ROLE OF INFORMATION TECHNOLOGY IN INTEGRATING BIOLOGICAL DATABASES INTO METABASES A REVIEW

Shweta Agrawal
Dr. Rahore Pragya

Abstract

Life science researchers are experiencing a "data explosion" due to vast amounts of raw data being produced as genomic, proteomic and structural representing experimentation. This exponential growth have is expected to continue. These explosion of data are kept in different databases scattered across the diverse industries, different universities and research institutes throughout the world. There is a dire need for the creation of intelligent database of these databases to query, collate, integrate, analyze and give biological sense to these different data to tackle more and more relevant and challenging scientific problems. An important resource for finding biological databases is a special annual issue of the journal Nucleic Acids Research (NAR) which is freely available and categories many of the publicly available on the databases related to biology and bioinformatics. Although a huge variety of databases are available on the, PubMed maintained by the National Center for Biotechnology Information (NCBI) Ensembl maintained by the European Bioinformatics Institute (EBI) and the Wellcome Trust; the UCSC Genome Browser an anonymous and raw genome browser maintained by David Hausler's group at the University of California at Santa Cruz; FlyBase maintained by the FlyBase Consortium; WormBase and the Gene Ontology (GO) databases exemplify some of them are biological databases. Many researchers are not aware of their availability and consequently these databases were not effectively used. A better solution would to generate a database of databases which contains a plethora of information of various biological databases containing these scientific and publically independence of databases by allowing information data to contain it to be easily integrated to enable across database queries. This study deals with the features of four major metabases of biological databases — DBD (database of biological databases) http://www.biodb.info/, DOD — Database of Databases (http://www.progenebio.in/DoD/DoD.htm), MetaBase - Database (http://biodatabase.org) and DBcat (http://www.infobiogen.fr/services/dbcat)

Keywords: Databases, biological databases, Database of biological Databases, bioinformatics glossary

1 Introduction – Biological databases are libraries of information concerned with life sciences collected from scientific experimentation published literature high-throughput experiment techniques and computational analysis. They contain

Lecture Department Biotechnology SIM S Indore
HOD Department Biotechnology SIM S Indore

information from research areas including genomic microarray gene expression and phylogenetics. Gb for Biotechnology Information, the EMBL (European database and DDBJ DNA DataBank of Japan), a of databases accessible through the Internet. Beg databases such as the proteomic databases SwissPr Expression Database), KEGG (he Kyoto Encyclopedia Genome Sequence Database GSDB), National Center be queried via Web based interfaces (Andrzej Polan momentous tool in assisting scientist to understa biological phenomena from the structure of biomolec chemistry of biomolecules of organisms and understanding the fight against diseases, development of medical among species (tpen.wikipedia.org/wikiBiol In the last two decades storage of biological data increasingly common and these databases have grown 2011). The amount of information about the world do number of databases probably increases at a similar 2003. Although several databases are reported in l a are otherwise available and these databases w Babu et al 2005) efforts are directed toward for convenient identification and correlation of in common platform for various biological databases.

sp rotomics ismetabolomics, ein 2004), NCBI (National Center Molecular Biology Laboratory) the best known of the hundreds eon 2002. Besides these archival of UniProt, CGED (Cancer Gene of Genes and Genomes, the for Genome Resources, and ski 2007). These databases are nd, explain and explore various udes and hence can be exploring the evolution of species, sheddi ons and in discovering phylogeny gical database). in public databases has become exponentially, Cynthia Gibas, ubles every 20 months these and pace (Sushmita M in, features many researchers are not even effectively utilized (Ajay generation of database of databases potent queries besides providing a

1. DBD: Database of Biological Bioinformatics Databases

DBD: Database of Biological Bioinformatics Databases is a comprehen covering various databases related to molecular mod microbial databases. Exceptionally, the databases based on technical terms (keywords) http://omicsw ebako.org/bw webglossary.htm l the desired database (figure 1.1 and 1.2). Siva retrieval of information regarding "Gene" just "Gene", entails a number of databases related to ge (http://www.bioinform atics.org/forum sforum.php?fo

