마이크로어레이 데이터 공유 시스템

Microarray Data Sharing System

윤지희, 홍동완, 이종근
한림대학교 컴퓨터공학과

Jee-Hee Yoon(jhyoon@hallym.ac.kr), Dong-Wan Hong(dwhong@hallym.ac.kr),
Jong-Keun Lee(jeikei@hallym.ac.kr)

요약

최근, 마이크로어레이 실험 데이터의 품질과 재생산성에 대한 신뢰도가 증가하고 있어 마이크로어레이 데이터의 공유 및 활용에 대한 요구가 급속히 증가하고 있다. 그러나 공개되어 있는 국내, 외 마이크로어레이 데이터는 실험 방식, 플랫폼 등에 따라 서로 다른 데이터 항목과 포맷을 가지므로 데이터의 실제적 접근 및 활용이 어려운 상황이다. 본 논문에서는 실험 플랫폼, 데이터 포맷, 정규화 기법, 분석 방식 등이 서로 다른 기존의 마이크로어레이 데이터를 효율적으로 검색, 공유, 통합할 수 있는 마이크로어레이 데이터 공유 시스템을 제안한다. 제안된 시스템은 웹 서비스 기반 기술을 이용하여 분산된 마이크로어레이 데이터를 통합하며, 각 사이트의 사용자는 UDDI를 통하여 검색한 데이터를 표준 MGED 기반의 공통 데이터 구조로 자동 변환하여 다운 받을 수 있다. 정의된 공통 데이터 구조는 IDF, ADF, SDRF, EDF로 구성되어 다양한 구조의 마이크로어레이를 통합할 수 있는 템플릿 역할을 수행하며, MAGE-ML, MAGE-TAB, XML Schema 문서로 저장할 수 있다. 또한 제안된 시스템의 자동 데이터 제출기, 파일 관리자 등은 마이크로어레이 데이터 공유를 위한 다양한 부가 기능을 제공한다.

■ 중심어 : | 마이크로어레이 데이터 | 데이터 공유 | 웹 서비스 기술 | 데이터 표준 포맷 |

Abstract

Improved reliability of microarray data and its reproducibility lead to recent increment in demand of data sharing and utilization among laboratories, but house-keeping and publicly opened microarray experimental data can hardly be accessed and utilized since they are in heterogeneous formats according to the various experimental methods and microarray platforms. In this paper, we propose a microarray sharing method which can easily retrieve and integrate microarray data from different experiment platforms, data formats, normalization methods, and analysis methods. Our system is based on web-service technology. The biologists of each site are able to search UDDI(Universal Description, Discovery, and Integration) registry, and download microarray data with common data structure of standard format recommended by MGED(Microarray Gene Expression Databases) society. The common data structure defined in this paper consists of IDF(Investigation Design Format), ADF(Array Design Format), SDRF(Sample and Relationship Format), and EDF(Expression Data Format). These components play role as templates to integrate microarray data with various structure and can be stored in standard formats such as MAGE-ML, MAGE-TAB, and XML Schema. In addition, our system provides advanced tools of automatic microarray data submitter and file manager to manipulate local microarray data efficiently.

■ keyword : | Microarray Data | Data Sharing | Web-service | Data Standard Format |

* 이 논문은 2007년도 정부재원(교육인적자원부 학술연구조성 사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음 (KRF-2007-531-D00019).

접수번호 : #090421-003
접수일자 : 2009년 04월 21일
심사완료일 : 2009년 06월 23일
교신저자 : 윤지희, e-mail : jhyoon@hallym.ac.kr

Ⅰ. 서 론

마이크로어레이 데이터의 공유를 위하여 국내, 외 유명 저널에서는 실험에 사용된 마이크로어레이 데이터를 의무적으로 공용 데이터베이스에 공개하도록 규정하고 있다[1]. 또한 최근에는 MAQC(Microarray Quality Control)[2] 프로젝트의 수행 결과를 바탕으로 마이크로어레이 실험 데이터의 품질(quality)과 재생산성(reproducibility)에 대한 신뢰도가 증가하고 있어 금후 마이크로어레이 데이터 저장소의 규모 및 그 활용이 급속히 증가할 것으로 예상되고 있다.

그러나 마이크로어레이 실험은 one-channel 칩과 two-channel 칩 등의 실험 방식에 따라 실험 결과가 다르며, 플랫폼에 따라 실험 데이터가 서로 다른 포맷을 가지므로 공유를 위한 실제적 접근 및 활용이 어렵다. 또한, 대표 마이크로어레이 저장소인 GEO(Gene Expression Omnibus)[3]나 ArrayExpress[4] 등의 경우에도 관련 데이터의 상호 참조가 어려워, 생물학 실험자는 각 사이트를 개별 방문하여 데이터를 수집, 통합하여야 하며, 이 경우 각 시스템의 검색 방법 및 데이터의 획득 방법이 상이하여 생물학 실험자에게는 상당히 부담스러운 일이 될 수 있다.

마이크로어레이 데이터에 대한 표준화 작업은 MGED(Microarray Gene Expression Data) 협회[5]를 중심으로 이루어지고 있으며, 4개의 워킹 그룹 중 하나인 MIAME(Minimum Information About a Microarray Experiment)[6]는 check-list를 기초로 한 마이크로어레이 데이터 표현을 제안하였다. 최근에는 MIAME 기반의 MAGE-ML(Microarray Gene Expression Markup Language)[7], SOFT(Simple Omnibus Format in Text)[8], MINiML(MIAME Notation in Markup Language)[9], MAGE-TAB[10] 등의 데이터 교환 포맷이 사용되고 있으나, 아직 실용화, 범용화 단계에 이르지 못하고 있다.

본 논문에서는 실험 플랫폼, 데이터 포맷, 정규화 기법, 분석 방식 등이 서로 다른 기존의 마이크로어레이 데이터를 효율적으로 검색, 공유, 통합할 수 있는 마이크로어레이 데이터 공유 시스템을 제안한다. 본 시스템의 데이터 통합 범위는 각 생물학 실험실에서 보유하고 있는 마이크로어레이 데이터와 대표적인 마이크로어레이 저장소의 데이터로 한다. 제안하는 시스템은 웹 서비스 기반 기술을 이용하여 분산된 마이크로어레이 데이터를 통합하며, 이질의 마이크로어레이 실험 데이터를 공유하기 위하여 표준 MGED 기반의 공통 구조(common structure)를 정의한다. 정의된 공통 구조는 실험에 대한 일반적인 정보를 제공하는 IDF(Investigation Design Format), 어레이 디자인을 기술하는 ADF(Array Design Format), 샘플, 어레이, 실험 데이터를 연결하는 SDRF(Sample and Relationship Format), 마이크로어레이 실험의 발현 값을 표현하는 EDF(Expression Data Format)로 구성되는데 이들은 다양한 구조의 마이크로어레이를 통합할 수 있는 템플릿 역할을 하며, UDDI 등록기(Universal Description Discovery and Integration registry) 구현에 사용되어 이질의 마이크로어레이 실험 데이터의 통합 검색을 가능하게 한다. 각 사이트의 사용자는 UDDI를 통하여 검색한 데이터를 공통 데이터 구조로 자동 변환하여 다운 받을 수 있으며, 본 시스템에서 개발한 자동 프로파일러가 그 역할을 담당한다. 또한 본 시스템의 자동 데이터 제출기, 파일 관리자 등은 마이크로어레이 데이터 공유를 위한 다양한 부가 기능을 제공한다.

본 논문의 구성은 다음과 같다. 제 Ⅱ 장에서는 관련 연구로서 공용 마이크로어레이 데이터베이스의 현황을 살펴보고, 마이크로어레이 실험 데이터의 품질 평가 및 표준화 방식에 대하여 간단히 설명한다. 제 Ⅲ 장에서는 본 논문에서 제안하는 마이크로어레이 데이터 공유 방식과 시스템의 전반적인 구조에 대하여 기술한다. 제 Ⅳ 장에서는 시스템 구현을 위한 각 구성 모듈의 개발 방법과 특징을 사용자 인터페이스와 함께 자세히 설명한다. 제 Ⅴ 장에서는 결론과 향후 연구 과제를 기술한다.

