에이전트 기반의 벼 기능 유전자 통합 데이터베이스

An agent-based integrated database for rice functional genomics

이기열*, 신문수*, 안수영**, 정동훈**, 안진흥**, 정무영*

* 포항공과대학교 산업경영공학과/제품생산기술연구소
** 포항공과대학교 생명과학과/생명공학연구센터

Abstract

In the field of rice research, insertional mutants have become a valuable resource for studies of gene function. However, a well-designed database yet in the area of rice functional genomics. The relevant data are widely distributed and independently managed by the individual research groups. Heterogeneous data format in the distributed database systems causes many problems related to redundancy and compatibility. In this research, integration of the distributed databases using agent technology is pursued. In particular, a data integration agent, an ontology agent, a comparison agent, and resource agents are designed, whereby the integrated database is maintained. Moreover a framework for the web-based information system, which provides information to biologists and permits biologists to add new data to the database, is proposed. To establish an interoperable data format, an XML-based data model is also developed adopting ontology concept.

Keyword: rice gene, functional genomics, agent technology, database integration, ontology

1. Introduction

기능 유전체학은 유전자와 유전자 산물의 기능에 관한 연구를 하는 학문 분야이다. 유전자의 기능을 알아내는 방법으로는 생물학적 접근법과 생화학적 접근법이 있다. 생물학적 접근법은 실험실에서 사용되는 모델 동물로부터 특정 유전자를 제거하여 생리작용이 변화하는 상태를 관찰하는 방법이다. 이 방법은 유전자의 발현으로 나타나는 생태를 알고자 할 때 주로 사용한다.

벼는 유전체의 크기가 작고 진정한 2n이며 유전체의 구성이 타 작물과 유사하기 때문에 단자엽 식물, 특히 작용 식물의 모델로 연구되고 있다. 이미 미국, 유럽, 일본 등을 중심으로 벼의 전체 유전체 서열 분석이 종료 단계에 이르고 있어 벼의 유전체 연구는 더욱 가속화 되고 있다. 현재 포항공과대학교 생명과학과의 식물 기능 유전체 연구실[8]에서는 knock-out 변이체를 이용하여 벼의 유전자 기능을 밝히는 연구를 진행하고 있다. 이 연구 결과를 TIGR(The Institute for Genomic Research)[14]와 같은 기존 데이터베이스에서 BLAST[1] 검색한 결과 자료를 이용하여 RISD(Rice T-DNA Insertion Sequence Database)[11]를 구축하고 있다. 다른 연구 단체에서도 벼의 유전자 기능을 밝히는 연구를 수행하며 개별 데이터베이스 시스템을 구축하여 데이터베이스가 분산되어 존재한다. 하지만 벼의 유전자를 연구하는 생물학자는 분산된 데이터베이스 시스템을 사용함으로써 다음과 같은 문제점에 직면하게 된다. 1) 연구 결과가 이질적인 방식으로 분산 저장됨으로써 동일한 결과임에도 불구하고 사용자에게 혼란을 야기한다. 2) 여러 단체에서 공통적으로 연구가 이루어진 유전자의 경우 사용자가 쉽게 찾을 수 있는 반면 특정 연구 단체에서 연구한 유전자의 경우 그 유전자에 대한 정보를 검색하기 힘들다. 이러한 문제를 해결하기 위해 벼 기능 유전체 데이터베이스의 통합이 필요하다.

본 연구에서는 에이전트 기술을 이용하여 분산된 데이트베이스를 통합한다. 에이전트 기술은 과거 시스템 통합에 성공적으로 적용된 사례가 있으며[3, 4, 5, 13], 특히 생물정보학 자원 통합에 적합하다고 알려져 있다[6]. 본 논문에서는 분산된 벼 기능 유전자 데이터베이스를 바탕으로 웹 기반의 정보 시스템(web-based information system)을 구축한다. 이 정보 시스템을 사용자 인터랙티브한 환경으로 구축함으로써 생물학자에게 사용 편의성을 제공하며, 데이터 활용도를 높인다. 또한 생물학자가 자신의 연구 성과를 등록하는 기능을 제공하여 연구 결과를 공유할 수 있도록 한다. 그리고 분산된 데이터베이스에 저장되어 있는 유전자의 정보를 서로 비교하여 제시함으로써 생물학자가 개별 웹 사이트를 검색하여 비교하는 수고를 줄인다.