ases." DBD: Database of sive compilation of 200 databases cluding, HIV, human genome, present DBD the alphabetizing of
viz.
This makes very quick access to
Kiran R R 2009). To exemplify, for click alphabet "G" and search for keyword
ne.
rum _id=5542 (Figure 1.3)
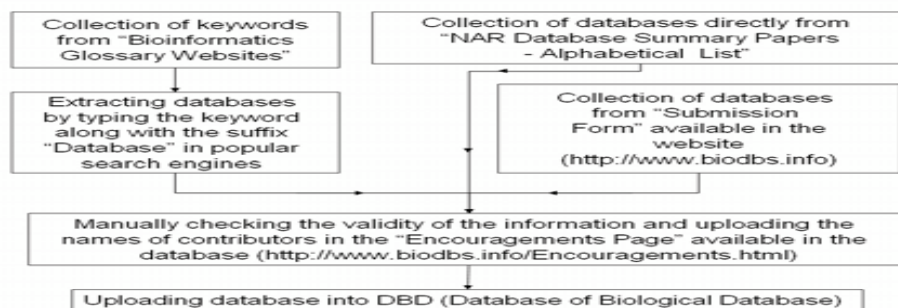
Figure 1.1 Structure of DBD (Database of Biological Databases)

Table 1: Alphabetical listing of number of keywords with total number of databases linked to DBD (Database to Biological Data base)

Alphabets	Total keywords	Total number of databases	Some major keywords with total number of database linked to DBD		
A	16	64	Alignment-8, Arabidopsis-2, Spring-2, Annotation-12, Agricultural-4		
B	20	68	Bacteria-B, Binding-4, Biochemical Pathway-9, Biological database-11		
C	48	88	Caenorhabditis elegans-5, Cancer-14, Chromosome-9, Cytogenetics-3, Cluster-4		
D	33	76	Disease-12, Domain-9, Drosophila-5, DNA-18		
E	25	86	*Ec oli* -14, Enzyme-16, Eukaryote-8, EST-1, Expression database-5, EBI-3, Evolution-3		
F	12	19	Fish-3, Fly-3, Functional Genomics-4		
G	15	77	Genome-38, Gene & Gene Expression-16, Genomics-7, Germ-6		
H	25	93	HIV-21, Homologous & Homology-17, Human-18, *Haemophilus*-3, Hightaxon-5		
I	23	56	Inhibitor-10, Intron-6, Isotope-4		
J	Nil	Nil	Nil		
K	4	4	Kinases-K, KEGG-1		
L	9	32	Ligand-23, Liver protome-1		
M	41	258	Motif-10, Metabolic Pathway-1, Microarray-22, Microbial-5, Mitochondria-5, Mouse-90, Mutation – 40		
N	11	34	NMR-3, Nuclei-N, Nucleotide-6		
O	11	38	Oligo-38, Organism-5, Ontology-9		
P	52	297	Panel-10, Pathway-12, PCR-8, Peptide-34, Probe-13, Promoter-15, Protease-12, Protein- Phenotype-9, Plant-Plasmid-10, Primer-0, 5, Proteome-18		
Q	3	11	QSAR-6, Quantitative-4		
R	19	113	Radiohybrid-20, Retrovirus-, RNA-39, Ribosomal RNA site-17		
S	34	132	Sanger-5, Sequence-4, Signal Transduction-8, SNP-16, Species-Structure-3, Synapse-3		
T	19	47	Taxonomy-8, Transcript-RNA-3, Transgenic-3, Tumor-9, Tree-3		
U	3	7	Unigene-4, UnPro-2		
V	5	31	Variation-9, Viral-6, Vector-4		
W	3	4	Worm-2		
X	3	4	X chromosome-2, *Xenopus laevis* -1		
Y	2	25	Yeast-23		
Z	1	5	*Zea mays* -5		

Figure 2 Screenshot Database of Biological Database (DBD)

DBD
DATABASE OF BIOLOGICAL DATABASE

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Proceed by clicking the first alphabet of the keyword