II. 관련 연구

1. 공용 마이크로어레이 데이터베이스

마이크로어레이 실험 데이터를 저장하는 대표적인 저장소로 GEO, ArrayExpress, SMD(Stanford Microarray Database)[11] 등을 들 수 있다. 1999년 NCBI에서 개발을 시작한 GEO는 플랫폼, 샘플, 시리즈 별로 데이터를 관리하며, 데이터 셋을 압축된 BLOB 형태로 관리하고 있다. 2002년 EBI(European Bioinformatics Institute)에서 개발을 시작한 ArrayExpress는 마이크로어레이 실험 데이터 저장소의 역할 외에 그래프 기반 분석 툴 등의 다양한 데이터 분석 기능을 제공하고 있다. 스탠포드 대학을 중심으로 개발된 SMD는 최근 Array XML working group[12]과 함께 데이터 교환 표준 포맷에 관한 연구를 진행 중이다. [표 1]은 대표적 마이크로어레이 데이터베이스와 그 특징을 보인다. 그러나 이들 공용 데이터베이스는 개별적으로 마이크로어레이 데이터의 저장, 검색 기능을 지원하고 있으나 관련 데이터의 상호 검색이 어렵고, 검색 환경도 매우 제한적이다. 즉, 이들 분산 이질의 마이크로어레이 데이터베이스로부터 마이크로어레이 데이터를 통합 검색할 수 있는 검색 환경이 제공되지 못하고 있는 실정이다.

2. 마이크로어레이 데이터 품질 평가

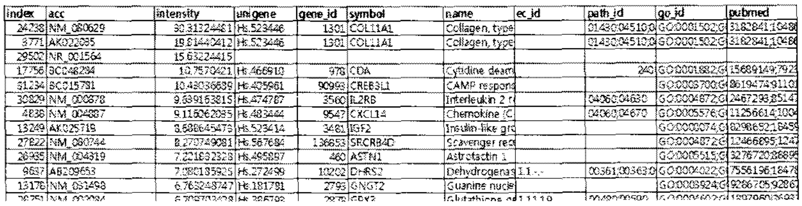
일반적으로 같은 목적 및 실험 환경, 동일 실험자가 진행한 반복적인 마이크로어레이 실험에서도 각 실험 결과는 차이를 보인다. 최근 미국 식품의약국(FDA: Food and Drug Administration)에서 마이크로어레이 실험 결과의 품질 보증을 위한 MAQC 프로젝트를 수행하였다[2]. MAQC 프로젝트에서는 원거리에 있는 서로 다른 3개의 사이트에서 주어진 동일 실험을 수행하여 마이크로어레이 플랫폼의 성능, 각 실험실의 숙련도, 다양한 실험 절차의 장점 등을 평가하였다. 프로젝트 수행 결과, 참고 문헌 [2]는 각 사이트에서 신뢰할 수 있는 유전자 탐지가 이루어졌음을 보고하고 있으며, 이는 마이크로어레이 실험 데이터의 품질과 재생산성에 대한 신뢰도를 재평가할 수 있는 새로운 결과로 주목 받고 있다.

3. 마이크로어레이 실험 데이터의 표준화 현황

마이크로어레이 실험 데이터의 표준화 작업은 데이터의 표준 규정을 제안하는 것과 컴퓨터의 데이터 포맷을 개발하는 것으로 구분된다. 2007년 5월 MGED 협회에서 제안한 MIAME v2.0[13]은 다음 6개의 요소로 구성된다. ① 각 혼성화(hybridization)단계에서의 Raw 데이터 처리 방법, ② 혼성화 후의 최종 처리 데이터 구조, ③ 실험 요소와 그 값들을 포함하는 샘플의 주석(annotation) 처리, ④ 실험 원리(예. one-channel 또는 two-channel)에 따라 샘플 데이터 간의 관계를 정의하는 실험 디자인, ⑤ 프로브 시퀀스나 데이터베이스 accession number 등을 포함하는 어레이 디자인에 대한 주석 정보, ⑥ 실험 방식이나 정규화 등의 데이터 처

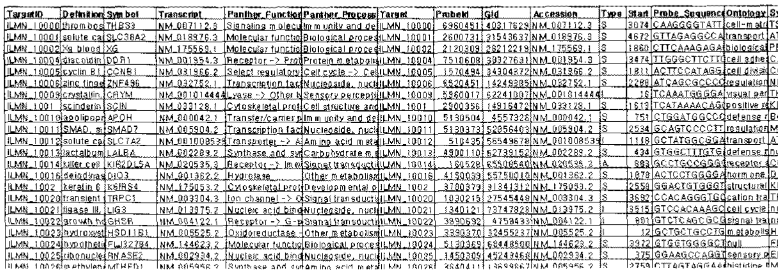
표 1. 마이크로어레이 데이터베이스[3][4][11]

Database	Organization	Description	URL
ArrayExpress	European Bioinformatics Institute (EBI)	public data deposition and public queries (coming soon)	http:// www.ebi.ac.uk/arrayexpress
Dragon	Johns Hopkins University	public queries	http://pevsnerlab.kennedykrieger.org/dragon.htm
ExpressDB	Harvard University	public queries of E. coli and yeast data	http://arep.med.harvard.edu/ExpressDB
GeneX	NCGR	local installation, public data deposition, and public queries of E. coli and yeast data	http://genex.sourceforge.net http://www.ncgr.org
GEO	National Center for Biotechnology Information (NCBI)	public data deposition and public queries	http://www.ncbi.nlm.nih.gov/geo
SMD	Stanford University	local installation and public queries	http://smd-www.stanford.edu



index	acc	intensity	unigene	gene_id	symbol	name	ec_id	path_id	go_id	pubmed
24238	NM_080629	90.31324461	Hs.523446	1301	COL11A1	Collagen, type		01430;04510;0	GO:0001502;G	3182841;1048
3771	AK022085	19.81440412	Hs.523446	1301	COL11A1	Collagen, type		01430;04510;0	GO:0001502;G	3182841;10486
29502	NR_001564	15.63224415								
17756	BC048284	10.7570421	Hs.466910	978	CDA	Cytidine deam		240	GO:0001882;G	15689149;7923
61234	BC015781	10.48036689	Hs.405961	90993	CREB3L1	CAMP respons			GO:0003700;G	8619474;9110
30829	NM_000878	9.639163815	Hs.474787	3560	IL2RB	Interleukin 2 r		04060;04630	GO:0004872;G	2467298;8514
4838	NM_004887	9.115062085	Hs.483444	9547	CXCL14	Chemokine (C		04060;04670	GO:0005576;G	11256614;1004
13249	AK025719	8.688645479	Hs.523414	3481	IGF2	Insulin-like gro			GO:0000074;G	8298652;18459
27822	NM_080744	8.270749081	Hs.567684	136853	SRCRB4D	Scavenger rec			GO:0004872;G	12466895;1247
26935	NM_004319	7.221882328	Hs.495897	460	ASTN1	Astrotactin 1			GO:0005515;G	3276720;8869
9637	AB209653	7.080185925	Hs.272499	10202	DHRS2	Dehydrogenas	1.1.-.-	00361;00363;0	GO:0004022;G	7556196;18478
13176	NM_051498	6.763248747	Hs.181781	2793	GNGT2	Guanine nucle			GO:0003924;G	9286705;9286

(a) Operon (OpArray Human Whole Genome 35K: two-channel 칩)



