Design and Implementation of a Directory System for Disease Services

Myung-Ho Yeo*
Department of Computer and Communication Engineering
Chungbuk National University, Cheongju, Korea.

Yoon-Kyeong Lee
Department of Biochiemistry
Chungbuk National University, Cheongju, Korea.

Kyu-Jong Roh, Hyeong-Soon Park, Hak-Sin Kim, Jun Ho-Park, Tae-Ho Kang
Department of Computer and Communication Engineering
Chungbuk National University, Cheongju, Korea.

Hak-Yong Kim,
Department of Biochiemistry
Chungbuk National University, Cheongju, Korea.

Jae-Soo Yoo
Department of Computer and Communication Engineering
Chungbuk National University, Cheongju, Korea.

ABSTRACT

Recently, biological researches are required to deal with a large scale of data. While scientists used classical experimental approaches for researches in the past, it is possible to get more sophisticated observations easily with the convergence of information technologies and biology. The study on diseases is one of the most important issues of the life science. Conventional services and databases provide users with information such as classification of diseases, symptoms, and medical treatments through the Web. However, it is hard to connect or develop them for other new services because they have independent and different criteria. It may be a factor that interferes the development of biology. In this paper, we propose integrated data structures for the disease databases. We also design and implement a novel directory system for diseases as an infrastructure for developing the new diseases services.

Keywords*: Disease Database, Disease directory System, Data Integration.*

1. PREPARATION OF PAPERS

Recently, bioinformatics has been recognized as one of the important fields in life science. Bioinformatics is the application of information technology and computer science to the field of molecular biology[1][2]. Many scientists have constructed various databases that includes gene sequences and amino-acid sequences in order to accelerate the researches of bioinformactics fields. In recent, databases related with diseases have also been built for various disease information services. The study of diseases are the important theme of biologists, and it also gives interests to all people. Nowadays, various databases related with diseases have been constructed and serviced using the web. However, the disease databases were constructed for other objectives each other and have other data formats. Therefore, it is very difficult to get various disease information services in the existing disease databases. Thus, it is necessary to integrate the independent data formats with the disease related research efficiently.

The data integration are helpful to reduce the time and economic loss of disease related research. In this paper, we propose data structures to construct the integrated disease

[*] *Corresponding author. E-mail : yjs@chungbuk.ac.kr*

Manuscript received Feb. 16, 2010 ; accepted Mar. 17, 2010

database. We also design and implement a novel directory system as an infrastructure for the new diseases services. Our disease directory system provides a variety of disease information services through Web.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 describes the major functions of the proposed disease directory system. Section 4 implement the proposed disease directory system. Finally, section 5 presents the conclusions.

2. RELATED WORK

The representative disease databases are CHE[3], Gastro net[4], Findis[5], AID[6], 3DinSight[7], OMIM-Morbid Map[8], and DiseaseDatabase[9]. CHE is a medicine and disease database. It offers simple information about chemical medicine and 180 kinds of human disease. Through CHE, we can get information about the lists of matters occurring diseases and the occurring extent with 3 levels(Strong, Good, Limited Evidence). Strong evidence is approved by medical science groups. It is indicated when the evidence is enough as appeared in a text book. Good evidence is indicated when it has a little evidence appeared by people, or it was improved strongly through animal experiments. Limited/conflicting evidence is indicated when it appeared by people or animal experiment weakly.

Gastro net is a medical information site that is offered on line by patients and medical specialist. Principally, it offers information about the disease related to the stomach, intestines, and the digestive organ. We get through Gastro net a simple explanation about disease, symptom, treatment, and so on. Findis is a database that scientist and doctor examined the paper and collected only reliable data. It offers each transference of disease, explanation about occurrence-rate and clinical symptom and gene transference information. AID analyzes more than 50,000 related papers recorded in MEDLINE and offers gene information related to self-immunity diseases as an accumulated database. We can get information about self-immunity in AID. It is linked to Entrez Gene, Ensemble, and SwissProt. So, we can confirm additional information about the related gene and get papers about it through PMID. In 3DinSight, we can search information by using a keyword, protein name, gene name and disease name. This site offers information about molecules related to specific disease as an integrated database about function of creature molecules, character, structure, mutation and disease. OMIM-Morbid Map offers gene information related to the disease in OMIM and cytogenetic map location of genes. We also can get additional biological information about gene and the information of the related paper because it is linked in OMIM. However, it is very difficult to integrate the databases and develop other new services using them because they have independent and different formats. This may be a factor that interferes the development of biology.