Keywords	Database Description	Website
Abstract	Deja Vu: A Database of Highly Similar and Duplicate Citations	http://spore.swmed.edu/dejavu/
AceDB	A hierarchical database system for displaying genomic data (*Caenorhabditis elegans*)	http://www.acedb.org/
Activity	Functional DNA/RNA site activity	http://wwwmgs.bionet.nsc.ru/mgs/systems/activity/
Acyrthosiphon pisum	AphidBase - Genomic database for the pea aphid (*Acyrthosiphon pisum*)	http://www.aphidbase.com/
Aedes sp..	Aedes aegypti Database	http://www.broad.mit.edu/annotation/disease_vector/aedes_aegypti/
Alignment	Homologous Structure Alignment Database (HOMSTRAD)	http://www-cryst.bioc.cam.ac.uk/~homstrad/
	Protein Alignments organised as Structural Superfamilies	http://caps.ncbs.res.in/campass/pass2.html
	EMBL-ALIGN Information for Submitters	http://www.ebi.ac.uk/embl/Submission/alignment.html
	Standards Alignment Database Homepage	http://standards.netacad-currdev.net/home/default.asp
	BAliBASE (version 2.0): A benchmark alignment database.	http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE2/
	The SCANPS VERSION 2.3.9 protein sequence scanning package - User Guide.	http://www.compbio.dundee.ac.uk/Software/Scanps/scanps.html
	SAD-A normalized structural alignment database: improving sequence-structure alignments	http://portal.acm.org/citation.cfm?id=1092906.1093101
	A Multiple Alignment Database	http://www.daimi.au.dk/~besen/TBiB2006/projects/project2.html
	CDD: Curated alignments of protein domains from Pfam, SMART and COG databases	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
	CE: Combinatorial Extension method to compute and review 3D protein structure alignments	http://cl.sdsc.edu/ce.html
Allele	Allele frequencies and DNA polymorphisms	http://alfred.med.yale.edu/
Alternative Splicing	ASDB-Alternative Splicing Database	http://hazelton.lbl.gov/~teplitski/alt/
	EDAS: EST Derived Alternative Splicing Database	http://www.ig-msk.ru:8005/EDAS/
	ASTD: Alternative Splicing and Transcript Diversity	http://www.ebi.ac.uk/astd/main.html;jsessionid=D55E5F0D5C1455A4C56AC1004EBDD67F
	Human Genes with Alternative Splicing	http://moult.umbi.umd.edu/human2004/
	Alternative Splice Site Predictor (ASSP)	http://www.es.embnet.org/~mwang/assp.html
	Alternative Splicing website	http://www.exonhit.com/index.php?page=59
	Hollywood RNA Alternative Splicing Database	http://hollywood.mit.edu/Login.php
	Plant Alternative Splicing Database	http://pasdb.genomics.org.cn/
	H-DBAS: Human Database of Alternative Splicing	http://h-invdb.jbic.or.jp/h-dbas/
	Manually Annotated AlternativelySpliced Events (MAASE) database system	http://maase.genomics.purdue.edu/
	Putative Alternative Splicing Database (PALS db)	http://ymbc.ym.edu.tw/palsdb/
	ProSplicer : An Alternative Splicing Database based on Protein, mRNA, and EST Sequences	http://prosplicer.mbc.nctu.edu.tw/
	ECgene: Genome Annotation for Alternative Splicing	http://genome.ewha.ac.kr/ECgene/
	ASTRA (Alternative Splicing and TRanscription Archives) is the database of elementary patterns of alternative splicing and transcriptional initiation.	http://alterna.cbrc.jp/
	TassDB: TAndem Splice Site DataBase (TassDB stores extensive data about alternative splice events at GYNGYN donors and NAGNAG acceptors)	http://helios.informatik.uni-freiburg.de/TassDB/
	FastDB: Friendly Alternative Splicing and Transcripts Database	http://www.fast-db.com/fastdb2/frame.html
	Alternative Splicing Database	http://www.bioinformatics.ucla.edu/ASAP2/
	Alternative Splicing in Plants	http://www.plantgdb.org/ASIP/

Database Submission Form: Submit Databases to DBD in the following format. After verification it will be posted in DBD in 2 or 3 weeks. For further queries contact rrskiran@msrit.edu.

Enter Database Website :

Database Description:

Suitable Keyword:

Institution Name

Official Email Address

[Submit Database]

DBD: Database of Biological Database team are R.R. Siva Kiran, MVN Setty, Department of Biotechnology, MS Ramaiah Institute of Technology, MSR Nagar, Bangalore, India and G. Hanumantha Rao, Center for Biotechnology, Department of Chemical Engineering, Andhra University, Visakhapatnam-530003, India. DBD consists of 1200 Database entries covering wide range of databases useful for biological researchers.