2. Overall System Architecture

통합 데이터베이스 시스템의 전체적인 구조는 그림 1과 같다. 통합 데이터베이스를 구축하기 위한 Database Integration System(DIS), 외부 사용자가 통

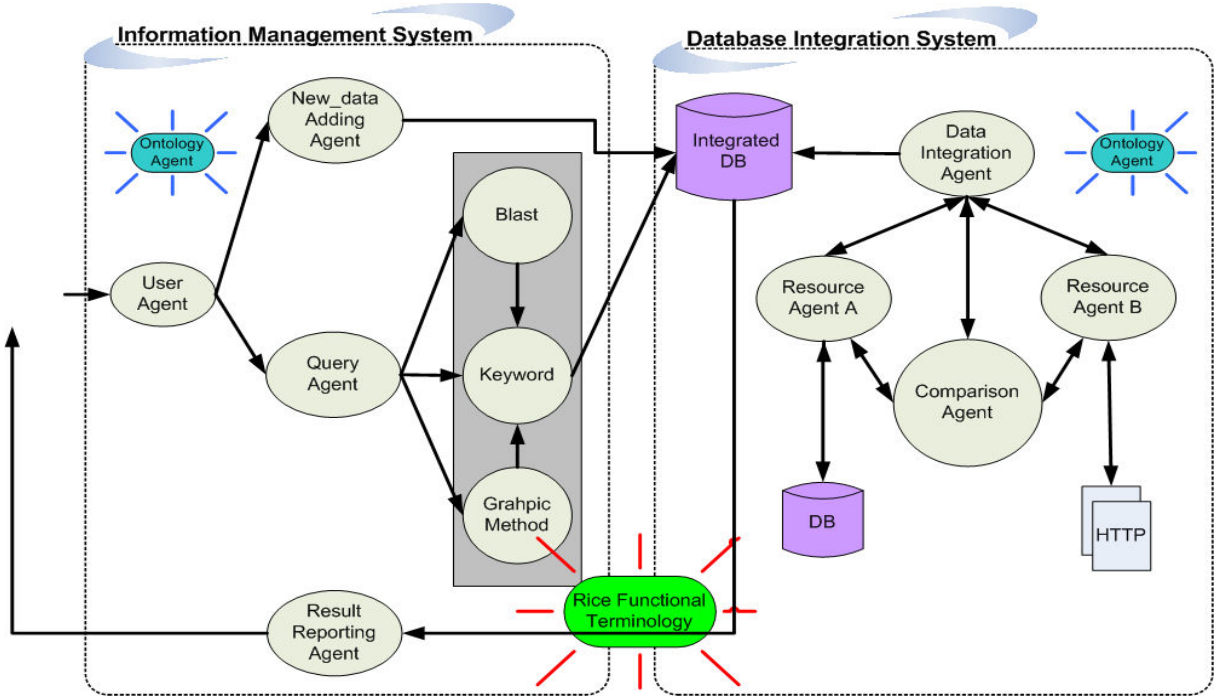

그림 1. Overall System Architecture

합 데이터베이스를 사용하기 위한 웹 기반의 정보 시스템인 Information Management System(IMS), 그리고 두 시스템 및 각 시스템의 Agent 사이의 의사소통을 위한 온톨로지로 전체 시스템이 구성된다.