Therefore, we propose an integrated data structure for the disease databases. We also design and implement a novel directory system for diseases as an infrastructure for developing the new diseases services.

3. DESIGN OF THE DISEASE DIRECTORY SYSTEM

3.1 Containment Structure

Existing disease information systems supports the classification of disease data and provides users with the data through the web. However, each service uses different ways in classifying data each other. Also, there is no way of linking a new service which requires high-dimensional processing. In this paper, we propose a new directory system that solves the problems of existing disease information systems. Figure 1 shows the structure of the proposed directory system.

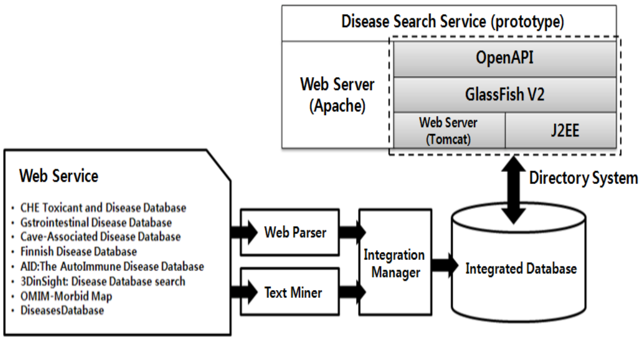


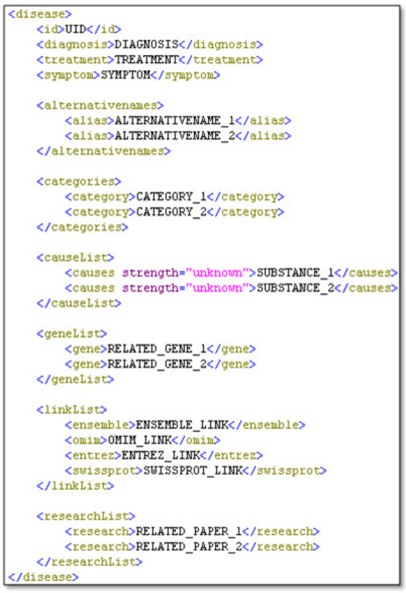
Fig. 1. The proposed directory system structure

In order to integrate the existing disease databases, we collected disease information from many web sites. Each data was collected in required forms in the system using the web parser and the text miner. The collected data through the web parser and the text miner were stored in the integration database schema by the integration manager. We give users various operations such as search, addition, modification, and deletion through the application server, called Glass Fish V2 and provides Open API.

3.2 Integration of Databases

To integrate the existing disease databases, we first analyze disease information provided on the existing disease search services. They provide services about duplicate attributes and other attributes. CHE provides services such as category, accuracy, causes, and related papers about diseases. We implement proper web parser and text miner in order to extract information from the existing disease databases. And then, integration manager creates a disease identifier and generates the integration database based on the identifier.

Users access the integrated data that are represented by XML documents through Open API. Figure 2 shows XML DTD that is designed for integrating data. By expressing the integrated database using the XML structure, it is very easy to search disease information by each attribute. Classifying data attributes obtained from each web service makes data management easier. It reduces the waste of unnecessary data spaces by constructing the integrated databases for specific disease information.



```
<disease>
    <id>UID</id>
    <diagnosis>DIAGNOSIS</diagnosis>
    <treatment>TREATMENT</treatment>
    <symptom>SYMPTOM</symptom>

    <alternativenames>
        <alias>ALTERNATIVENAME_1</alias>
        <alias>ALTERNATIVENAME_2</alias>
    </alternativenames>

    <categories>
        <category>CATEGORY_1</category>
        <category>CATEGORY_2</category>
    </categories>

    <causeList>
        <causes strength="unknown">SUBSTANCE_1</causes>
        <causes strength="unknown">SUBSTANCE_2</causes>
    </causeList>

    <geneList>
        <gene>RELATED_GENE_1</gene>
        <gene>RELATED_GENE_2</gene>
    </geneList>

    <linkList>
        <ensemble>ENSEMBLE_LINK</ensemble>
        <omim>OMIM_LINK</omim>
        <entrez>ENTREZ_LINK</entrez>
        <swissprot>SWISSPROT_LINK</swissprot>
    </linkList>

    <researchList>
        <research>RELATED_PAPER_1</research>
        <research>RELATED_PAPER_2</research>
    </researchList>
</disease>
```

Fig. 2. XML,DTD for integrating data.

