was just allowed on the server side, can be built on the user side using Web services.

3. The Concept of the Developed System

3.1 The Functionality of the System

Next generation sequencing techniques like RNA-Seq have enabled to measure expression levels of tens of thousands of genes at a time. Comparative analysis like fold-change measurements reveals which genes are over-expressed or under- expressed. Among tens of thousands of genes, the analysts usually focus on those distinctively expressed genes and sometimes search for pathways which are involved with them. The pathway databases have the interface to retrieve pathways in which query genes are placed. When gene list is long, the analyst needs to carry out bothersome tasks, such as repetitive query submission and saving the results, to get the whole results. Even when a pathway map is given, it is sometimes not easy to locate the designated genes on the map due to complicated map structure.

A software tool is developed, which retrieves metabolic pathway maps which include some of interesting genes and edits them in order to highlight the interesting genes for easy location, once a transcriptome data set is given. It stores all retrieved and edited pathway maps in a local database and allows to browse them through the graphical user interface. It makes use of the Web services provided by the KEGG database in order to access and manipulate pathways archived in the database.

3.2 The Selection of Interesting Genes Based on Fuzzy Filters

In experimental study, there are usually two groups of expression data one of which is a case group and the other of which is its control group to be compared. When biologists look into a transcriptome data, they pay attention to how expression level of genes has been changed between case and control groups. As genes of interest for analysis, the following 4 classes are considered in the tool: *over-expressed, suppressed, appeared*, and *disappeared*. The over-expressed class indicates a group of genes whose expression level is significantly higher in case than in control. The suppressed class means the opposite situation of the over-expressed class. The appeared class indicates the genes which are expressed meaningfully in the experiment samples but not detected in the control samples. The disappeared class indicates the opposite of the appeared one.

For easy selection of interesting genes from transcriptome data, a kind of fuzzy filter is proposed which selects the genes for the 4 classes, each of which is defined by membership functions. A transcriptome data set can be expressed as a matrix $E = (e_{ij})$ of which element e_{ij} indicates the expression level of a gene g_i in sample s_j. Suppose that there are one case sample s_1 and one control sample s_2. The fold change fc_i of gene g_i is defined as $log_2(e_{i1}/e_{i2})$ and thus it is close to zero when there is no significant difference between case and control. For

over-expressed and suppressed classes, we adopt two corresponding membership functions $\mu_{overexpressed}(fc)$ and $\mu_{suppressed}(fc)$ which are defined on the fold change values fc . For appeared and disappeared classes, we use the membership functions $\mu_{notexpressed}(e)$, $\mu_{expressed}(e)$ defined over the expression level value e. The interesting genes are collected using the following criteria:

g_i is $over\text{-}expressed$ if $\mu_{overexpressed}(fc_i) \geq \theta_o$.
g_i is $suppressed$ if $\mu_{suppressed}(fc_i) \geq \theta_s$.
g_i is $appeared$ if $\mu_{notappeared}(e_{i2}) \geq \theta_d$ and $\mu_{appeared}(e_{i1}) \geq \theta_a$.
g_i is $disappeared$ if $\mu_{notappeared}(e_{i1}) \geq \theta_d$ and $\mu_{appeared}(e_{i2}) \geq \theta_d$.

Here θ_o, θ_s, θ_a and θ_d represent the threshold for over-expressed, suppressed, appeared, and disappeared class, respectively. Depending on the threshold values which the analyst specifies, the number of selected genes for each group vary. The analyst just needs to submit the transcriptome data set and the threshold values to the tool, and then she gets the pathway maps which contain the genes likely to be interesting to her.

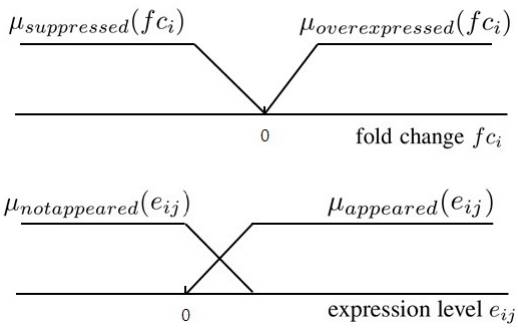

Figure 3: The membership functions for filtering

3.3 The Operations of the System

In order to retrieve all pathways containing genes of interest and construct their pathway maps of which interesting genes are marked in designated colors, the tool takes the following steps:

Once a transcriptome data set is given, it applies the fuzzy filters to choose genes belonging to one of 4 classes. The chosen genes for each class are stored in the local database for later use.

The gene names of transcriptome data are different from those used in the KEGG database which uses UniGene name scheme. Hence, the tool converts the gene names into UniGene identifiers(IDs) by using KEGG web service which maps various kinds of gene names into UniGene IDs.

By issuing a query for each chosen gene to the KEGG database, it retrieves the IDs of pathway maps in which the gene is included, and keeps them in the local database. After that, it takes the union for all retrieved pathway map IDs, which now becomes the set of pathway map identifiers that contain some of interesting genes.