TargetID	Definition	Symbol	Transcript	Panther_Function	Panther_Process	Target
ILMN_10000	thrombos	THBS3	NM_007112.3	Signaling molecu	Immunity and de	ILMN_10000
ILMN_10001	solute ca	SLC38A2	NM_018976.3	Molecular functio	Biological proces	ILMN_10001
ILMN_10002	Xg blood	XG	NM_175569.1	Molecular functio	Biological proces	ILMN_10002
ILMN_10004	discoidin	DDR1	NM_001954.3	Receptor -> Prot	Protein metaboli	ILMN_10004
ILMN_10005	cyclin B1	CCNB1	NM_031966.2	Select regulatory	Cell cycle -> Cel	ILMN_10005
ILMN_10006	zinc finge	ZNF496	NM_032752.1	Transcription fac	Nucleoside, nuc	ILMN_10006
ILMN_10009	crystallin,	CRYM	NM_00101444	Lyase -> Other	Sensory percepti	ILMN_10009
ILMN_1001	scinderin	SCIN	NM_033128.1	Cytoskeletal prot	Cell structure and	ILMN_1001
ILMN_10010	apolipopr	APOH	NM_000042.1	Transfer/carrier p	Immunity and de	ILMN_10010
ILMN_10011	SMAD, m	SMAD7	NM_005904.2	Transcription fac	Nucleoside, nuc	ILMN_10011
ILMN_10012	solute ca	SLC7A2	NM_00100853	Transporter -> A	Amino acid meta	ILMN_10012
ILMN_10013	lactalbum	LALBA	NM_002289.2	Synthase and sy	Carbohydrate m	ILMN_10013
ILMN_10014	killer cell	KIR2DL5A	NM_020535.3	Receptor -> Imm	Signal transducti	ILMN_10014
ILMN_10016	deiodinas	DIO3	NM_001362.2	Hydrolase	Other metabolis	ILMN_10016
ILMN_1002	keratin 6	K6IRS4	NM_175053.2	Cytoskeletal prot	Developmental p	ILMN_1002
ILMN_10020	transient	TRPC1	NM_003304.3	Ion channel -> O	Signal transducti	ILMN_10020
ILMN_10021	ligase III,	LIG3	NM_013975.2	Nucleic acid bind	Nucleoside, nuc	ILMN_10021
ILMN_10022	growth ho	GHSR	NM_004122.1	Receptor -> G-p	Signal transducti	ILMN_10022
ILMN_10023	hydroxyst	HSD11B1	NM_005525.2	Oxidoreductase -	Other metabolism	ILMN_10023
ILMN_10024	hypotheti	FLJ32784	NM_144623.2	Molecular functio	Biological proces	ILMN_10024
ILMN_10025	ribonucle	RNASE2	NM_002934.2	Nucleic acid bind	Nucleoside, nuc	ILMN_10025

TargetID	ProbeId	Gid	Accession	Type	Start	Probe_Sequence	Ontology	Syn
ILMN_10000	6960451	40317629	NM_007112.3	S	3074	CAAGGGGTATT	cell-matr	TSP
ILMN_10001	2600731	31543637	NM_018976.3	S	4672	GTTAGAGGCCA	transport	ATA
ILMN_10002	2120309	28212219	NM_175569.1	S	1860	CTTCAAAGAGA	biological	PBD
ILMN_10004	7510608	39927631	NM_001954.3	S	3474	TTGGGCTTCTT	cell adhe	CAK
ILMN_10005	1570494	34304372	NM_031966.2	S	1811	ACTTCCATAGG	cell divisi	CCN
ILMN_10006	6520451	14249385	NM_032752.1	S	2288	ATCAGCGCCCC	regulation	NIZF
ILMN_10009	5960017	62241007	NM_00101444	I	16	TCAAATGGGGA	visual per	THB
ILMN_1001	2900356	14916472	NM_033128.1	S	1613	TCATAAAACAG	positive re	KIAA
ILMN_10010	5130504	4557326	NM_000042.1	S	751	CTGGATGGCCC	defense r	BG2
ILMN_10011	5130373	52856403	NM_005904.2	S	2534	GCAGTCCCCTT	regulation	MAD
ILMN_10012	510435	56549678	NM_00100853	I	1118	GCTATGGCGGA	transport	ATR
ILMN_10013	4900110	62739152	NM_002289.2	S	434	GTGGCTTTGTG	defense r	null
ILMN_10014	150528	65508540	NM_020535.3	A	983	GCCTGCCGGGG	receptor	CD1
ILMN_10016	4150039	55750010	NM_001362.2	S	1878	ACTCCTGGGGA	hormone	D3;
ILMN_1002	3780379	31341312	NM_175053.2	S	2558	GGACTGTGGGT	structural	KRT
ILMN_10020	1030215	27545449	NM_003304.3	S	3692	CCACAGGGTGG	cation tra	TRP
ILMN_10021	1340121	73747828	NM_013975.2	I	3515	GTCCACAAAGC	cell cycle	null
ILMN_10022	3990692	4758433	NM_004122.1	I	801	GTCTCAGCGCG	signal tra	null
ILMN_10023	3390370	32455237	NM_005525.2	I	12	GCTGCTGCCTG	metabolis	HDL
ILMN_10024	5130369	68448500	NM_144623.2	S	3372	GTGGTGGGGCT	null	FLJ
ILMN_10025	1450309	45243468	NM_002934.2	S	375	GGAAGCCAGGT	sensory p	EDN

(b) Illumina (Human Whole Genome 6_V2.0 R2: one-channel 칩)

그림 1. 마이크로어레이 실험 데이터의 예

리에 대한 프로토콜이다. 또한 MGED 협회는 데이터 교환 포맷으로 MAGE-ML을 채택하고 있다. MAGE-ML은 XML (eXtensible Markup Language)에 기반을 둔 마이크로어레이 실험 데이터의 전용 데이터 포맷으로 대부분의 생물학적 정보들을 포함하고 있다. 그러나 MAGE-ML의 복잡성으로 인하여 사용의 실효성을 거두지 못하고 있는 실정이다. 이에 따라 데이터 표현 및 교환을 쉽게 할 수 있는 데이터 포맷의 개발이 이어지고 있으며, SOFT, MINiML, MAGE-TAB 등이 대표적 포맷이라 할 수 있다.

III. 마이크로어레이 데이터 공유 방식

본 장에서는 마이크로어레이 데이터의 효율적인 공유 방식을 제안한다. 제 III.1절에서는 마이크로어레이 데이터의 특성을 분석하고, 제 III.2절에서는 마이크로어레이 데이터 공유를 위한 공통 데이터 구조를 정의한다. 제 III.3절에서는 제안하는 마이크로어레이 공유 시스템의 전반적인 구조를 보인다.

1. 마이크로어레이 데이터 구조

마이크로어레이 실험 데이터는 실험 방법 및 실험 환경 등에 대한 기초 정보와 발현 값 등에 대한 실험 결과로 구성된다. 그러나 마이크로어레이 실험은 동일한 샘플에 대한 동일 실험을 수행하는 경우에도 실험 조건에 따라 그 결과가 다를 수 있으며, 특히 다른 회사의 마이크로어레이를 사용하는 경우에는 제작 회사마다 프로브의 특성 및 종류, 분석 플랫폼 등이 모두 다르기 때문에 실험 결과가 다르게 주어진다. 마이크로어레이 칩은 Affymetrix[14], Illumina[15], Agilent [16], Operon[17], Digital genomics[18] 등 다수의 제조회사에 의하여 제공되고 있다. 이 들 제조회사에서 제공하는 마이크로어레이 칩은 크게 발현 값의 비율을 측정하는 two-channel 칩과 발현 데이터 값을 그대로 사용하는 one-channel 칩으로 구분된다.

다음의 [그림 1]은 동일 샘플에 대한 마이크로어레이 실험을 서로 다른 2종의 플랫폼(Operon, Illumina)에서 수행하여 얻어진 실험 데이터의 예를 보인다 (여기에서 보인 데이터는 실험의 기초 정보에 해당하는 헤더 정보의 일부를 나타낸다). 이와 같이 실험 결과는 일반적으로 텍스트 혹은 엑셀 파일의 형태로 제공되며, 동일 샘플을 이용한 동일 실험의 경우에도 각 제조회사로부터 각각 다른 데이터의 항목과 구조로 표현된 실험 결과를 제공받게 된다.

서로 다른 플랫폼에서 얻어지는 실험 데이터를 상호

```
<?xml version="1.0" encoding="euc-kr"?>
<xs:schema xmlns="http://jasmine.ce.hallym.ac.kr/IDF_DTD" elementF
ormDefault="qualified" targetNamespace="http://jasmine.ce.hallym.ac.kr/I
DF_XSD" xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="IDF_Format">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Status" />
        <xs:element ref="Protocol" />
        <xs:element ref="Investigator" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="Status">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Title" />
        <xs:element ref="Sample_Type" />
        <xs:element ref="Source_Name" />
        <xs:element ref="Organism" />
      .....
```

(a) IDF 정의 XML Schema

```
<xs:element name="ADF_Format">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="MetaRow" />
      <xs:element ref="MetaCol" />
      <xs:element ref="Column" />
      <xs:element ref="Row" />
      <xs:element ref="Reporter_Identifier" />
      <xs:element ref="Reporter_Name" />
      <xs:element ref="Reporter_Comment" />
      <xs:element ref="Reporter_BioSequence_DatabaseEntry" />
      <xs:element ref="Reporter_BioSequence_Type" />
      <xs:element ref="Reporter_BioSequence_Polymer_Type" />
      <xs:element ref="Reporter_BioSequence" />
      <xs:element ref="Reporter_Group" />
      <xs:element ref="Reporter_Control_Type" />
    </xs:sequence>
  </xs:complexType>
  .....
```