3.3 Disease Services based on Web

In this paper, we propose a web service system for human disease services. The proposed web service system is divided into disease search services, disease addition services, disease renewal services, and disease deletion services. Each web service provides developers with APIs for disease services. The proposed system provides communication functions for disease services and conducts communication through SOAP/HTTP with web servers.

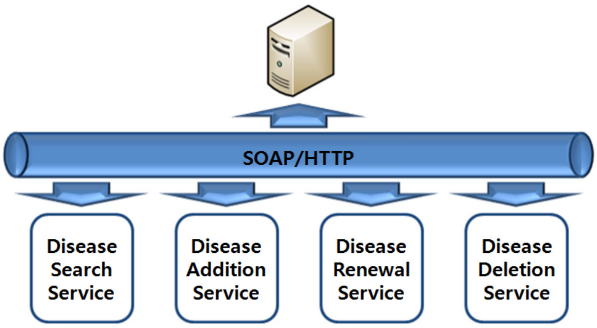


Fig. 3. The proposed web services

Disease search services allow users to search by disease identifier, disease name, disease category, medical diagnosis, treatment, symptom, causes, genes, paper, and so on. Also, the searched disease information is provided in the forms of XML data format. For the constantly found new disease information, the disease addition services allow users to add disease information easily. For so many different classifying methods for one disease information, disease renewal services made it possible to change the existing disease information to standardized disease information. Finally, when the same disease information is classified differently by another classification method, disease deletion services integrate the same disease information and delete unnecessary disease information.

4. IMPLEMENTATION OF THE DISEASE DIRECTORY SYSTEM

4.1 Implementation Environment

The proposed directory system is implemented using J2EE 1.4 and J2SDK in CentOS 5.2 environments. We use MySQL 5.0 as a database management system. We also implement various disease services to show the utilization of the directory system using AJAX(Asynchronous Javascript and XML) and PHP.

4.2 Database Structure for Integrating Databases

Figure 4 shows the structure of a disease database. The disease database consists of 7 tables such as diseases, category, gene, causes, alternativenames, reasearches and links to manage disease information. The diseases table keeps disease information. It consists of unique disease ID, diagnosis information, medical prescription and symptom information for a specific disease. The category table is composed of disease ID and category information. The gene table consists of disease ID and gene information. The causes table consists of disease ID, disease causing substance, and the intensity information of causing substance. The alternativenames table consists of disease ID and substitution disease information. The links table consists of disease ID, link provider, and link information. The researches table manages disease ID, name, title, author, year, and the language information of paper.

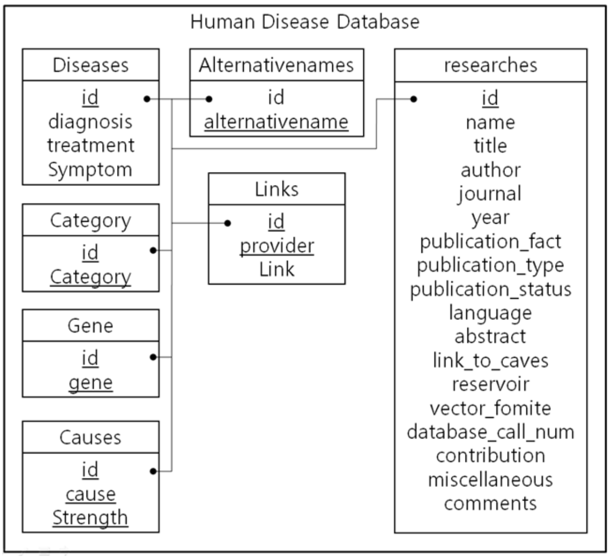


Fig. 4. Relation diagram of integration databases

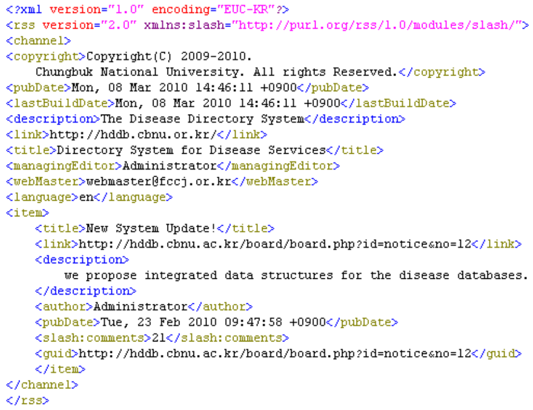
Table 1 shows the number of data that are included in DB table. The number of diseases is 24 and the number of alternative names is 19,644. The number of related Gene is 22864 and the number of Category is 392. The number of causes is 2760, the number of links is 8336, and the number of Researches is 209.

