International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:1, No:1, 2007

A Survey on Life Science Database Citation Frequency in Scientific Literatures

Hendry Muljadi, Jiro Araki, Satoru Miyazaki, and Asao Fujiyama

***Abstract*—**There are so many databases of various fields of life sciences available online. To find well-used databases, a survey to measure life science database citation frequency in scientific literatures is done. The survey is done by measuring how many scientific literatures which are available on PubMed Central archive cited a specific life science database. This paper presents and discusses the results of the survey.

***Keywords*—**Life science, database, metadatabase, PubMed Central.

I. Introduction

NOWADAYS, there are so many databases of various fields of life sciences available online. With the increasing numbers of publicly available databases, metadabases such as the NAR Molecular Biology Database Collection [1], Pathguide [2], Semantic Metadatabase (SEMEDA) [3], etc have emerged to provide service to support end-users in finding appropriate data for their researches from appropriate databases.

The Japanese Bioportal site (http://www.bioportal.jp) also provides a metadatabase (http://www.ps.noda.tus.ac.jp/biometadb/index.html) [4]. The main characteristic of the Japanese Bioportal Metadatabase is that it provides the databases' access methods including parameters accepted by CGIs provided by the publicly available database sites [5]. As a database of databases, currently the metadatabase contains links to more than 28,000 records of more than 600 publicly available databases. Almost all of the collected databases have been introduced in the 2003-2005 Database Issue of Nucleic Acids Research Journal (http://nar.oxfordjournals.org).

However, when a keyword search is done, the metadatabase will show all the reference databases in alphabetical order. For example, when "genome" is submitted as the keyword, the search engine will retrieve 179 databases and show them in the alphabetical order as the search result. The metadatabase will be more useful to support end-users if it can show results in recommending order, rather than in alphabetical order.

Recommendation order of the databases can be constructed based on the use frequency ranking order. It is possible to use the cited rank of databases in scientific literatures to create well-used database ranking order.

A survey to measure citation frequency of the collected databases in science literatures is done. The purpose of the survey is to measure the use of the collected databases in year 1990-2006 as well as to analyze in what fields the databases are being used. The survey is done by measuring how many scientific literatures available on PubMed Central archive (http://www.pubmedcentral.nih.gov/) cited each collected database.

The structure of this paper is as follows. In section 2, the process of measuring the citation frequency of life science databases in scientific literatures and the results are described. Section 3 discusses the results. Section 4 states the conclusion and future works of this research.

II. Measuring Life Science Database Citation in the Scientific Literatures

A. Scope of the Scientific Literatures

As mentioned in section 1, the survey is done to measure citation frequency of life science databases in scientific literatures available on PubMed Central archive. The survey is limited to literatures published in year 1990-2006.

As the PubDate of the PubMed Central archive is set between (1990-01-01) and (2006-10-20), there are 394,314 full text papers from 284 journals available.

Since there are many other journals that are not covered by PubMed Central, it is arguable that the survey has limitation in measuring the use of the databases. However, since only internationally recognized life science journals are covered by PubMed Central, we can measure the use of the databases only in researches that are internationally recognized.

B. The Measuring Process

Fig. 1 illustrates the process of measuring the citation frequency of life science databases in the scientific literatures.

To measure how many scientific literatures cited each database, the name of databases which are recorded in the Japanese Bioportal Metadatabase are used as the search keywords. Then, the keywords are submitted to the NCBI

Manuscript received November 16, 2006.

H. Muljadi is with the Research Organization of Information and Systems, Tokyo, Japan (phone: 81-3-4212-2664; fax: 81-3-3556-1916; e-mail: hendry@nii.ac.jp).

J.Araki is with the Mitsubishi Research Institute. Currently, he is also with the Department of Informatics, Graduate University of Advanced Studies, Tokyo, Japan (e-mail: jiro@grad.nii.ac.jp).

S.Miyazaki is with the Faculty of Pharmaceutical, Tokyo University of Science, Tokyo, Japan (e-mail: smiyazak@rs.noda.tus.ac.jp).

A.Fujiyama is with the National Institute of Informatics, Tokyo, Japan (e-mail:afujiyam@nii.ac.jp).

Entrez Cross-Database Search to search the PubMed Central archive (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pmc).