Encouragements List (or list of database contributors)

Andhra University

All Rights Reserved, Maintained by R.R. Siva Kiran, Department of Biotechnology, MS Ramaiah Institute of Technology, Bangalore, India

Last update: 30-July-2008

Figure 3 Screenshot Search for database on gene

DBD
DATABASE OF BIOLOGICAL DATABASE

(A) (B) (C) (D) (E) (F) (G) (H) (I) (J) (K) (L) (M) (N) (O) (P) (Q) (R) (S) (T) (U) (V) (W) (X) (Y) (Z)

Proceed by clicking the first alphabet of the keyword

Keywords	Database Description	Website
Gamete	Eggcorns Database	http://eggcorns.lascribe.net/english/572/gamete/
GEML	Gene Expression Markup Language (GEML)	http://xml.coverpages.org/geml.html
GenBank	GenBank Database	http://www.psc.edu/general/software/packages/genbank/genbank.html
Genome	List of Genome Database	http://www.gdb.org/
	Yeast Genome Database	http://www.yeastgenome.org/
	Mouse Genome Database	http://www.informatics.jax.org/
	Tumor Gene Database	http://www.tumor-gene.org/tgdf.html
	Cannadian Genome Database	http://www.candidagenome.org/
	Plant Genome Research Through PlantGDB	http://www.plantgdb.org/
	Tetrahymena thermophila genome	http://www.ciliate.org/
	Rat Gene Database	http://rgd.mcw.edu/
	Gene Knockout Database	http://www.bioscience.org/knockout/knochome.htm
	Rat Genome Database	http://rgd.mcw.edu/
	Maize Genome Database	http://www.maizegdb.org/
	HGVbase (the Human Genome Variation Database)	http://hgvbase.cgb.ki.se/
	The Gene Expression Database (GXD)	http://www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml
	UCSC Genome website	http://www.ucsc.edu/
	The Expressed Gene Anatomy Database (EGAD)	http://www.tigr.org/tdb/egad/egad.shtml
	Pseudomonas Genome Database	http://www.pseudomonas.com/
	Fusarium Database	http://mips.gsf.de/genre/proj/fusarium/
	CERB Target Gene Database	http://natural.salk.edu/CREB/
	The Arabidopsis Lipid Gene Database	http://lipids.plantbiology.msu.edu/
	Ribosomal protein gene database	http://ribosome.med.miyazaki-u.ac.jp/
	The Blood Group Antigen Gene Mutation Database (BGMUT)	http://www.ncbi.nlm.nih.gov/projects/mhc/xslcgi.fcgi?cmd=bgmut/home
	Rickettsia Genome Database	http://www.igs.cnrs-mrs.fr/mgdb/Rickettsia/
	Leukemia Gene Database	https://www.bioinformatics.org/legend/legend.htm
	Gene Ontology Database	http://www.geneontology.org/
	Genome Database List	http://grain.jouy.inra.fr/gendatabasemirror.html
	Organism Specific Genome Database	http://restools.sdsc.edu/biotools/biotools10.html
	Complete Genome Information Resource	http://www.tigr.org/tdb/
	5D-Genomics - Structural annotations for complete proteomes	http://www.sbg.bio.ic.ac.uk/5dgenomics
	Animal Genome Size Database	http://www.genomesize.com/
	ArkDB-Genome databases for farm and other animals	http://www.thearkdb.org/
	CoGenT++ Complete Genome Tracking database	http://cgg.ebi.ac.uk/cgg/cpp_sitemap.html

2 Databases of Database (DoD): DoD enlists 14 major categories comprising 719 databases. The number of databases under each category given in Table 1 and the subcategory in http://www.progene.in/DoD/DoD.html. Table 2.2 gives Description of different databases used in DoD. java script and html code were integrated and the DoD supports java script browser. A search function of the databases enables data access in each subcategory. Link 'Keyword help' keyword used to search specific database (RA jay Babu, 2005)

Construction of DoD: The Nucleic Acid Research 2005 Database issue has been referred to construct DoD due to the segregated highly structured categorization of databases. The modes of access of the database are given below:

DoD provides direct access to data from major type of the database. A direct link to the database home page of databases requiring manual input, search from data accessor download able content, Database requiring sequence, structure or category search are linked to database home page.

Table 2.1: List of database categories with number of database under each category.