(b) ADF 정의 XML Schema

```
<xs:element name="SDRF_Format">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Sources" />
      <xs:element ref="Samples" />
      <xs:element ref="Extracts" />
      <xs:element ref="Labeled_Extracts" />
      <xs:element ref="Hybridizations" />
      <xs:element ref="Scans" />
      <xs:element ref="Array_Data_Files" />
      <xs:element ref="Normalizations" />
      <xs:element ref="Derived_Array_Data_File" />
    </xs:sequence>
  </xs:complexType>
</xs:element>
  .....
```

(c) SDRF 정의 XML Schema

```
<xs:element name="EDF_Format">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Control" />
      <xs:element ref="Treatment" />
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="Control">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="1" maxOccurs="unbounded" ref="Sa
mple" />
    </xs:sequence>
  </xs:complexType>
  .....
```

(d) EDF 정의 XML Schema

그림 2. IDF, ADF, SDRF, EDF 정의 XML Schema 문서

비교하기 위하여, 세계적으로 사용 빈도가 높은 다음 4가지 종류의 플랫폼(Affymetrix Illumina, Agilent, Operon)에서 제공되는 실험 데이터의 구조 및 특성을 비교하였다. 다음 [표 2]에 각 실험 데이터의 항목명을 상호 비교한 결과의 일부를 보인다. 이 예에서 보이는 바와 같이 실험 결과는 각각 다른 항목명을 갖는 서로 다른 구조로 제공되며, 같은 의미를 가지는 항목도 각기 항목명으로 표기되어 있다. 예를 들어, "유전자명(gene name)"을 뜻하는 항목에 대해서 Illumina에서는 "Definition"이라는 항목명을 사용하며, Operon이나 Agilent에서는 "Name"이라는 항목명을 사용하고 있다. 이러한 실험 데이터의 구조적 차이는 실험 데이터를 공유하여 분석하고자 하는 사용자들에게 많은 어려움을 줄 수 있다.

2. 마이크로어레이 데이터의 공통 구조

표 2. 플랫폼에 따른 마이크로어레이 실험 데이터의 항목 비교

Illumina	Operon	Affymetrix	Agilent
TargetID	Block	Probe Set Name	Block
Definition	Column	Stat Pairs	Column
Symbol	Row	Stat Pairs Used	Row
Transcript	Name	Signal	Name
Panther_Function	ID	Detection	ID
Panther_Process	X	Detection p-value	X
Target	Y	Stat Common Pairs	Y
ProbeId	Dia.	Signal Log Ratio	Dia.
Gid	F635 Median	Signal Log Ratio Low	F635 Median
Accession	F635 Mean	Signal Log Ratio High	F635 Mean
Type	F635 SD	Change	F635 SD
Start	B635 Median	Change p-value	F635 CV
Probe_Sequence	B635 Mean	Positive	B635
Ontology	B635 SD	Negative	B635 Median
Synonym	% > B635+1SD	Pairs	B635 Mean
...	...	...	...

서로 다른 구조를 갖는 마이크로어레이 실험 데이터를 공유, 활용하기 위하여 다음과 같이 공통 데이터 구조를 정의한다. 우선, 각 플랫폼에서 산출되는 데이터의 구성 요소를 분류한다. 각 플랫폼의 마이크로어레이 실험 데이터로부터 유전자 서열 데이터베이스의 정보와 연관되는 유전자 주석(gene annotation) 정보, 실험에 사용된 정보를 표현하는 샘플 주석(sample annotation) 정보, 어레이 형태로 유전자의 발현 정보를 표현하는 유전자 발현 데이터를 분류한다. 다음, 이들 데이터를 IDF, ADF, SDRF, EDF의 4가지 포맷으로 정의한다. 각 포맷의 특징은 다음과 같다.

(1) IDF : 실험에 대한 일반적인 정보를 제공한다. 실험자의 이름, 연락 정보, 논문(bibliographic) 참조, 실험 방식(protocol) 등을 포함하고 있다.

(2) ADF : 스프레드시트나 스프레드시트 셋 안의 어레이 디자인을 설명하기 위한 것이다. 어레이 디자인을 설명하기 위해 다음과 같은 세 가지 정보 필요하다. ① 어레이 상의 각 스팟에서의 핵산(nucleic acid)의 위치를 나타내는 어레이 정보(feature on the array), ② 어레이 상에 위치한 시퀀스 정보를 나타내는 리포터 시퀀스(reporter sequence), ③ 유전자와 엑손(exon)과 같은 생물학 개체를 측정하기 위한 리포터 시퀀스의 집합인 합성 요소(composite element)이다. ADF로 정의되는 데이터는 마이크로어레이 실험 장비 자체에서 실험자에게 제공할 발현 값, 분석 정보 등을 추출, 처리하는데 사용되는 메타 데이터라 할 수 있다.

(3) SDRF : 마이크로어레이 레이아웃을 전체적으로 설명하기 위해서는 어레이의 특징, 리포터 시퀀스, 합성요소와 이것들 사이의 관계를 제공하여야 한다. SDRF는 샘플, 어레이, 실험의 객체 간 연결 정보를 제공한다.

(4) EDF : 바이너리나 ASCII 파일로 제공되는 Affymetrix의 CEL, Agilent의 TXT, GenePix의 GPR 파일과 같은 Raw 데이터 파일이나 정규화 등의 공정을 거친 Data Matrix 파일 등의 실제 실험 결과 데이터를 포함한다.

본 연구에서 제안하는 공통 구조의 IDF, ADF, SDRF는 MGED의 표준안을 따르도록 정의하였으며, EDF는 실제 발현 데이터의 구조 정의를 위하여 추가로 정의하였다. 기존의 생물학 실험 장비에서 나온 데이터들은 일반적으로 마이크로소프트 엑셀이나 ASCII 형식의 텍스트 데이터 포맷을 가지고 있으나, 최근 이들 장비의 업그레이드 과정에서 내부 데이터의 표현이 표준 XML 형태로 변환해 가고 있는 추세에 있다. 또한 마이크로어레이 실험 데이터 중 발현 데이터도 점차 XML 데이터 형태로 표현, 교환되고 있는 추세이다.

이와 같이 정의된 공통 구조는 MGED 협회의 표준 데이터 교환 포맷으로 채택된 MAGE-ML로 표현된다. MAGE-ML은 현재 문서 구조 정의에 DTD (Document Type Definition)를 사용하고 있다. 그러나 본 연구에서는 최근의 표준화, 실용화 경향을 참조하여 XML 문서의 구조 정의를 위하여 XML Schema를 이용한다. 다음의 [그림 2]에 IDF, ADF, SDRF, EDF의 XML Schema 문서의 일부를 보인다.

3. 시스템 구조

바이오 데이터를 통합하기 위한 기존의 시스템 구현 방식으로 링크 기반 통합, 뷰 기반 통합, 데이터웨어하우징 통합 방식 등을 들 수 있다. 그러나 이 들 방식은 데이터 소스의 손실 및 변경이 일어날 경우 데이터 검색이 불가능하거나 변경된 데이터의 실시간 반영이 어려운 단점이 있다[19]. 이러한 단점들을 극복할 수 있는 방안으로서 웹서비스 기반의 데이터 통합 방법이 주목받고 있다. 웹서비스는 플랫폼에 상관없이 시스템의 상호운영이 가능한 표준기술로서 다음과 같이 서비스 등록자(publisher), 서비스 사용자(consumer), 서비스 저장소(registry)로 구성된다. 서비스 등록자는 자신이 제공하고자 하는 데이터를 통합 시스템에 업로드하지 않고, 자신이 보유한 데이터의 기본 정보만을 서비스 저장소에 등록하며, 서비스 사용자는 서비스 저장소에서 정보 검색을 수행하여 자신이 원하는 서비스를 찾게 된다. 서비스 사용자는 이와 같이 검색한 해당 서비스를 서비스 제공자(등록자)에게 직접 요청하게 되며, 서비스 등록자는 이에 응답하여 서비스 사용자에게 원하는 정보를 제공하게 된다.

본 시스템에서는 웹 서비스 기술을 기반으로 하여 분산 이질의 마이크로어레이 데이터를 통합한다. 본 연구에서 제안하는 시스템의 전반적인 구조를 [그림 3]에 보인다. 그림에 보이는 바와 같이 웹 상에는 각 실험실에서 생성된 수많은 소규모의 마이크로어레이 데이터가 존재하며, 또한 GEO나 ArrayExpress 등의 공용 마이크로어레이 저장소에 저장된 대규모의 마이크로어레이 데이터가 존재한다. 본 시스템에 의한 마이크로어레이 데이터 검색 및 공유/활용 과정을 간단히 설명하면 다음과 같다.