However, databases which use simple terms as their names, such as All, Protein, are not submitted to the NCBI Entrez Cross-Database Search since,

1) developing an algorithm to determine whether a word refers to a database name or not is beyond the scope of this research,
2) checking a very large amount of literatures manually, if not impossible, is a laborious, expensive and time-consuming task.

The results are sorted based on the publishing year of the citing literatures and the journal classification of the citing literatures, since it is possible to obtain the journal's name and the publishing year of the citing literatures,

NCBI annotates journals in PubMed Central archive by using 125 different subject terms (http://www.nlm.nih.gov/bsd/journals/subjects.html). The NCBI's annotations of the journals in PubMed Central archive are used to determine the journal classification of the citing literatures.

C. The Results

There are 176 databases that have been cited in more than 10 scientific literatures in year 1990-2006. Due to space limitation, all the results cannot be shown in this paper.

Table I shows a list of databases that were cited in more than 10 scientific literatures published in year 1990-1995. Table II shows a list of databases that have been cited in more than 130 scientific literatures published in year 1990-2006. Table III shows a list of databases that were cited in more than 80 scientific literatures published in year 2001-2005. Table IV shows a list of new databases, which are databases that were not cited at all before year 2001, but have been cited in more than 10 scientific literatures published after year 2000.

Table V shows a list of databases shown in Table III sorted based on the journal classification of the citing literatures. Only 10 fields are shown in Table V.

III. Discussion

From the results, the following points can be stated.

1) From Table II, we discover that GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) and EMBL (http://www.ebi.ac.uk/embl/index.html) have been the most cited databases in year 1990-2006. However, since year 2002, EMBL has been ranked below NCBI.
2) There are 24 databases listed on Table I. Five of them are not listed in Table III.
   -NBRF (http://pir.georgetown.edu/nbrf/),
   -EcoSys (http://ukcrop.net/perl/ace/search/EcoSys),
   -HerVD (http://herv.img.cas.cz/),
   -Haemophilia B database (http://www.kcl.ac.uk/ip/petergreen/haemBdatabase.html),
   -BLOCKS database (http://blocks.fhcrc.org/).
   In other words, there are databases that have become less popular.

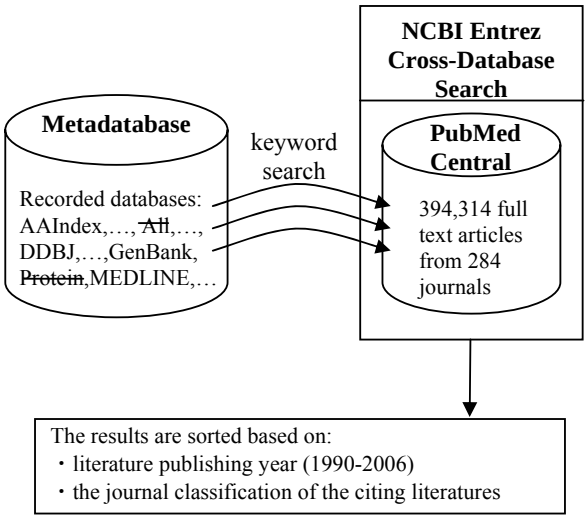

Fig. 1 Illustration of the process of measuring the usage of life science databases in scientific literatures

TABLE I
DATABASES THAT ARE CITED IN MORE THAN 10 SCIENTIFIC LITERATURES IN YEAR 1990-1995

Database Name	Total	1990	1991	1992	1993	1994	1995
GenBank	8045	1010	1277	1530	1621	1663	944
EMBL	5244	1184	877	966	907	869	441
DDBJ	872	101	126	193	164	174	114
Swiss-Prot	664	32	92	131	140	178	91
NBRF	481	112	157	96	65	37	14
MEDLINE	438	56	59	67	73	96	87
HUGO	267	47	45	31	52	56	36
NCBI	208	0	5	13	36	88	66
PROSITE	193	7	15	36	54	52	29
PDB	188	28	29	28	22	39	42
Genpept	138	5	14	33	31	36	19
Islander	77	8	10	7	17	19	16
GDB	52	4	5	5	14	13	11
MIPS	49	7	5	9	9	8	11
Entrez	26	0	0	2	4	11	9
SGD	25	2	4	5	4	6	4
EcoSys	19	0	1	0	0	2	16
dbEST	18	0	0	0	5	4	9
OMIM	18	1	3	3	6	3	2
FlyBase	15	0	0	3	2	8	2
HERVd	14	3	6	1	1	3	0
haemophilia B database	14	1	3	3	3	3	1
ExPASy	12	0	0	3	3	6	0
BLOCKS Database	11	0	1	0	1	5	4