Table 1. The number of data included in DB tables

DB table	The number of data
Alternativename	19,644
Category	392
Causes	2,760
Diseases	24
Genes	22,864
Links	8,336
Researches	209

4.3 The Disease Directory System

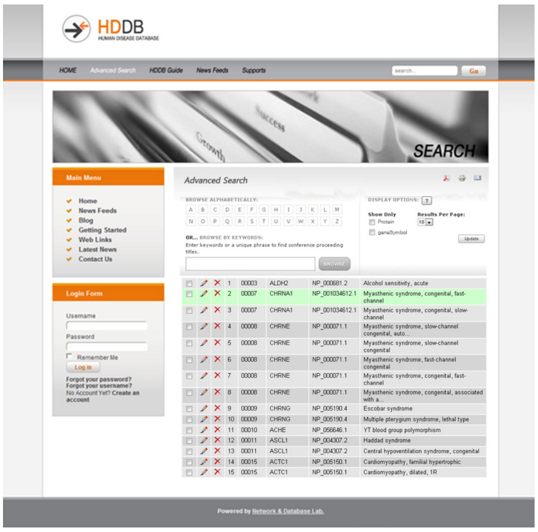
The disease directory system were implemented using AJAX(Asynchronous Javascript and XML) and PHP in order to provide various disease services. Our system integrates our disease databases with existing integrated databases, processes XML element efficiently, and provides the processed result through various interfaces. It supports Web interfaces for providing disease information. It also provides hyperlinks which makes it possible to use various disease information via the paths such as RSS feed, e-mail, and so on. Figure 5 shows a RSS document to provide RSS feed. Our web feeds provide users by letting them syndicate contents automatically. The users who subscribe to get the updated information in the RSS service system from our websites and geneDatabase can read the updated information using tools such as "RSS reader" and web browser.



```
<?xml version="1.0" encoding="EUC-KR"?>
<rss version="2.0" xmlns:slash="http://purl.org/rss/1.0/modules/slash/">
<channel>
<copyright>Copyright(C) 2009-2010.
    Chungbuk National University. All rights Reserved.</copyright>
<pubDate>Mon, 08 Mar 2010 14:46:11 +0900</pubDate>
<lastBuildDate>Mon, 08 Mar 2010 14:46:11 +0900</lastBuildDate>
<description>The Disease Directory System</description>
<link>http://hddb.cbnu.or.kr/</link>
<title>Directory System for Disease Services</title>
<managingEditor>Administrator</managingEditor>
<webMaster>webmaster@fccj.or.kr</webMaster>
<language>en</language>
<item>
    <title>New System Update!</title>
    <link>http://hddb.cbnu.ac.kr/board/board.php?id=notice&no=12</link>
    <description>
        we propose integrated data structures for the disease databases.
    </description>
    <author>Administrator</author>
    <pubDate>Tue, 23 Feb 2010 09:47:58 +0900</pubDate>
    <slash:comments>21</slash:comments>
    <guid>http://hddb.cbnu.ac.kr/board/board.php?id=notice&no=12</guid>
    </item>
</channel>
</rss>
```

Fig. 5. RSS document for RSS feed

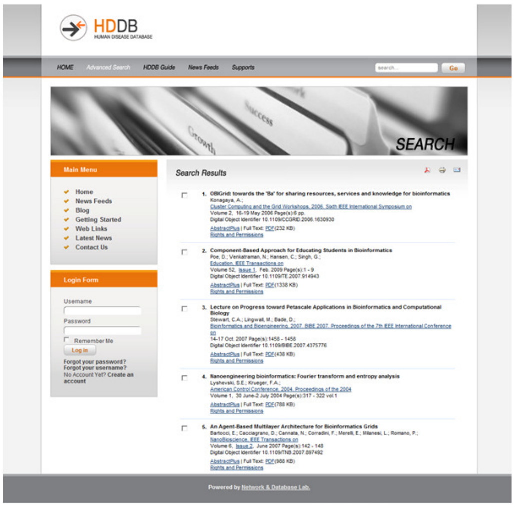
Figure 6 shows the interface of an advanced searching tool. This tool provides various searching options (geneSymbol, HPRD_ID, content types and so on). It supports the substring matching for genes, proteins, and Diseases. If “cancer” is searched, the tool returns documents that include words “cancer”, “breath cancer”, “lung cancer” and so on.