S No.	Database Category (No of Entries)
1.	Nucleotide Sequence Database (76)
2.	RNA Sequence Database (89)
3.	Protein Sequence Database (120)
4.	Structure Database (71)
5.	Genomic Database (non-vertebrate) (162)
6.	Metabolic & Signaling Pathway (50)
7.	Human and other Vertebrate Genome (77)
8.	Human Genes and Diseases (89)

9.	M icroarray D ataand the G eneExpression D atabases	(48)
10.	Proteomics Resources(10)	
11.	O ther M olecular B iology D atabase(22)	
12.	O rganelle D atabase(19)	
13.	Plant D atabase(58)	
14.	Im m unological D atabase(21)	

Figure1 Screen-shot of the D oD show ing sub-category is of N ucleotide Sequence D atabases

S N o.	L ink to D atabase	D escription	Exam ples
1.	M ultimsearch	represents to a database that has m ore than one search option	N U R E B A SE [R uau *et al*, 2004]PRO D O R IC [M ünch *et al* 2003]
2.	D ow nload A ccess data	to access or dow nload data	PLA C E, G O SU B PlantProm [Shahm urodov *et al* 2003]
3.	Sequencesearch	to brow se for a sequence data	CD D [M archler-B auer *et al* 2005]
4.	Structure search	to search a carbohydrate	G lycan

		structures	
5.	Category search	search query against different categories	eFst
6.	Text and Sequence search	used to search both keyword and sequences	TIGRFAM [Haft *et al* 2003]

Table2: Description of different database links used in DoD.

A pop-up window (Fig 23) carries a brief description of the databases under the categories which provision to view full text of the published article and direct link to database home page. The respective home pages of the databases are linked through search option under link to database. During the search, results are displayed in a separate window enabling further databases can work with in the home page.

Figure 23 Screenshot of the DoD displaying GenBank pop-up window under Nucleotide Sequence Database.

3. MetaBase: MetaBase is designed to index and categorize more than 850 databases and approximately 700 data resources available on the internet at http://biodatabase.org. This related list of databases now exceeds much the Nucleic Acids Research Database Summaries' database (2006) and the Web Server Summaries' database (2006). Oxford University Press has kindly permitted us to share the content of these two databases.

MetaBase is a user-contributed resource permitting anyone to feel welcome to contribute and update. MediaWiki technology that has been used in the popular WikiPedia website has been used to run MetaBase ensuring the capacity to enrich the content and scope. Additional users who understand the power of the MediaWiki templating system can redesign in any aspect of the database organization.

More than 80 contributors who have been included in the author line of the MetaBase publication (Bolser et al., 2012)

MB

Main Page

Page | Discussion | Read | View source | View history | Go | Search

Welcome to MetaBase, the database of biological databases!

Navigation
Main Page
Recent changes
Random page
Help

See also:
Related sites

Toolbox
What links here
Related changes
Special pages
Printable version
Permanent link
Browse properties

Using MetaBase

- **Search** for resources.
- **Browse** resources by category.
- **Create** a new resource.

Introduction

What is MetaBase?
A short description of the aims, scope and status of the MetaBase project.
Documentation.
For detailed information about MetaBase, please see the documentation pages.
Related work.
Please see the section describing related projects.

Contact

- Via email here, or
- Visit the team as we relax in irc.freenode.net/#Bioinformatics (or via a web portal here).
- Tell us what you think!

The 10 'newest' resources are;
1. MERYB
2. Database for gene expression evolution (Bgee)
3. PROTICdb
4. Bioinformatist.org
5. Biopeople.org
6. OpenGenome.kr
7. Tigergenome.org
8. FemaleGenome.org
9. OpenGenome.net
More... RSS

Log in / create account with OpenID

MB

Special page

genome

Go | Search

Search results

genome

Search

Content pages   Multimedia   Help and Project pages   Everything   Advanced

Showing below up to **20** results starting with **#1**.

Create the page "**Genome**" on this wiki!

Page title matches

Chloroplast Genome Database/NAR
Chloroplast **genome** database
386 B (52 words) - 09:41, 18 January 2009

Genome Database for Rosaceae/NAR
220 B (29 words) - 13:44, 18 December 2008

TMBETA-GENOME
|name=TMBETA-**GENOME** ... Further, for a **genome** that have different chromosomes
(for example, human **genome** has 24 chromosomes) the data for each chromosome
is given individually. Fur
3 KB (473 words) - 04:54, 11 August 2009

Chloroplast Genome Database
|name=Chloroplast **Genome** Database |nar description=Chloroplast **genome**
database
741 B (103 words) - 07:00, 11 August 2009