TABLE II
DATABASES THAT ARE USED OR CITED IN MORE THAN 130 SCIENTIFIC LITERATURES PUBLISHED IN 1990-2006

Database Name	Total	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
GenBank	39170	1010	1277	1530	1621	1663	944	781	835	2130	2322	3050	3142	3271	3523	3884	4191	2401
EMBL	15000	1184	877	966	907	869	441	286	213	699	659	931	892	862	837	1008	1007	448
NCBI	9498	0	5	13	36	88	66	43	50	191	188	486	685	927	1226	1683	2097	1714
PDB	4766	28	29	28	22	39	42	54	42	107	157	300	342	421	704	763	916	627
MEDLINE	4376	56	59	67	73	96	87	105	124	226	274	303	294	340	363	430	503	458
DDBJ	3691	101	126	193	164	174	114	42	52	217	216	416	377	402	339	277	292	153
Swiss-Prot	2765	32	92	131	140	178	91	69	49	132	101	150	179	215	284	307	316	266
TIGR	2699	1	0	0	1	1	3	1	8	52	61	203	272	321	400	498	533	340
pubmed	1856	0	0	0	0	0	0	0	2	28	33	76	91	163	228	316	446	473
PFAM	1679	0	0	1	0	0	0	0	0	9	13	66	114	211	236	330	411	284
Gene Ontology	1578	0	0	0	0	0	0	0	0	0	0	3	21	69	172	308	496	509
HUGO	1568	47	45	31	52	56	36	29	14	30	20	35	46	43	66	70	116	79
PROSITE	1364	7	15	36	54	52	29	18	19	85	81	129	124	146	148	182	145	92
ExPASy	1333	0	0	3	3	6	0	6	8	35	53	91	130	149	194	260	267	128
UniGene	1284	0	0	0	0	0	0	1	4	11	29	75	118	131	170	247	291	205
Ensembl	1262	0	0	1	0	0	0	0	1	0	0	5	28	81	183	253	407	303
Entrez	1164	0	0	2	4	11	9	9	10	34	22	69	68	127	128	166	253	251
RefSeq	944	0	0	0	0	0	0	0	0	0	0	4	18	65	139	197	277	244
TrEMBL	865	0	0	0	0	0	1	2	1	13	19	35	61	104	131	157	180	160
MIPS	853	7	5	9	9	8	11	3	12	18	16	34	59	110	143	152	163	78
dbEST	840	0	0	0	5	4	9	10	29	70	47	95	84	79	88	126	109	85
NBRF	822	112	157	96	65	37	14	5	3	8	9	15	7	4	7	3	3	2
FlyBase	775	0	0	3	2	8	2	2	6	18	16	30	51	99	97	122	193	126
Saccharomyces Genome Database	756	0	0	0	0	0	0	3	11	34	30	48	52	56	90	134	155	143
OMIM	722	1	3	3	6	3	2	1	10	24	35	46	58	78	86	100	138	123
Islander	713	8	10	7	17	19	16	33	15	27	35	24	29	53	79	80	85	106
InterPro	648	0	0	0	0	0	0	0	0	1	1	6	23	75	111	119	167	140
LocusLink	602	0	0	0	0	0	0	0	0	0	1	10	32	76	128	140	141	74
KEGG	575	1	0	0	0	0	0	0	0	3	4	17	23	29	65	85	156	192
SGD	571	2	4	5	4	6	4	3	5	17	10	27	38	42	61	89	103	108
TRANSFAC	568	0	0	0	0	0	3	1	3	16	18	24	37	59	63	113	120	111
dbSNP	427	0	0	0	0	0	0	0	0	0	2	12	20	37	56	87	114	99
TAIR	426	1	2	0	1	1	3	1	0	1	1	4	24	33	58	93	103	65
UniProt	386	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	149	201
Genpept	357	5	14	33	31	36	19	11	6	22	8	16	35	21	31	29	21	19
WormBase	335	0	0	0	0	0	0	0	0	0	0	1	17	28	46	80	79	84
GDB	296	4	5	5	14	13	11	7	11	31	17	36	26	17	17	20	24	11
COG database	292	0	0	0	0	0	0	0	0	0	0	8	24	30	36	42	71	81
ProDom	257	0	0	0	0	1	0	0	0	5	8	24	31	39	48	40	37	24
SCOP database	217	0	0	0	0	0	0	0	1	4	1	7	14	26	27	38	38	61
Entrez Gene	173	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	60	106
EcoCyc	173	0	0	0	0	0	0	1	1	7	4	11	14	13	20	25	31	46
AceDB	170	0	0	0	3	2	5	5	5	14	5	19	16	22	17	28	20	9
HIV Sequence Database	162	3	1	2	0	2	1	0	0	2	4	11	15	19	27	28	30	16
NCBI Taxonomy	147	0	0	0	0	1	0	0	1	2	1	9	9	14	22	20	33	35
PlasmoDB	133	0	0	0	0	0	0	0	0	0	0	0	6	13	23	36	34	21
RDP-II	131	0	0	0	0	0	0	0	0	1	2	4	10	16	19	23	33	23