2.1 Database Integration System (DIS)

DIS는 분산된 벼 기능 유전자 데이터베이스를 통합하기 위한 시스템이다. Kropbase[7], RAP-DB(Rice Annotation Project DataBase)]9], RMD(Rice Mutant Database)[10] 등의 외부 데이터베이스와 식물 기능 유전체 실험실 자체의 실험 결과로 구축하는 RISD를 통합함으로써 하나의 통합된 데이터베이스를 구축한다. 이 시스템은 DIA(Data Integration Agent), RA(Resource Agent), CA(Comparison Agent), OA (Ontology Agent)로 구성된다. 각 에이전트의 기능과 동작은 4장에서 살펴본다.

2.2 Information Management System (IMS)

IMS는 외부 사용자가 통합된 데이터베이스를 사용하기 위한 웹 기반 정보시스템이다. 이 시스템은 이용하여 외부 사용자는 통합 데이터베이스에서 정보를 검색, 추출하거나 새로운 정보를 등록할 수 있다. 이 시스템은 UA(User Agent), NAA(New-data Adding Agent), RRA(Result Reporting Agent), QA (Query Agent)의 4가지 에이전트로 구성된다. 특히 QA는 검색 방식에 따라 3가지의 하위 그룹을 가진다.

2.3 Rice Functioanl Terminology (RFT)

RFT는 IMS와 DIS의 내부 온톨로지와 분산된 데이터베이스의 온톨로지를 포함하는 최상위 온톨로지이다. IMS에는 시스템 사용자의 요청을 통합 데이터베이스에 전달하며, 데이터베이스의 검색 결과를 사용자에게 전달하는 과정에서의 불일치성을 줄이기 위한 온톨로지가 존재한다. DIS에는 분산된 데이터베이스와 통합 데이터베이스의 의사소통을 위한 온톨로지가 존재한다. 두 시스템의 온톨로지는 RFT를 기반으로 하며, 온톨로지의 변화 내용은 RFT에 전달한다. RFT는 BioMOBY[2]처럼 형식 또는 스키마에 상관없이 벼 유전자 데이터의 다양한 소스들과 상호작용이 가능한 클라이언트를 연결한다.

3. Structure of Database

통합 데이터베이스 시스템의 전체적인 구조는 그림 2와 같이 mutant lines, local system, web-system, phenotype, flanking sequence, reference 6개의 테이블을 가지며 모든 테이블은 line no.라는 primary key를 가진다.

- Mutant Lines: 이 테이블은 mutant line를 screening 한 사람, report gene, variety, seed information 등의 정보를 가진다. 이 테이블은 모든 mutant line의 line no.를 가지고 있다. report gene은 insertional mutant의 종류를 나타내는 것으로 T-DNA, TOS 17, DS 등이 있다. 그리고 분산된 데이터베이스의 위치를 알려주는 location 항목이 존재하여, local database와 web-based database를 구별할 수 있다.
- Local System and Web System: mutant lines의 위치 정보에 따라 두 가지 테이블로 구분된다. local system의 경우 RISD에서 제공하는 insertional mutant line의 정보를 보유하며, web system의 경우 mutant line정보를 제공하는 Kropbase, RAP-DB, RMD 등 웹사이트 정보와 report gene의 정보를 가진다.
- Phenotype: 생물체에 돌연변이가 발생하며 식물의 생태에 변화가 발생한다. Morphology는 돌연변이로 인한 식물의 형태 변화와 관련된 사진 자료