우선 생물학 실험자는 보유하고 있는 실험 데이터를 제공하기 위하여 UDDI 등록기를 통해 UDDI 저장소에 웹 서비스를 등록하게 된다 ([그림 3]-②의 과정). 일반적으로 마이크로어레이 실험의 특성을 분류, 파악하기 위해서는 실험에 관한 세부 정보가 필요하며, 이는 일반적으로 제공되는 실험 데이터의 헤더 정보만으로는 불충분하다. 따라서 본 시스템에서는 UDDI 등록기를 통해 실험에 관한 기본 정보를 입력받아 UDDI의 등록 정보로 사용한다.

이제 마이크로어레이 실험 데이터를 검색하고자 하는 일반 사용자는 UDDI 저장소를 통해 정보를 검색하게 되며 ([그림 3]-③의 과정), 검색 결과로서 원하는 데이터 저장소의 위치인 종점(end-point)을 획득하게 된다 ([그림 3]-④의 과정). 다음, 사용자는 이와 같이 획득한 종점에 직접 데이터를 요청하게 되며([그림 3]-⑤의 과정), 실 데이터를 보유하고 있는 종점의 서비스 등록자로부터 실험 데이터를 얻을 수 있게 된다. 그러나 이와 같이 제공되는 마이크로어레이 실험 데이터는 이질의 구조/포맷을 가지고 있어 이를 직접 활용하기 어렵다. 따라서 본 시스템에서는 이 문제를 해결하기 위하여 검색된 실험 데이터를 제 III.2절에서 제안한 공통 데이터 구조로 자동 변환하여 사용자에게 제공하고 있으며, 이와 같은 데이터의 변환 작업은 자동 프로파일러가 담당한다 ([그림 3]-⑥의 과정).

이와 같이 원격 사이트에서 자주 검색하게 되는 실험 데이터는 매번 그 사이트에서 검색하지 않고, 자체적으로 각 생물학 연구실의 내부 데이터베이스에 저장하여 두는 것이 편리하다. 하지만, 일반 생물학 전문가들에게 자체적으로 데이터베이스를 구축하여 활용하는 일은 부담스러운 일이 될 수 있다. 따라서 본 시스템에서는 사용자들의 편의성을 고려하여 이와 같은 마이크로어레이 실험 데이터를 손쉽게 저장, 관리할 수 있는 파일 관리자(file manager)를 구현하여 제공한다([그림 3]-⑦의 과정). 파일 관리자는 자동 프로파일러를 통하여 생성된 IDF, ADF, SDRF, EDF 구조의 XML Schema 문서 데이터를 로컬 컴퓨터에 손쉽게 저장, 관리하도록 설계, 구현되어 있으며, XQuery를 표준 질의어로 사용할 수 있다.

또한 사용자의 편의성을 고려한 시스템의 부가 기능으로서 본 시스템에서는 실험 데이터 자동 제출 기능을 제공한다. 최근 주요 생물학 분야의 저널에서는 실험에 사용된 마이크로어레이 데이터를 의무적으로 공용 데이터베이스에 공개하도록 규정하고 있으나, 생물학 실험자가 처음으로 공용 데이터베이스에 실험 데이터를 제출할 경우, 각 제출 단계에서 입력해야 하는 사항을 적절히 기입하는 작업은 일반적으로 쉽지 않다. 따라서 본 시스템에서는 생물학 실험자들이 공용 데이터베이스에 실험 데이터를 제출할 경우, 각 제출 단계마다 참조하여 정보를 입력할 수 있도록 각 단계의 가이드 파일을 자동으로 생성, 참조하도록 하는 자동 제출기(auto submitter)를 제공한다 ([그림 3]-⑧의 과정). 자동 제출기의 가이드 파일 생성을 위한 실험의 기초 정보는 마이크로어레이 실험 데이터와 UDDI 등록 정보로부터 자동 추출함을 원칙으로 하여, 사용자로 하여금 같은 정보를 여러 번 입력해야 하는 불편함을 최소화한다.

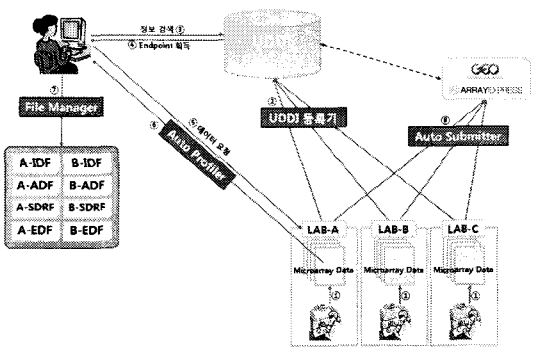


그림 3. 웹 서비스 기반의 시스템 구조도

Ⅳ. 시스템 구현

본 장에서는 마이크로어레이 공유 시스템의 구현 방법에 대하여 기술한다. 시스템을 구성하는 주요 모듈의 기능과 구현 방법에 대하여 설명하고, 또한 실행 결과의 예를 이용하여 그 유용성을 보인다.

1. UDDI 저장소의 서비스 등록 및 검색

제안된 시스템에서는 마이크로어레이 데이터의 상호 공유를 위하여 서로 다른 구조의 실험 데이터를 공통 데이터 구조로 변환한다. 그러나 실험 데이터만을 이용하여 공통 데이터 구조의 데이터를 생성하는 경우, 실험에 관한 정보가 부족하여 실제 데이터의 공유 및 활용에 어려움을 겪게 된다.

이 문제를 해결하기 위하여 본 시스템에서는 서비스 등록 과정에서 필요한 정보를 생물학 실험자로부터 부가적으로 입력 받고 있다. 즉, 생물학 실험자가 UDDI 저장소에 실험 정보를 등록하는 과정에 있어, 실험 데이터에 포함되어 있지 않은 실험에 관한 기본 정보를 수집하게 되며, 수집된 추가 정보와 실험 데이터를 기반으로 공유 데이터 구조의 데이터를 생성한다.

다음의 [그림 4]는 본 시스템에서 제공하는 UDDI 등록기의 사용자 인터페이스 화면을 나타낸다. 이 예는 [그림 1]-(a)에서 보인 Operon의 마이크로어레이 실험 데이터를 UDDI 저장소에 등록하는 과정을 나타낸다. 사용자는 우선 [그림 4]에 보이는 바와 같이 UDDI 등록기를 통하여 실험 데이터 파일과 URL 주소를 입력하게 되며, 다음 실험 데이터(헤더 정보)에 포함되어 있지 않는 실험에 관한 기본 정보 (Investigation Title, Organism, Manufacturer, Protocol, 사용자 정보 등)를 입력한다. 이와 같이 입력된 UDDI의 등록 정보는 UDDI 서비스 검색을 위한 기본 정보로 활용될 뿐 아니라, 외부 사이트에 의한 데이터 전송 요구 시, 자동 프로파일러에서 공통 데이터 구조를 따르는 데이터 생성에 활용된다.

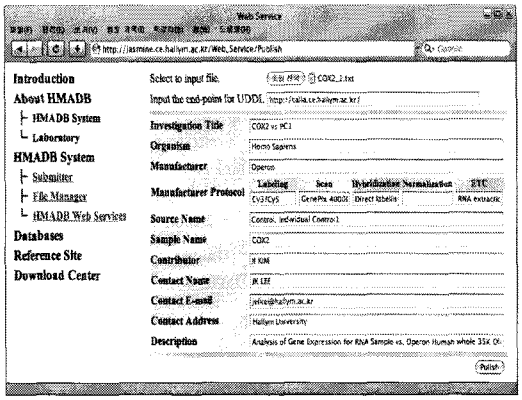


그림 4. UDDI 등록기 사용자 인터페이스의 예

한편, 마이크로어레이 실험에 대한 정보를 얻고자 하는 일반 사용자는 UDDI 저장소를 통해 정보 검색을 수행하게 된다. 사용자가 검색하고자 하는 서비스에 관련된 키워드를 입력하면, 시스템은 UDDI 저장소의 등록 정보를 대상으로 관련 서비스를 검색하게 되며, 검색 결과로서 서비스 관련 정보와 데이터 저장소의 위치를 반환하게 된다. 다음, 사용자는 검색 결과를 바탕으로 데이터 저장소에 데이터를 요청하게 되며, 실 데이터를 보유하고 있는 서비스 등록자로부터 원하는 형태의 실험 데이터를 얻을 수 있게 된다.