TABLE III
DATABASES THAT ARE CITED IN MORE THAN 80 SCIENTIFIC LITERATURES IN YEAR 2001-2005

Database Name	Total	2001	2002	2003	2004	2005
GenBank	18011	3142	3271	3523	3884	4191
NCBI	6618	685	927	1226	1683	2097
EMBL	4606	892	862	837	1008	1007
PDB	3146	342	421	704	763	916
TIGR	2024	272	321	400	498	533
MEDLINE	1930	294	340	363	430	503
DDBJ	1687	377	402	339	277	292
PFAM	1302	114	211	236	330	411
Swiss-Prot	1301	179	215	284	307	316
pubmed	1244	91	163	228	316	446
Gene Ontology	1066	21	69	172	308	496
ExPASy	1000	130	149	194	260	267
UniGene	957	118	131	170	247	291
Ensembl	952	28	81	183	253	407
PROSITE	745	124	146	148	182	145
Entrez	742	68	127	128	166	253
RefSeq	696	18	65	139	197	277
TrEMBL	633	61	104	131	157	180
MIPS	627	59	110	143	152	163
FlyBase	562	51	99	97	122	193
LocusLink	517	32	76	128	140	141
InterPro	495	23	75	111	119	167
Saccharomyces Genome Database	487	52	56	90	134	155
dbEST	486	84	79	88	126	109
OMIM	460	58	78	86	100	138
TRANSFAC	392	37	59	63	113	120
KEGG	358	23	29	65	85	156
HUGO	341	46	43	66	70	116
SGD	333	38	42	61	89	103
Islander	326	29	53	79	80	85
dbSNP	314	20	37	56	87	114
TAIR	311	24	33	58	93	103
WormBase	250	17	28	46	80	79
COG database	203	24	30	36	42	71
ProDom	195	31	39	48	40	37
UniProt	185	0	0	0	36	149
SCOP database	143	14	26	27	38	38
Genpept	137	35	21	31	29	21
HIV Sequence Database	119	15	19	27	28	30
PlasmoDB	112	6	13	23	36	34
GDB	104	26	17	17	20	24
EcoCyc	103	14	13	20	25	31
AceDB	103	16	22	17	28	20
RDP-II	101	10	16	19	23	33
NCBI Taxonomy	98	9	14	22	20	33
GeneCards	97	13	16	18	20	30
HomoloGene	89	3	6	12	27	41
IMGT	85	12	12	14	20	27
Rfam	84	0	0	7	28	49
RegulonDB	83	11	13	13	25	21