□	✎	✕	1	00003	ALDH2	NP_000681.2	Alcohol sensitivity, acute
□	✎	✕	2	00007	CHRNA1	NP_001034612.1	Myasthenic syndrome, congenital, fast-channel
□	✎	✕	3	00007	CHRNA1	NP_001034612.1	Myasthenic syndrome, congenital, slow-channel
□	✎	✕	4	00008	CHRNE	NP_000071.1	Myasthenic syndrome, slow-channel congenital, auto...
□	✎	✕	5	00008	CHRNE	NP_000071.1	Myasthenic syndrome, slow-channel congenital
□	✎	✕	6	00008	CHRNE	NP_000071.1	Myasthenic syndrome, fast-channel congenital
□	✎	✕	7	00008	CHRNE	NP_000071.1	Myasthenic syndrome, congenital, fast-channel
□	✎	✕	8	00008	CHRNE	NP_000071.1	Myasthenic syndrome, congenital, associated with a...
□	✎	✕	9	00009	CHRNG	NP_005190.4	Escobar syndrome
□	✎	✕	10	00009	CHRNG	NP_005190.4	Multiple pterygium syndrome, lethal type
□	✎	✕	11	00010	ACHE	NP_056646.1	YT blood group polymorphism
□	✎	✕	12	00011	ASCL1	NP_004307.2	Haddad syndrome
□	✎	✕	13	00011	ASCL1	NP_004307.2	Central hypoventilation syndrome, congenital
□	✎	✕	14	00015	ACTC1	NP_005150.1	Cardiomyopathy, familial hypertrophic
□	✎	✕	15	00015	ACTC1	NP_005150.1	Cardiomyopathy, dilated, 1R

Fig. 6. Interface of an advanced searching tool.

Figure 7 shows database search results in the disease directory system. It provides the integrated database search results when the disease is entered in the system. Detailed information for search results is given through hyperlinks. When a user requests a query, our system returns XML documents through the integrated database. It generates various forms of service pages by dynamically processing the XML documents.



Search Results

□ 1. OBIGrid: towards the 'Ba' for sharing resources, services and knowledge for bioinformatics
Konagaya, A.;
Cluster Computing and the Grid Workshops, 2006. Sixth IEEE International Symposium on
Volume 2, 16-19 May 2006 Page(s):6 pp.
Digital Object Identifier 10.1109/CCGRID.2006.1630930
AbstractPlus | Full Text: PDF(232 KB)
Rights and Permissions

□ 2. Component-Based Approach for Educating Students in Bioinformatics
Poe, D.; Venkatraman, N.; Hansen, C.; Singh, G.;
Education, IEEE Transactions on
Volume 52, Issue 1, Feb. 2009 Page(s):1 - 9
Digital Object Identifier 10.1109/TE.2007.914943
AbstractPlus | Full Text: PDF(1338 KB)
Rights and Permissions

□ 3. Lecture on Progress toward Petascale Applications in Bioinformatics and Computational Biology
Stewart, C.A.; Lingwall, M.; Bade, D.;
Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on
14-17 Oct. 2007 Page(s):1458 - 1458
Digital Object Identifier 10.1109/BIBE.2007.4375776
AbstractPlus | Full Text: PDF(438 KB)
Rights and Permissions

□ 4. Nanoengineering bioinformatics: Fourier transform and entropy analysis
Lyshevski, S.E.; Krueger, F.A.;
American Control Conference, 2004. Proceedings of the 2004
Volume 1, 30 June-2 July 2004 Page(s):317 - 322 vol.1
AbstractPlus | Full Text: PDF(788 KB)
Rights and Permissions

□ 5. An Agent-Based Multilayer Architecture for Bioinformatics Grids
Bartocci, E.; Cacciagrano, D.; Cannata, N.; Corradini, F.; Merelli, E.; Milanesi, L.; Romano, P.;
NanoBioscience, IEEE Transactions on
Volume 6, Issue 2, June 2007 Page(s):142 - 148
Digital Object Identifier 10.1109/TNB.2007.897492
AbstractPlus | Full Text: PDF(988 KB)
Rights and Permissions

Fig. 7. Disease retrieval service results

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an integrated data structure for different disease databases. We also have designed and implemented a novel disease directory system that provides an infrastructure to easily link other new services. To achieve this, we analyzed existing databases and constructed integrated databases based on XML with various attributes using a web parser and the text miner. It can be used a web services using a SOAP/HTTP communication. Our disease directory system provides various disease search services. In the near future, we will extend our directory system in order to support various functions such as disease mechanism analysis, relationship among diseases, and more meaningful information extraction.