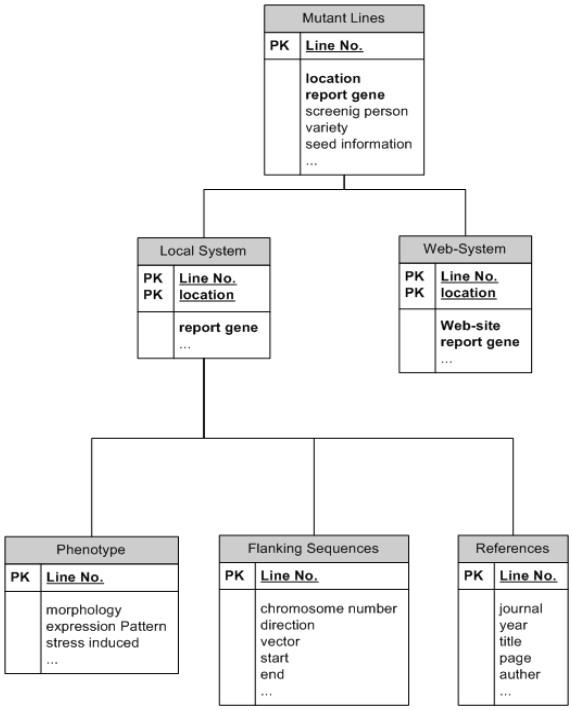

그림 2. Structure of Database

및 서술 내용이며, expression pattern은 실험을 통해 유전자의 발현 경향 및 식물에서의 발현 위치 정보이다. 이 두 정보는 유전자의 기능을 규명하기 위한 중요한 자료로 이용된다. 또한 돌연변이로 인해 식물이 화학적, 물리적 자극과 같은 외부 자극에 대한 내성가지기도 한다. Stress indused 항목은 돌연변이 식물이 외부 자극에 대한 내성의 변화를 관찰한 자료를 보관한다. 이 정보를 이용하여 식물체에서 내성을 향상시키는 유전자를 과다발현 시켜 외부 자극에 강한 식물을 개발할 수 있다. 포항공과대학교 식물 세포 생물학 연구실[15]에서는 이와 같이 식물에서 중금속에 내성을 가지는 유전자를 과다발현 시켜 환경정화 식물을 개발하고 있다.

- Flanking Sequences: 이 테이블은 insertional mutant가 벼 유전자에 삽입된 위치에 대한 상세한 정보를 저장한다. 삽입된 벼 유전자의 chromosome number와 삽입된 방향, insertional mutant의 시작 위치와 마지막 위치 등의 정보를 보관한다.
- References: 이 테이블은 insertional mutant와 관련된 논문 등의 문헌 정보 및 각종 참고자료를 저장한다. 특정 유전자의 돌연변이로 인해 벼의 생태에 변화가 발생하고, 이 현상을 실험을 통해 기작을 밝힌 논문이 있다면 논문이 발표된 저널 명칭, 발행연도, 저자, 제목 등을 저장한다.

4. Agents in Database Integration System

이 장에서는 DIS에서 사용하는 에이전트들의 역할과 동작 과정에 대하여 서술한다.

4.1 Data Integration Agent (DIA)

DIA는 데이터베이스 통합을 총괄하는 에이전트이다. 각 분산된 데이터베이스에서 정보를 종합하여 통합 데이터베이스 시스템에 저장한다. 우선 벼의 1번 유전자부터 scan하여 insertional mutant의 정보를 검색하고 mutant 정보가 없으면 RA에게 mutant 정보를 요청한다. 그리고 mutant 데이터의 update 시기를 확인한 후 update 여부를 결정한다. 그리고 최종적으로 다른 에이전트로부터 받은 메시지를 바탕으로 데이터베이스를 update한다. DIA는 여러 RA에게 mutant 정보를 요청할 수 있다.

4.2 Resource Agent (RA)

이 유형의 에이전트들은 각 데이터베이스를 개별적으로 접근하여 데이터베이스에서 insertinoal mutant의 정보를 검색한다. 검색된 결과가 존재할 경우 CA에 보낸다. 검색 결과가 없을 경우 DIA에 "Not Found" 메시지를 전달한다.

4.3 Comparison Agent (CA)

CA는 RA로부터 제공받은 insertional mutant의 유전자의 정보를 비교한다. RA로부터 받은 용어가 서로 상이한 경우 OA에 보내 통일된 용어를 제공받는다. DIS에 이미 존재하는 정보일 경우 DIA에 하위 데이터베이스가 "Not Updated" 메시지를 전달한다. 동일한 위치에 삽입된 insertional mutant일 경우 중복을 제거하여 하나의 데이터를 DIA에 제공하고, 서로 다른 데이터일 경우 모든 데이터를 DIA로 보낸다.