Non-Human Genome Segmental Duplication Database
|name=Non-Human **Genome** Segmental Duplication Database ... presented at
http://projects.tcag.ca/humandup. Analyses based on previous **genome** assemblies

4 DBcat The DBcat (http://www.infobiogen.fr/services/dbcat) is a comprehensive catalog of biological databases, maintained and curated by a dedicated database, SRS. Infobiogen contains more than 500 database entries, classified by eight application domains: DNA, RNA, Protein, Protein structure, Genome, Mapping, Literature and Miscellaneous. The DBcat structured file data can be searched by means of an SRS Server or a dedicated Web interface. The file is available for downloading from Infobiogen anonymous ftp server (Figure 41)

AC	DBC00001
NAME	DBcat, Catalog of biological DataBases
DOMAIN	Literature
DESCRIPTION	The purpose of this catalog is to list all the existing
DESCRIPTION	biological databases, including relevant information
DESCRIPTION	about these databases
CHECKED	YES
AUTHORS	Claude Discala, Marion Ninnin
RA	Discala C., Ninnin M., Achard F., Barillot E.,
RA	Vaysseix G.
RT	DBcat: a catalog of biological databases
RL	Nucleic Acids Research, Vol. 27, 10-11, 1999.
ORIGINAL-SITE	INFOBIOGEN
ADDRESS	7, rue Guy Moquet
ADDRESS	94801 Villejuif cedex
ADDRESS	FRANCE
CONTACT	discala@infobiogen.fr
CONTACT	mninnin@infobiogen.fr
SUBMIT	http://www.infobiogen.fr/servic es/dbcat/file/dbcat_form.html
URL-FTP	ftp://ftp.infobiogen.fr/pub/db/ dbcat/
URL-WWW	http://www.infobiogen.fr/servic es/dbcat/
URL-QUERY	http://www.infobiogen.fr/srs
RELEASE	Daily
UPDATES	Daily
COMMENTS	Corrections and additions will be greatly appreciated
COMMENTS	as well as spontaneous submission
OTHER-SITE	-
ADDRESS	-
ADDRESS	-
URL-FTP	-
URL-WWW	-
URL-QUERY	-
UPDATES	-
COMMENTS	-
//	

The DB can provide the users with a variety of mode of access:

Download the file of info biogen ftp   ub/db/dbcat

Web interface home page with simple query by name interface:
http://www.infobiogen.services/dbcat
SRS server http://www.infobiogen.fr/

5. Conclusion: There are numerous public databases of biological data available on the internet. Generally, each of them stores a particular type of biological data such as nucleic acid sequence, primary, secondary and tertiary structures of proteins, compound based therapeutic molecular target, potential adverse related process involving metabolism DNA and protein expression analysis, mo ng others. Since such data are on vast and different databases, the integration become areal challenge. The primary aim of these four major Databases of Biological Data abase is to integrate information regarding wide variety of databases in a single database and provide an easy access to a number of databases through web sites that based on keywords of Brief description about the database is to enabling the user to know the contents of the database with specific keywords and database link. These databases of biological database should be updated periodically for the benefit of researchers and more bioinformatics biology keywords and other relevant information should also be added to the database.

References

1. Andrzej Polonski, K (2007) *Bioinformatics* Springer, New York.
2. B egeron, B (2002) *Bioinformatics computing.* Prentice Hal
3. Bolld D M & et (2012) MetaBase— the new ki database of biological *Nucleic Acid Research Vol 40,* 1,250-1,254.
4. Cynthia Gibas (2001) *Developing bioinformatics computer skills* OReily Media.
5. Galperin M Y (2004) The Molecular Biology Database Collection: *Nucleic Acid Research* .
6. kram-Ul-Haq M M (2006) Sugarcane bagasse pret reatment Attempt to enhance the production potential of cellulase by Hu m icola insolens TAS-13. *BIO KEM ISTRI 18(2)* 83-88.
7. P.A. jry Babu R B (2005) DooDD atabase Database se updated molecular biology databases. *In Silico Biology 50,054* .
8. Siva Kumar R R S M (2009) Bioinformatics glossary based Database Biological Database DBD. *J Biochem Tech 1(3)* 88-90.
9. Sushma M et (2003) *Data Mining Methods for the Life Soft Computing and Bioinformatics* John Wiley & Sons, Hoboken, New Jersey.