Meanwhile the tool allows the analyst to set the foreground and background colors for each class, which are used to shade the nodes corresponding to those classes. For each pathway map identifier in the union, it requests the KEGG Web service to generate a pathway map in which the interesting genes are shaded according to the colors designated for each class. The service request is made with the information about pathway map ID, gene IDs with their corresponding foreground and background colors determined on the basis of their class. The generated pathway maps are fetched from the KEGG site and stored into the local database. The retrieved pathway maps could be browsed through graphical user interface from the local database.

Figure 4 show the sequence diagram of the operations for the developed system. After the analyst feeds a transcriptome data file and the threshold values for interesting gene selection, the system conducts a sequence of operations some of which are the Web service invocation to the KEGG database, and finally constructs a database of pathway maps in which interesting genes are marked according to her color assignment.

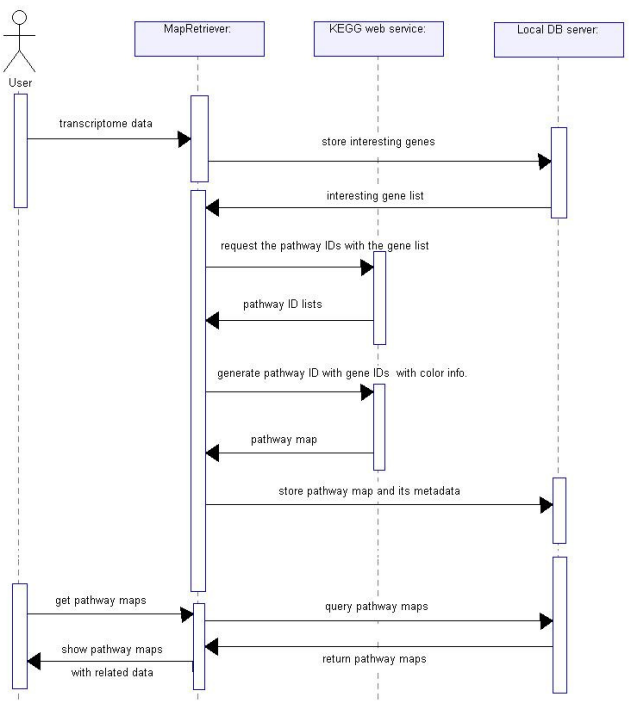

Figure 4: The sequence diagram of the system

4. Implementation

The pathway retrieval tool has been implemented as a

stand-alone system to use the KEGG Web Service APIs for KEGG database access. The KEGG Web service APIs are available for retrieving the pathway IDs of which map contains a specified gene, for shading specific nodes on the map according to the requested direction, and so on. It uses Java language for its portability across different platforms and MySQL for local database management. The system provides the graphical user interface through which the users can easily request pathway maps by submitting the transcriptome data and the threshold values. A transtriptome data set needs to be prepared in a matrix format where each row and each column corresponds a gene and a sample, respectively. In the matrix, the first row contains the list of sample names separated by a tab character, and each row for a gene starts with the gene name and contains expression level values in the corresponding samples.

Figure 5 shows the snapshot of the developed system, in which the center pane is the area to display a retrieved pathway map. In the map, the color shaded nodes indicate the genes identified by the gene filtering membership functions and their color indicates the class to which the corresponding gene belongs.

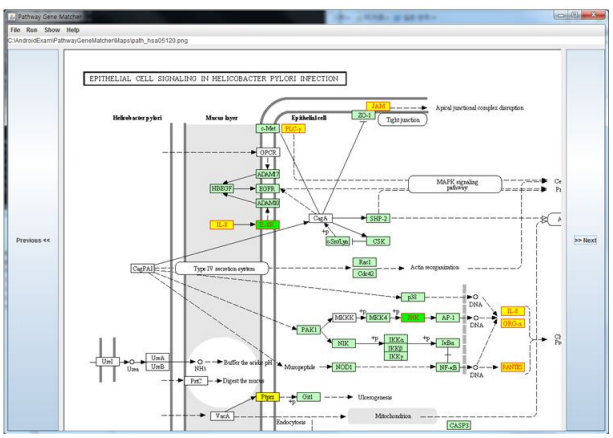

Figure 5: The developed system

5. An Application of the Developed System

The developed system can be used as a front-end analysis for transcriptome study in biology. It has been successfully applied to a practical biological study conducted by one of a biologist. An RNA-Seq data set was obtained with Illumina Genome Analyzer which is an RNA-Seq sequencer to understand the cellular transcriptome change following human cytomegalovirus infection. The data set was fed into the system and the threshold values were set so that the genes with the fold-change of magnitude greater than around 2 were grouped into the over-expressed genes, the genes of which fold-change is less than around -1 were into the suppressed genes, the genes expressed only in the control were into the appearing genes, and the genes not detected in the sample were into the disappeared genes.