본 시스템에 의한 서비스 검색 과정을 예를 이용하여 설명하면 다음과 같다. [그림 5]는 UDDI 저장소의 서비스 검색을 위한 사용자 인터페이스 화면을 나타낸다. 유전자 'CYCLOOXYGENASE 2; COX2'와 관련된 마이크로어레이 실험 데이터의 검색 과정을 보인 예로서, 생물학 실험자가 검색 키워드로 'COX2'를 입력하여, 등록된 자료들이 검색된 결과를 나타낸다. UDDI 저장소에는 마이크로어레이에 관한 다양한 실험 정보가 등록 정보로서 보관되어 있다. 따라서 입력된 키워드에 대한 관련 정보 검색은 이들 전체 등록 정보를 대상으로 포괄적으로 수행되며, 그 결과로서 검색된 서비스에 대한 간단한 설명 정보와 함께 데이터 저장소가 연결되고 있다. 다음의 [그림 6]은 데이터 수집 과정을 나타낸다. 기존의 시스템에서는 검색된 결과 중 선택된 하나의 데이터에 대한 정보수집만을 요청할 수 있으나, 본 시스템에서는 생물학 실험자가

수집하고자 하는 마이크로어레이 데이터를 다중 선택할 수 있다. [그림 6]은 [그림 5]의 검색된 결과에 대하여 2개의 서비스를 선택한 경우의 결과 화면을 나타낸다. 이와 같이 사용자는 다중 검색된 결과의 상호 비교가 가능하며, 이 들 결과를 원하는 포맷으로 변환하여 열람하거나 다운로드 할 수 있다.

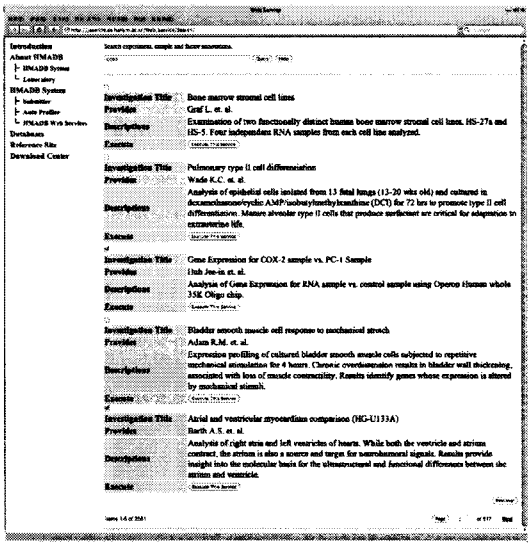


그림 5. UDDI 검색 결과 화면

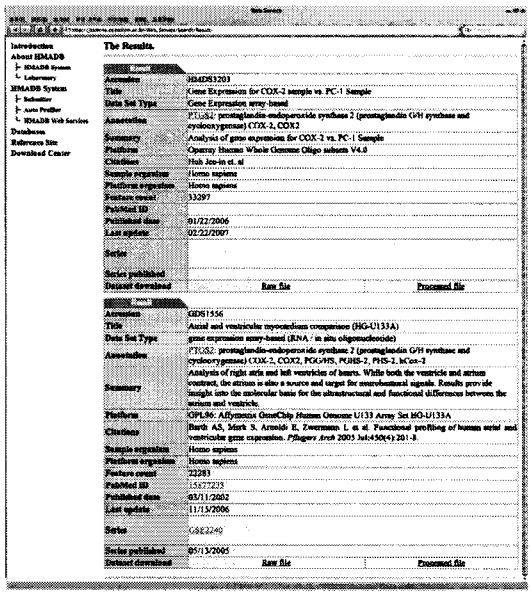


그림 6. 다중 선택된 마이크로어레이 데이터 수집

2. 공통 구조 프로파일링

생물학 실험자가 보유하고 있는 이질의 마이크로어레이 데이터를 공통 구조의 마이크로어레이 데이터로 자동 변환하는 역할은 자동 프로파일러가 제공한다. 자동 프로파일러는 검색된 마이크로어레이 실험 결과로부터 일반 정보 (IDF), 어레이 관련 데이터 (ADF), 각 데이터의 연결 정보를 포함한 데이터 (SDRF), 실험 데이터 (EDF)를 자동으로 추출하여 표준 문서로 저장한다.

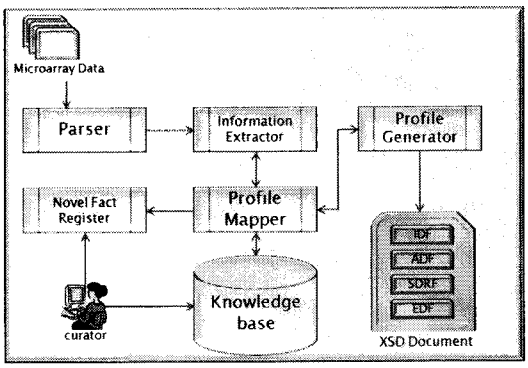


그림 7. 자동 프로파일러의 구성 요소

[그림 7]에 자동 프로파일러의 내부 구성 모듈과 데이터의 흐름을 보인다. 일반적인 마이크로어레이 실험 업체들은 원시 실험 결과와 정규화 및 정제 과정을 거친 실험 결과를 텍스트 형태의 아스키 파일이나 마이크로소프트 엑셀 파일 형태로 제공한다. 자동 프로파일러에 입력된 이 들 마이크로어레이 실험 데이터는 파서를 통하여 실험 데이터 정보와 주석 정보로 파싱된다. 정보 추출기(information extractor)는 이 들 정보로부터 실험에 대한 기초 정보, 샘플 정보, 외부 데이터베이스와의 연계를 위한 인식 번호(ID number) 등을 나타내는 항목 데이터를 추출하여 프로파일 맵퍼(profile mapper)로 보낸다.

다음, 프로파일 맵퍼는 추출된 정보를 IDF, ADF, SDRF, EDF의 카테고리로 구분하여 프로파일 생성자(profile generator)로 전송하는 역할을 수행한다. 추출된 각 항목 데이터를 공통 구조의 각 카테고리로 분류하기 위하여 지식베이스를 구축하여 활용하며, 다음

IDF:= {Investigation Title, Experimental Design, Experimental Design Term Source REF, Experimental Design Term Accession Number, Experimental Factor Name, Experimental Factor Type, Experimental Factor Term Source REF, Experimental Factor Term Accession Number, Person Last Name, Person First Name, Person Mid Initials, Person Email, Person Phone, Person Fax, Person Address, Person Affiliation, Person Roles, Person Roles Term Source REF, Person Roles Term Accession Number, Quality Control Type, Quality Control Term Source REF, Quality Control Term Accession Number, Replicate Type, Replicate Type Term Source REF, Replicate Type Term Accession Number, Normalization Type, Normalization Type Term Source REF, Normalization Type Term Accession Number, Date of Experiment, Public Release Date, PubMed ID, Publication DOI, Publication Author List, Publication Title, Publication Status, Publication Status Term Source REF, Publication Status Term Accession Number, Experiment Description, Protocol Name, Protocol Type, Protocol Description, Protocol Parameters, Protocol Hardware, Protocol Software, Protocol Contact, Protocol Term Source REF, Protocol Term Accession Number, SDRF File, Term Source Name, Term Source File, Term Source Version, Comment[<user-defined tag>]}

ADF:= {MetaColumn, MetaRow, Column, Row, Reporter Identifier, Reporter Name, Reporter BioSequence DatabaseEntry [database_code], Reporter BioSequence Type, Reporter BioSequence Polymer Type, Reporter BioSequence [Actual Sequence], Reporter Group [role], Reporter Control Type}

그림 8. 지식베이스의 예

[그림 8]에 IDF, ADF 카테고리 분류를 위하여 사용되는 지식베이스의 일부의 예를 보인다. 그러나 제 Ⅲ.1절에서 설명한 바와 같이 서로 다른 플랫폼에서 얻어지는 실험 데이터는 각각 다른 항목명을 가지며, 또한 같은 의미의 항목이라도 각각 다른 항목명으로 표기되는 경우가 많다. 따라서 본 시스템에서는 각 플랫폼에 따라 각기 다르게 제공되는 항목명을 표준 항목명으로 자동 변환하는 지식베이스를 구축하여 활용하고 있다. 또한 지식 베이스에서 검색되지 않은 새로운 항목을 위하여, 이들 정보를 표준 항목으로 변환하기 위한 항목 등록기(novel fact register)를 제공하고 있다. 이와 같이 등록된 새로운 항목정보는 지식 베이스에 등록되어 차후 활용된다. 마지막으로 프로파일 생성자는 프로파일 맵퍼에서 전송된 데이터를 입력 받아 공통 데이터 구조의 형식을 따르는 IDF, ADF, SDRF, EDF 문서를 생성한다. 생성된 문서는 MAGE-ML, MAGE-TAB, XML Schema 문서 등으로 선택하여 저장할 수 있다.