TABLE IV
NEW DATABASES

Database Name	Total	2001	2002	2003	2004	2005	2006
UniProt	386	0	0	0	36	149	201
Entrez Gene	173	0	0	0	7	60	106
Fantom2	81	0	2	33	13	19	14
DBTSS	70	0	3	5	18	23	21
Genolevures	57	0	8	6	18	13	12
Inparanoid	51	0	0	4	7	12	28
MaizeGDB	50	0	0	2	15	21	12
dictyBase	47	0	1	2	7	19	18
HPRD	37	0	0	1	6	12	17
BayGenomics	34	0	0	7	9	13	5
HOMSTRAD	33	1	1	2	13	10	6
iProClass	30	3	5	4	5	8	5
probeBase	30	0	3	7	6	8	6
HGVBASE	30	0	5	6	11	6	2
PlantsP	25	2	7	5	7	2	2
PlantsT	24	1	2	6	5	1	1
coliBase	23	0	0	1	3	11	8
Integr8	23	0	0	0	0	11	12
GtRDB	22	1	1	7	8	5	0
UniVec	22	1	0	2	8	4	6
ProtoNet	21	0	0	4	3	9	5
GeneNest	21	2	4	3	5	4	3
FANTOM3	20	0	0	0	0	0	20
Predictome	20	0	3	3	6	3	5
yMGV	19	1	6	3	5	3	1
SNP500Cancer	19	0	0	1	2	6	10
HIV Drug Resistance Database	19	0	0	0	6	8	5
RTCGD	18	0	0	1	4	8	5
UniProtKB	18	0	0	0	0	5	13
Worfdb	18	1	0	2	9	5	1
PRODORIC	18	0	0	3	4	3	8
ECgene	18	0	0	0	2	5	11
PubChem	18	0	0	0	0	8	10
CryptoDB	17	0	1	0	7	7	2
PseudoCAP	16	0	2	3	3	5	2
UniRef100	16	0	0	0	0	7	9
LGICdb	16	2	2	3	3	3	3
TargetDB	15	0	0	2	5	3	5
NucleaRDB	15	1	2	4	1	4	3
H-Invitational	15	0	0	0	5	6	4
ChEBI	15	1	1	0	1	6	5
BarleyBase	14	0	0	0	4	5	5
JSNP	14	0	5	0	4	2	3
Genome Reviews	14	0	1	0	0	6	7
MHCBN	13	0	1	0	0	7	5
TMPDB	12	0	0	2	2	4	2
RARGE	12	0	0	0	2	3	1
trGEN	11	3	1	1	3	3	0
trome	11	0	0	0	4	3	0
UniRef90	11	0	0	0	0	5	6
Oryzabase	11	0	1	1	2	5	2
Nuclear Protein Database	11	0	2	2	2	5	0
GENSAT	11	0	1	1	1	3	5

TABLE V
MAJOR DATABASES IN YEAR 2001-2005 SORTED BASED ON THE JOURNAL CLASSIFICATION OF THE CITING LITERATURES

Database Name	Biochemistry	Genetics	Library Science	Medical Informatics	Family Practice	Internal Medicine	Laboratory Techniques and Procedures	Microbiology	Public Health	Molecular Biology
GenBank	2996	1633	24	253	0	0	189	6384	7	4691
EMBL	3669	411	4	75	0	0	19	1248	2	2855
NCBI	1271	1413	36	369	0	0	34	1807	0	912
PDB	922	151	1	187	0	0	4	167	6	685
TIGR	249	378	0	56	0	1	10	236	0	189
MEDLINE	181	33	982	270	135	115	2	66	176	32
DDBJ	464	114	0	15	0	0	23	699	0	912
PFAM	363	279	0	100	0	0	1	96	0	151
Swiss-Prot	681	268	0	142	0	0	3	180	0	238
pubmed	305	97	178	171	30	9	1	34	36	45
Gene Ontology	353	435	2	240	0	0	0	14	0	85
ExPASy	145	83	0	20	0	0	6	113	0	164
UniGene	246	373	1	84	0	0	0	2	0	107
Ensembl	272	420	1	108	0	0	0	5	0	100
PROSITE	265	91	0	42	0	0	5	98	0	184
Entrez	262	195	20	102	0	0	0	73	1	65
RefSeq	272	309	1	103	0	0	0	13	0	36
TrEMBL	284	181	0	98	0	0	1	35	0	44
MIPS	155	130	0	56	0	0	2	21	1	118
FlyBase	136	242	0	39	0	0	0	4	0	187
LocusLink	168	175	2	74	0	0	0	0	0	30
InterPro	226	158	1	56	0	0	0	27	3	32
Saccharomyces Genome Database	108	157	0	64	0	0	0	20	0	250
dbEST	130	258	1	25	0	0	0	3	0	75
OMIM	194	137	15	48	0	0	0	0	3	46
TRANSFAC	183	102	0	45	0	0	1	3	0	65
KEGG	181	123	0	90	0	0	0	25	0	14
HUGO	100	94	17	39	3	1	2	156	96	80
SGD	136	104	0	59	0	0	0	12	2	124
Islander	7	6	3	9	10	44	1	12	383	0
dbSNP	99	125	1	17	0	0	0	1	0	13
TAIR	59	78	0	14	0	0	0	4	2	21
WormBase	67	133	0	24	0	0	0	0	0	63
COG database	61	73	1	43	0	0	0	19	0	9
ProDom	86	41	0	18	0	0	0	18	0	13
UniProt	143	67	0	71	0	0	0	10	0	9
SCOP database	87	21	0	47	0	0	0	2	0	9
Genpept	58	37	0	7	0	0	1	25	0	46
HIV Sequence Database	17	1	1	1	0	0	1	18	0	2
PlasmoDB	28	20	0	2	0	0	0	7	0	24
GDB	64	56	11	8	0	0	0	2	1	29
EcoCyc	63	33	0	28	0	0	0	14	0	3
AceDB	39	67	0	5	0	0	0	4	0	23
RDP-II	15	9	0	14	0	0	0	81	0	2
NCBI Taxonomy	74	16	0	28	0	0	0	9	0	3
GeneCards	41	27	3	13	0	0	0	0	0	6
HomoloGene	47	42	0	15	0	0	0	0	0	7
IMGT	43	7	0	5	0	0	1	1	0	5
Rfam	35	24	0	15	0	0	0	6	0	9
RegulonDB	50	25	0	16	0	0	0	3	0	2