4.4 Ontology Agent (OA)

OA는 분산된 데이터베이스가 사용하는 하나 이상의 XML 기반의 온톨로지로 구성된다. 다른 에이전트로부터 용어 해석 요청이 들어오면 OA는 온톨로지에 찾고자 하는 용어가 있는지 검색한다. 검색된 경우 검색된 용어를 넘겨주고, 검색한 용어가 없는 경우 추론하여 적합한 용어를 각 에이전트에 제공한다. OA는 시스템의 공통 용어를 사용하고 다른 에이전트에게 그 용어를 제공함으로써 에이전트 사이에 발생할 수 있는 잠재적인 이질성(potential semantic heterogeneities)을 최소화한다. 그리고 OA는 도메인에 있는 어떤 에이전트와도 대화할 수 있다.

지금까지 살펴본 각 에이전트의 주요 기능을 표 1에 정리하였고, 그림 3은 DIS 전체 에이전트의 동작을 하나의 activity diagram으로 나타내었다.

표 1. 각 에이전트의 주요기능

Agent	Function
DIA	- request data from RAs - fetch the discovered data to DIS
RA	- search insertional mutant data from related database.
CA	- compare data delivered by RA. - send new data to DIA
OA	- match the terminology of distributed databases with the one of DIS

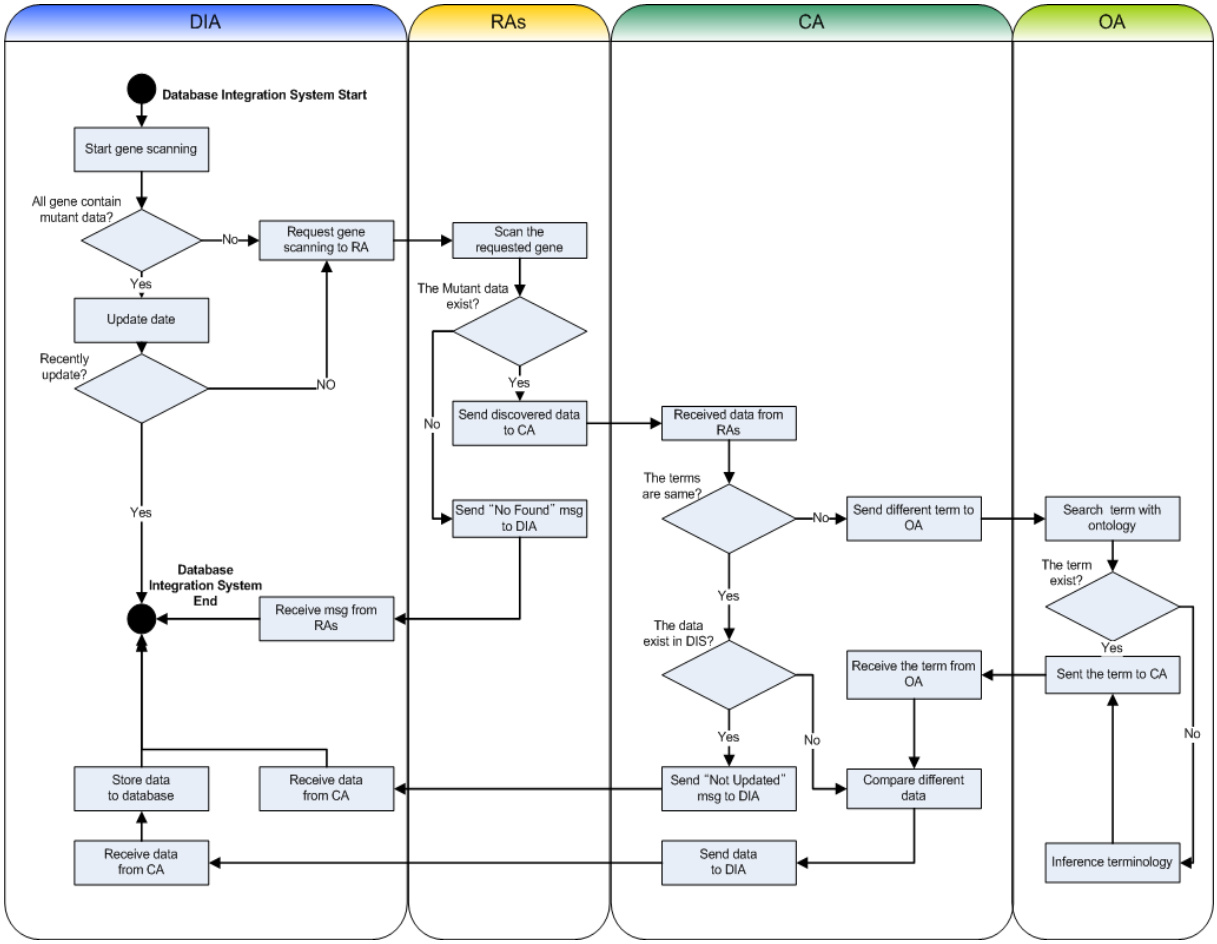

그림 3. Activity diagram of Database Integration System

5. Framework of Information Management System

이 장에서는 DIS을 기반으로 구축될 웹 기반의 정보시스템의 framework을 제시한다. IMS는 벼의 기능을 연구하는 생물학자에게 유용한 정보 제공을 목적으로 하며, 다양한 검색 가능을 제공한다. 또한 사용자의 연구 성과를 등록할 수 있는 기능을 제공한다. IMS를 구축하기 위해서는 시스템 사용자를 위한 온톨로지가 필요하다.