자동 프로파일링 과정을 [그림 1]-(a)의 Operon에서 산출된 마이크로어레이 실험 데이터의 예를 이용하여 간단히 설명하면 다음과 같다. 우선, 실험 데이터로부터 파서를 통해 "index", "acc", "intensity", "symbol", "name" 등의 항목 정보가 파싱되게 된다. 이렇게 파싱된 항목 정보는 프로파일 맵퍼를 통해 "index"는 "Probe ID"로, "acc"는 "Accession Number"로, "name"은 "Gene Name"이라는 표준 항목명으로 변경된다. 다음, 프로파일 생성자는 이들 정보와 UDDI 서비스 등록 시 입력 받은 실험의 기초 정보를 이용하여 공통 데이터 구조 형식을 따른 IDF, ADF, SDRF, EDF 문서를 생성하게 된다. 다음의 [그림 9]에 자동 프로파일러에서 생성된 실험 데이터의 일부(IDF, ADF 데이터)를 예로 보인다.

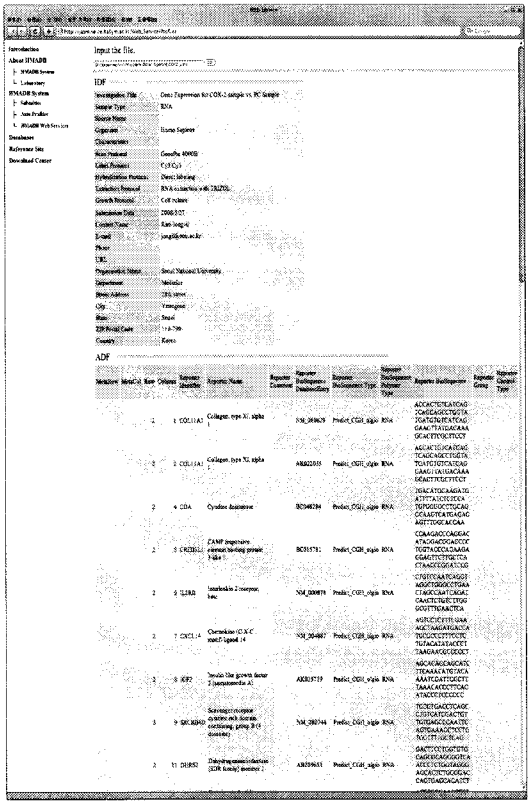


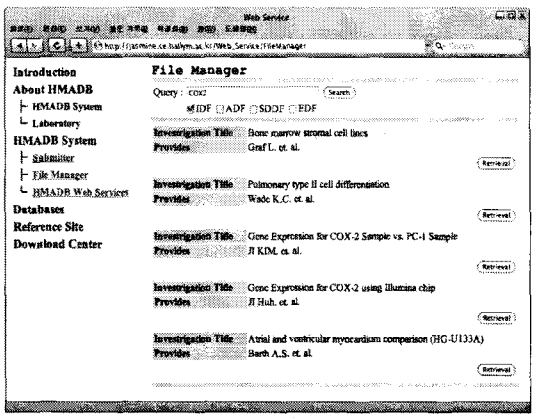
그림 9. 자동 프로파일러에서 생성된 공통 구조 데이터

3. 마이크로어레이 파일 관리자

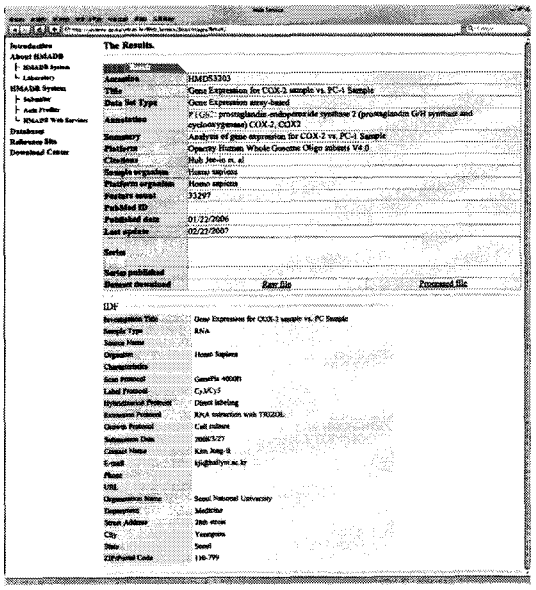
본 시스템에서는 마이크로어레이 실험 데이터를 효율적으로 저장, 관리할 수 있는 파일 관리자를 제공한

다. 파일 관리자는 원본 실험 데이터와 자동 프로파일러를 통해 생성된 IDF, ADF, SDRF, EDF 구조의 XML Schema 문서 데이터를 생물학 실험실의 로컬 컴퓨터에 저장, 관리한다. 본 시스템을 통해 다운로드한 데이터는 공통 데이터 구조의 XML 파일 형태로 관리되며, XML 파일을 검색하기 위하여 XQuery를 표준 질의어로 사용할 수 있다. 그러나 XQuery는 일반 생물학 사용자가 사용하기에 쉽지 않으므로 모든 질의는 사용자 친화적 GUI 환경에서 메뉴 선택 방식 등으로 입력되며, 이들 질의는 내부적으로 XQuery 형식으로 자동 변환되어 해석된다.

다음의 [그림 10]에 파일 관리자에 의한 정보 검색 과정의 예를 보인다. [그림 10]-(a)는 검색 키워드로서 'COX2'를 입력하여 관련 정보를 검색하는 사용자 인터페이스 화면을 나타낸다. 이 때 사용자는 원하는 정보 파일의 종류(IDF, ADF, SDRF, EDF)를 선택하여 검색할 수 있으며, 여기에서는 IDF 파일을 지정한 경우를 나타낸다. 이 예에서는 현재 5개의 검색 결과가 제시되고 있으며, 검색된 결과 중에서 원하는 데이터를 선택하면 그 마이크로어레이 데이터 파일에 대한 일반 정보와 선택된 공통구조 데이터를 보여준다. 검색 결과 중 세 번째 목록을 선택한 경우의 정보 검색 결과를 [그림 10]-(b)에 보인다.



(a) 검색 화면



(b) 검색 결과 화면

그림 10. 파일 관리자의 정보 검색 예

4. 마이크로어레이 데이터 자동 제출기

마이크로어레이 실험 데이터를 GEO 혹은 Array Express 등의 대표 마이크로어레이 데이터 저장소에 제출하고자 하는 경우, 여러 단계에 거쳐 입력하여야 하는 내용이 복잡하여 일반 실험자에게 실험 데이터 제출은 큰 부담이 될 수 있다[3]. 예를 들어 실험 데이터를 GEO에 제출하는 경우, GEO는 각 실험자가 보유하고 있는 실험 데이터 파일의 필드 명을 표준화하여 입력받기를 원한다. 예를 들어 생물학 실험자가 보유하고 있는 데이터 필드명이 'acc', 'intensity', 'unigene' 등으로 구성되어 있는 경우, 이들은 GEO에 실험 데이터 제출 시 각각 'GenBank accession number', 'Log ratio Cy3/Cy5', 'Unigene ID' 등의 GEO 표준 정의 필드명으로 변환, 입력되어야 한다.