3) From Table IV, we discover that there are many new databases developed. Among them, UniProt (http://www.ebi.ac.uk/uniprot/) and Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) have been cited in more than 100 literatures.
4) Comparing Table I, Table II, and Table III, we discover that there are databases such as TIGR (http://www.tigr.org), PFAM (http://www.sanger.ac.uk/Software/Pfam/), MIPS (http://mips.gsf.de), Ensembl (http://www.ensemble.org), etc, that were not major databases in year 1990-1995, but have become major ones in year 2000s.
5) Comparing Table II and Table V, we discover that though GenBank and EMBL have been the most used databases in year 1990-2006 as well as the most used databases in various fields, yet there are fields where these two major databases are not cited at all.

The survey gives results that can be used to create well-used database ranking order based on the cited rank of databases in scientific literatures. The ranking order can be constructed based on the publishing year of the citing literatures as well as on the journal classification of the citing literatures.

IV. Conclusion and Further Works

The survey shows that there are only 176 databases that have been cited in more than 10 scientific literatures in year 1990-2006. The results can be used to create well-used databases ranking order based on the cited rank of databases in scientific literatures as well as on the journal classification of the citing literatures.

However, to enable the construction of the recommendation order, further work needs to be done on how the journal classification can be associated with life science terms, since users may prefer to use life science terms as the search keywords rather than just selecting the provided but limited classifications.

And since there are more than 400 other databases that cannot be recommended based on well-used databases ranking order, manual analysis of the recorded databases is also being considered. It is intended to create metadata of each database that may be useful for the construction of the recommending order in the metadatabase.

Acknowledgment

Special thanks to Prof. Hideaki Takeda and Dr. Shoko Kawamoto for their valuable input related to this paper.

References

[1] M.Y.Galperin,"The Molecular Biology Database Collection: 2006 Update," *Nucl. Acids Res*., vol.34, Jan.2006, D3-5
[2] G.D.Bader, M.P.Cary, C. Sander,"Pathguide: a Pathway Resource List," *Nucl. Acids Res*., vol.34, Jan.2006, D504-506
[3] J. Kohler, S. Schulze-Kremer,"The Semantic Metadatabase (SEMEDA): Ontology Based Integration of Federated Molecular Biological Data Sources," *Silico Biology*, vol.2, no.3, March 2002, pp.219-231
[4] S.Miyazaki, H.Sugawara,"Japanese Bioportal: The Construction of Bio-metadatabase," presented at the 28th Annual Metting of the Molecular Biology Society, Fukuoka City, Japan, December 7-10, 2005, Poster 2P-0068 (in Japanese).
[5] S.Miyazaki, T.Asano, S.Kitadate, H.Sugawara,"META-database as the collection of access methods and sets of parameters to manipulate CGIs of the molecular biological databases on the Internet," presented at the 13th Int'l Conf. on Intelligent Systems for Molecular Biology, Michigan, June 25-29, 2005, Poster G-70.