- BLAST Search: 유전자의 기능을 규명하는 생물학자들은 유전자 matching program인 BLAST tool[1]에 익숙하다[16]. DNA sequence를 이용하는 BLASTn, amino acid sequence를 이용하여 검색하는 BLASTp 등을 적용한다. BLAST 검색 결과는 입력한 sequence와 비교하여 우선순위가 높은 순서로 총 10개의 line no.를 candidate으로 제공한다.
- Graph Search: GBrowse(Generic-Genome-Browser)[12]를 이용하여 각 유전자에 통합된 데이터베이스의 insertional mutant 위치를 보여준다. 사용자는 나타난 그림을 클릭함으로써 insertional mutant의 정보를 확인할 수 있다. Blast 검색하였을 경우 검색된 결과에 따라 유전자의 위치를 파악할 수 있지만 insertional mutant의 정보가 나타나지 않을 경우가 있다. 이 경우 graph search를 이용함으로써 사용자가 알고자 하는 유전자 주변의 insertional mutant 정보를 확인할 수 있다.
- Keyword Search: 이 방법은 사용자가 insertional mutant의 line no.를 알고 있을 경우 BLAST 검색, graph 검색보다 유용하게 사용할 수 있다. 통합 데이터베이스에서 line no.는 primary key이기 때문에 사용자가 찾고자 하는 정보를 빠르게 확인할 수 있다.
- Add New Data: 이 기능을 이용하여 사용자가 자신의 연구 성과를 데이터베이스에 등록할 수 있다. 사용자의 연구 성과를 등록할 경우 통합 데이터베이스의 온톨로지를 제공하여 시스템의 통일성을 유도한다. 그러나 이 기능을 구현하기 위해서는 사용자의 연구 성과가 사실임을 인증할 수 있는 장치가 필요하다.
- Ontology: IMS는 DIS의 정보 제공을 목적으로 하기 때문에 사용자의 용어를 DIS의 용어로 바꾸어야 한다. 따라서 모든 시스템을 총괄하는 온톨로지(RFT)가 필요하다. RFT는 사용자와 IMS, IMS과 DIS의 온톨로지뿐만 아니라 DIS내부의 OA의 온톨로지도 포괄한다.