ACKNOWLEDGEMENTS

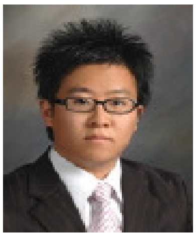
This work was supported by the Ministry of Education, Science and Technology Grant funded by the Korea Government"(The Regional Research Universities Program/Chungbuk BIT Research-Oriented University Consortium) and by the Ministry of Education, Science Technology (MEST) and Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Regional Innovation.

REFERENCES

[1] Luscomebe N.M and G. D, G. M., "What is bioinformatics? A proposed definition and overview of the filed", Methods Inf. Med 40:346-358, 2001.
[2] Lesk A. M. "Introduction to bioinformatics", pp.2-20, Oxford Iniversity Press, United Kingdom, 2002.
[3] CHE, http://database.healthandenvironment.org/
[4] Gastro net, http://www.gastro.net.au/gastrodiseases/
[5] Findis, http://www.findis.org/
[6] AID, http://www.uni-rostock.de/aidb/
[7] 3DinSight, http://gibk26.bse.kyutech.ac.jp/jouhou/3dinsight/
[8] OMIM-Morbid Map, http://www.ncbi.nlm.nih.gov/Omim/
[9] DiseaseDatabase, http://www.diseasesdatabase.com/

Myung Ho Yeo
He received the B.S., M.S. and Ph.D. in Information and Communication Engineering from Chungbuk National University, Korea in 2004 and 2010, respectively. He is now a researcher in the Agency for Defence Development, South Korea. His main research interests include main-memory database system, wireless sensor networks and bioinformatics.

Yoon Kyeong Lee
She received the B.S degree in Department of Biochemistry from Chungbuk National University, Korea in 2008. She is currently working towards M.S. degree on Department of biochemistry. Hers main research interests include bioinformatics, systems biology and signal transduction.



Kyu Jong Roh
He received the B.S and M.S degree in Department of Information and Communication Engineering from Chungbuk National University, Korea in 2008 and 2010 respectively. He is currently working in Samsung SDS, Korea. His main research interests include wireless sensor networks and database system.

Hyoung Soon Park

He received the B.S. and M.S. in Departments of Information and Communication from Chungbuk National University, Korea in 2008 and 2010 respectively. He is currently working in MacroImpact Inc, Korea. His main research interests include database system, wireless sensor network, LCMS and LMS.

Hak Sin Kim
He received the B.S degree in the departments of Computer Engineering ChungJu National University, Korea in 2008. And he received the M.S degree in departments of Information and Communication Engineering, Chungbuk National University, Korea. He is currently working in SK C&C, Korea. His main research interests include wireless ensor network and database system.

Jun Ho Park
He received the B.S. and the M.S degree in the Departments of Information and Communication Engineering, Chungbuk National University, Korea in 2008 and 2010 respectively. He is working towards Ph.D degree on Department of Information and Communication Engineering from Chungbuk National University, Korea. His main research interests are the database system, wireless sensor network, RFID system, LMS, semantics and bioinformatics.

Tae Ho Kang
He received the B.S. degree in the department of Computer Communication Engineering from Howon University in 1999. And he received the M.S. and Ph.D. degree in department of Computer and Communication Engineering, Chungbuk National University in 2002 and 2007. He is working a Post-Doc in the School of Electrical and Computer Engineering, Chungbuk National University, Cheonju, South Korea where his research interests are the database system, bioinformatics.

Hak Yong Kim

He received the B.S. and M.S. degree in the departments of Agricultural Chemistry and Chemistry, Chungbuk National University in 1985 and 1987 respectively. He received the Ph.D. degree in the Molecular Cell Biology, University of Connecticut, U.S.A in 1994. He is now a professor in the Department of Biochemistry, Chungbuk National University, Cheongju, South Korea, where his research interests are the signal transduction, protein networks and biodynamic model.

Jae Soo Yoo

He received the B.S. degree in Computer Engineering in 1989 from Chunbuk National University, Chunju, South Korea. And he received the M.S. and Ph.D. degrees in Computer Science in 1991 and 1995 from Korea Advanced Institute of Science and Technology, Taejeon, South Korea. He is now a professor in the department of Computer and Communication Engineering, Chungbuk National University, Cheongju, South Korea, where his research interests are the database system, multimedia database,