From the RNAseq data with 36,027 genes, 3,387 genes were selected for the four categories by the fuzzy filters as follows: 1,037 over-expressed ones, 625 suppressed ones, 1,385 appearing ones, and 339 disappeared ones. Out of 3387 genes, only 2,856 genes were registered in the KEGG database. The number of pathway maps which contain some of those genes was 209. Figure 6 shows one of pathway maps likely to attract attention to the biologist because all genes on the pathways are selected as being significantly changed among case and control samples. From the pathways on the map, experiment design was made and the biological experiments have been conducted to identify some meaningful association with cytomegalovirus infection. In the figure, the boxes with yellow background and red foreground indicate the over-expressed genes, those with green background and red foreground show the suppressed one, those with yellow background and black foreground correspond to the appearing ones, and those with yellow background and green foreground illustrate the disappeared ones.

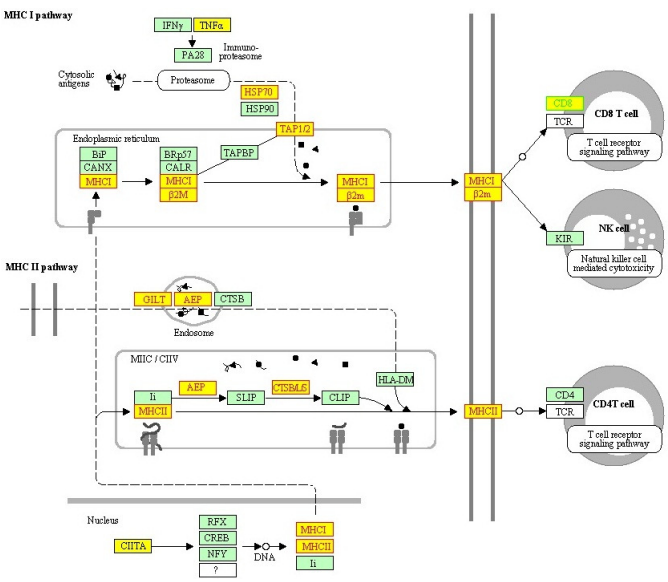

Figure 6: A pathway map with meaningful sequences of reactions by interesting genes

6. Conclusions

A bioinformatics tool has been developed to retrieve pathways that contain interesting genes from a transcriptome data set. It allows the analysts to choose interesting genes from the genes of the data set by specifying the threshold values for the membership functions which are used to define the 4 classes: *over-expressed, suppressed, appeared*, and *disappeared*. For the chosen genes, it collects the information for the pathway maps with some of

the genes, and edits the maps so that the interesting genes are highlighted according to their class. The edited pathway maps are maintained in its local database and hence the analyst can easily browse them with no access to external databases. It is unique in that it can build a local database with pathway maps related to interesting genes which are determined from a given transcriptome data set. The system had been successfully used in a practical application. The developed system is expected to be very useful in transcriptome study because it could considerably save the analysts' labors and time.

References

[1] D. W. Huang, B. T. Sherman and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol.4, No.1, 2009.

[2] R. D. Karp and R. Caspi, "A Survey of Metabolic Databases Emphasizing the MetaCyc Family," *Arch. Toxicol*, vol.85, pp.1015-1033, 2011.

[3] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic Acids Research*, vol.32, 2004.

[4] D. W. Huang, B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane and R. A. Lempicki, "The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists," *Genome Biology*, vol.8, Issue 9, 2007.

[5] M. Chen and R. Hofestadt, "Web-Based Information Retrieval System for the Prediction of Metabolic Pathways," *IEEE Trans. on Nanobioscience*, vol.3, no.3, 2004.

[6] Z. Wang, M. Gerstein, M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat. Rev. Genet.*, vol.10, no.1, pp.57-63, 2009.

[7] W3C Web Service Activity, http://www.w3.org/2002/ws/.

[8] A. L. Lehninger, D. L. Nelson and M. M. Cox, *Principles of Biochemistry*, New York, 1993.

[9] V. Hatzimanikatis, C. Li, J. A. Ionita and L. J. Broadbelt,"Metabolic networks: enzyme function and metabolite structure," *Current Opinion in Structural Biology 2004*, vol. 14, pp. 300–306, 2004.

[10] C. F. Schaefer, "Pathway Databases," *Ann N. Y. Acad. Sci.*, vol.1020, pp.77-91, 2004.

[11] H. W. Ma, A. P. Zeng, "Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms," *Bioinformatics*, vol.19, 270-277, 2003.

[12] J. A. Papin, N. D. Price, B.O. Palsson, "Extreme pathway lengths and reaction paprticipation in genome-scale metabolic network,", *Genome Res.*, vol.12, no.12, pp.1889-1900, 2002.

[13] DAVID, http://david.abcc.ncifcrf.gov.

[14] P. D. Karp, S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I. M. Keseler, R. Caspi, "Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology," *Briefings in Bioinformatics*, vol.2, no.1, pp.40-79, 2009.

Kyung Mi Lee

PhD course student, Department of Computer Science, Chungbuk National University
Research Area: soft computing, artificial intelligence applications
E-mail : kmlee07@cbnu.ac.kr

Keon Myung Lee

Professor, Department of Computer Science of Chungbuk National University
Research Area: machine learning, data mining, bioinformatics, big data mining
E-mail : kmlee@cbnu.ac.kr