이와 같은 문제점을 보완하기 위한 방법으로 본 시스템에서는 다음과 같은 자동 제출기를 설계, 구현하였다. 자동 제출기에서는 실험 데이터와 UDDI 등록 시 입력받은 기초 정보를 활용하여 안내 페이지를 자동 생성한다. 또한 각 플랫폼에 대하여 실험 데이터 파일의 필드에 대한 공용 마이크로어레이 데이터베이스의 표준 필드명 및 설명(description)을 지식베이스로 구축하여,

이들 정보를 기반으로 안내 페이지를 생성한다. 다음의 [그림 11]은 [그림 1]-(a)의 Operon 마이크로어레이 실험 데이터를 본 시스템의 자동 제출기가 생성한 안내 페이지의 예를 나타낸다. 자동 제출기는 Step 1의 해당 파일을 선택한 후, 지식베이스 등 실험의 기본 정보를 활용하여 Step 2에서 Step 4까지의 내용을 자동 생성한다. 즉, 생물학 실험자는 이와 같은 자동 제출기에서 생성된 안내 페이지를 참조하며 단계 별로 항목을 입력하여 실험 결과를 간편히 GEO에 제출할 수 있게 된다.

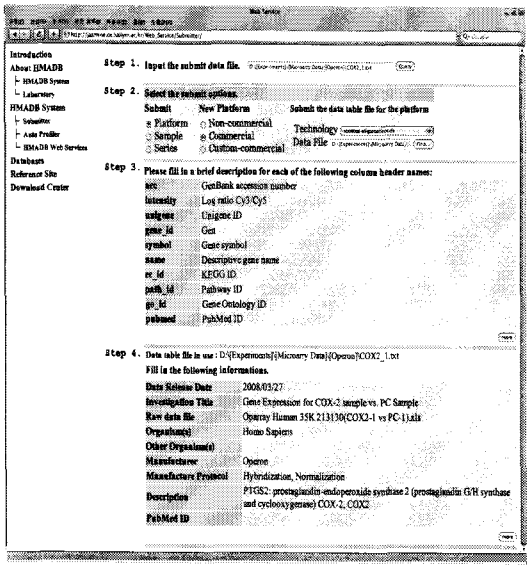


그림 11. 자동 제출기에서 생성된 안내 페이지

V. 결론 및 향후 연구과제

마이크로어레이 데이터 공유를 위한 다양한 표준화 작업이 진행되고 있으나, 현재까지 실효성을 거두지 못하고 있는 실정이다. 본 연구에서는 마이크로어레이 데이터가 표준화되지 못한 상황에서 데이터 공유를 가능하게 하는 효율적인 방법론을 제시하였다. 또한 본 연구에서는 이질 마이크로어레이 데이터를 위한 새로운 공통 데이터 구조를 제시하고 그 유용성을 보임으로써, 제안된 공통 데이터 구조를 새로운 표준 마이크로어레이 구조로서 제안하는 의미를 갖는다.

제안하는 시스템은 웹 서비스 기반 기술을 이용하여 분산된 마이크로어레이 데이터를 통합하며, 각 사이트의 사용자는 UDDI를 통하여 검색한 데이터를 표준 MGED 기반의 공통 데이터 구조로 자동 변환하여 다운받을 수 있다. 또한 각 실험실에서 수집한 실험 데이터를 효율적으로 관리하는 파일 관리자를 개발하였으며, 대표 마이크로어레이 저장소에 실험 데이터 제출 시 입력해야 하는 내용을 자동으로 생성하여 주는 자동 제출기를 개발하였다.

본 시스템은 웹 서비스의 응용 서버로 Sun Application Server Platform Edition 9.0을 사용하였으며, 개발 툴로 J2SE 5.0과 JSP(Java Server Page)를 활용하였다. 현재 본 시스템은 프로토타입의 개발 완료 후, 시험 운용 중으로 생물학 실험자들의 실질적인 사용 경험을 통한 의견을 시스템 개발 및 운영에 반영하고 있다. GEO와 ArrayExpress 등의 대표 마이크로어레이 저장소에서는 웹 서비스 기반 연결 서비스 함수를 공개하겠다고 이미 1-2년 전에 밝힌 바 있다. 그러나 아직 연결 서비스 함수가 공개되지 않고 있어 이 들 사이트와의 직접 연결이 불가능한 실정이다. 향후 단 기간 내에 대표 저장소에서 이들 연결 서비스 함수가 제공될 것으로 예상되며, 이를 기반으로 하여 본 시스템은 보다 효율적인 정보 공유를 지원할 수 있을 것으로 기대된다.

또한 제안된 정보 공유 시스템을 이용한 마이크로어레이 실험의 효율적인 정보 통합 및 분석 방법에 관한 연구를 수행 중에 있으며, 향후 특정 질병에 대한 표적 진단에의 활용을 목적으로 연구를 수행 하고 있다.

참 고 문 헌

[1] M. P. Charles, "Show me the data!," Nature Genet., Vol.29. p.373, 2001.

[2] MAQC Consortium, "MAQC project shows inter and intraplatform reproducibility of gene expression measurements," Nature Biotechnology, Vol.24, No.9, pp.1151-1161, 2006.

[3] Barrett, "NCBI GEO: mining millions of expression profiles - database and tools," Nucleic Acids Research, Vol.33, No.1, pp.562-566, 2005.

[4] Brazma, "Arrayexpress: a public database of gene expression data at EBI," Nucleic Acids Research, Vol.31, pp.68-71, 2003.

[5] C. A. Ball and A. Brazma, "MGED Standards: Work in Progress," OMICS, Vol.10, No.2, pp.138-144, 2006.

[6] Brazma, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," Nature Genet., Vol.29, No.4, pp.365-371, 2001.

[7] Spellman, "Design and implementation of microarray gene expression markup language (MAGE-ML)," Genome Biology, Vol.3, No.9, RESEARCH0046.1-0046.9, 2002.

[8] Y. Yi, C. Li, C. Miller, and A. L. George Jr., "Strategy for encoding and comparison of gene expression signatures," Genome Biology, Vol.8, 2007.

[9] D. Field and S. Sansone, "A Special Issue on Data Standards," OMICS: A Journal of Integrative Biology, Vol.10, No.2, pp.84-93, 2006.

[10] Rayner, "A simple spreadsheet- based, MIAME-supportive format for microarray data: MAGE-TAB," BMC Bioinformatics, Vol.7, pp.489-507, 2006.

[11] Gollub, "The Stanford Microarray Database: data access and quality assessment tools," Nucleic Acids Research 2003, Vol.31, pp.94-96, 2003.

[12] A. Hayes, "The Second International Meeting on Microarray Data Standards, Annotations, Ontologies and Databases," Wiley & Sons, Vol.17, pp.238-240, 2000.

[13] http://www.mged.org/Workgroups/MIAME/miame_2.0.html

[14] http://www.affymetrix.com/index.affx

[15] http://www.illumina.com/

[16] http://www.agilent.com/

[17] http://www.operon.com/default.aspx

[18] http://cms.gobizkorea.com/fileroot/media/8/8327/media/intro.htm

[19] Krestyaninova, "MolPAGE Data Warehouse," ISMB, Poster I13, 2007.

저 자 소 개

윤 지 희(Jee-Hee Yoon) 정회원

- 1982년 : 한양대학교 전자공학과 (학사)
- 1985년 : 일본 구주대학교 정보공학과(석사)
- 1988년 : 일본 구주대학교 정보공학과(박사)
- 1998년 ~ 1999년 : 미국 UCLA대학교 전산학과 방문교수
- 1988년 ~ 현재 : 한림대학교 컴퓨터공학과 교수

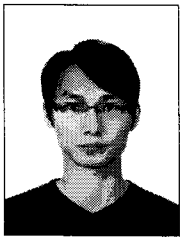
<관심분야> : 시계열 데이터베이스, 데이터 마이닝, XML, 공간 데이터베이스/GIS

홍 동 완(Dong-Wan Hong) 정회원

- 1996년 : 한림대학교 전자계산학과(학사)
- 1998년 : 한림대학교 컴퓨터공학과(석사)
- 2008년 : 한림대학교 컴퓨터공학과(박사)
- 2003년 ~ 2005년 : 송곡대학 미디어컨텐츠과 전임강사
- 2005년 ~ 2007년 : 송곡대학 미디어컨텐츠과 조교수
- 2008년 ~ 현재 : 한림대학교 바이오메디컬학과 겸임조교수

<관심분야> : 기가 시퀀싱, 유전체 변이, 서열 정렬, 마이크로어레이, 유전자 기능 분석

이 종 근(Jong-Keun Lee) **정회원**



- 2005년 : 한림대학교 컴퓨터공학과(학사)
- 2007년 : 한림대학교 컴퓨터공학과(석사)
- 2007년 ~ 현재 : 한림대학교 컴퓨터공학과 박사 과정(수료)

<관심분야> : 기가 시퀀싱, 유전체 변이, 서열 정렬, 마이크로어레이, 유전자 기능 분석