6. Conclusion

DIS(database integration system)은 웹 기반의 정보

시스템인 Information Management System(IMS)을 구축하기 위한 시스템이다. 본 논문에서는 DIS을 구축하기 위한 구조에 대하여 중점적으로 살펴보았다. DIS은 DIA(database integration system), RA(resource agent), CA(comparison agent), OA (ontology agent)의 네 종류의 에이전트로 구성되어 있으며 각 에이전트의 역할과 동작 과정을 살펴보았다. 그리고 DIS을 바탕으로 구성한 IMS의 framework를 제시하고 있다. 추후에는 제시된 framework을 바탕으로 IMS의 구현에 관한 연구가 필요하며, 특히 연구자가 등록한 데이터의 사실 여부를 검증하는 절차에 관한 연구가 필요하다.

References

[1] Altschul S. F., Madden T. L., et al.,(1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **25,** 3389-3402

[2] BioMOBY http://biomoby.org

[3] Bayardo, R., Bohrer, W., Cichocki, A., Fowler, j., Helal, A., Kashyap, V., Ksiezyk, T., Martin, G., Nodine, M., Rusinkiewicz, M., Shea, R., Unnikrishnan, C., Unruh, A. and Woelk, D.,(1997), InfoSleuth: agent-based semantic integration of information in open and dynamic environments. In *Proceedings of the ACM SIGMOD International Conference on Management of Data(SIGMOD 1997)*, May 13-15, Tuscon, AZ.ACM press

[4] Carey, M., Hass, L., Schwarz, P., Aryo, M., Cody, W., Ragin, R. Rlickner, M. Lahiewski, A., Niblack, W., Petkovic, D., Thomas II, J., Willians, J., and Wimmers, E.,(1995), *Towards heterogeneous miltimedia information systems: the garlic approach. In Proceedings of the Fifth International Workshop on Research Issues in Data Engineering-Distributed Object Management,* Taipei, Taiwan, March 6-7, *IEEE Computer Society press*, 124-131.

[5] Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Vassal, K., and Widom, J.,(1997), The TSI<<IS approach to mediation: data models and languages, *Journal of Intelligent Information Systems*, **8**(2), 117-132.

[6] Karasavvas, K.A. et al.,(2004), Bioinformatics integration and agent technology, *J.Biomed. Inform.*, 37(3), 205-219

[7] Kropbase, http://kropbase.snu.ac.kr

[8] Plant Functional Genomics Lab http://www.postech.ac.kr/life/pfg

[9] Rice Annotation Project DataBase, http://rapdb.lab.nig.ac.jp

[10] Rice Mutant Database, http://rmd.ncpgr.cn

[11] Rice T-DNA Insertion Sequence Database, http://141.223.132.44/pfg/index.php

[12] Stein L. D., Mungall C, Shu S., Caudy M., Mangone M., Nickerson E., Stajich J. E., Harris T. W., Arva A., Lewis S.,(2002) The generic genome browser: a building block for a model organism system database, *Genome Res.* 1599-1610

[13] Sycara, K., Paulucci, M., Van Velsen, M., and Giampapa, J.,(2001), The RETSINA MAS infrastructure. Technical Report *CMU-RI-TR-01-05*, Robotics Institute, Carnegie Mellon

[14] The Institute for Genomic Research, http://www.tigr.org

[15] The Plant Cell Biology Lab. http://pcb.postech.ac.kr

[16] Yan Z., Guyang M. H., and Liping W.,(2002), UniBlast: a system to filter, cluster, and display BLAST results and assign unique gene annotation, *BIOINFORMATICS APPLICATIONS NOTE,* **18**(9), 1268-1269