# BIOLOGICAL DATA MANAGING: A OPPORTUNITY FOR MOLECULAR BIOLOGY AND PHARMACEUTICALS.

***Sofiya Verma[1], Deepak Jain[2]**
[1]Research Scholar, Dept. of Pharmaceutics, Shri Ram Institute of Technology, Jabalpur (M.P.)
[2]Research Scholar, Dept.of Pharmacology, Shri Ram Institute of Technology, Jabalpur (M.P.)
***Email:** Sofiyaverma7@Gmail.Com

## ABSTRACT

Database management system (DBMS) A set of computer programs that control the creation, maintenance, and utilization of the databases of an organization DBMS uses application development: developing information systems by a process of investigation, analysis, design, implementation, and maintenance. Also called systems development life cycle (SDLC), information systems development. We demonstrate Biological Data Base Management (bdbm), an extensible database engine for biological databases. Bdbms started on the observation that database technology has not kept pace with the specific requirements of biological databases and that several needed key functionalities are not supported at the engine level. While bdbms aims at supporting several of these functionalities, this demo focuses on: (1) Annotation and provenance management including storage, indexing, querying, and propagation, (2) Local dependency tracking of dependencies and derivations among data items, and (3) Update authorization to support data curation. We demonstrate how bdbms enables biologists to manipulate their databases, annotations, and derivation information in a unified database system using the Purdue Ionomics Information Management System (PiiMS) as a case study.

**Keywords:** Database management system, Annotation, Manipulate, functionalities, implementation.

## INTRODUCTION

### I .Some Basic Term and Concept[1,2,13]

### Data Administration[1]

A data resource management function that includes responsibility for developing and maintaining the organization’s data dictionary, designing and monitoring the performance of databases, and enforcing standards for database use and security.

### Data Dictionary[1]

A software module and database containing descriptions and definitions concerning the structure, data elements, interrelationships, and other characteristics of an organization’s databases.

### Data Modeling[1]

A process where the relationships between data elements are identified and defined to develop data models.

### Data Planning[1]

A corporate planning and analysis function that focuses on data resource management. It includes the responsibility for developing overall information policy and data architecture for the firm’s data resources.

### Data Resource Management[1]

A managerial activity that applies information systems technology and management tools to the task of managing an organization’s data resources. Its three major components are database administration, data administration, and data planning.

### Database Administration[1]

A data resource management function that includes responsibility for developing and maintaining the organization’s data dictionary, designing and monitoring the performance of data bases, and enforcing standards for database use and security.

### Database Administrator[1]

A specialist responsible for maintaining standards for the development, maintenance, and security of an organization’s databases.

### Database Management Approach[1]

An approach to the storage and processing of data in which independent files are consolidated into a common pool, or database, of records available to different application programs and end users for processing and data retrieval.

### Database Management System (DBMS)[1]

A set of computer programs that controls the creation, maintenance, and utilization of the databases of an organization.

DBMS uses application development: developing information systems by a process of investigation, analysis, design, implementation, and maintenance. Also called systems development life cycle (SDLC), information systems development, or systems development.

### 1. Database Structures[2]

### Hierarchical

A logical data structure in which the relationships between records form a hierarchy or tree structure. The relationships among records are one-to-many, since each data element is related only tone element above it.

### Network

A logical data structure that allows many-to-many relationships among data records. It allows entry into a database at multiple points, because any data element or record can be related to many other data elements.

### Multidimensional

A database model that uses multidimensional structures (such as cubes or cubes within cubes) to store data and relationships between data.

## 2. Object-Oriented Structure[4]

### A. Relational

A logical data structure in which all data elements within the database are viewed as being stored in the form of simple tables. DBMS packages based on the relational model can link data elements from various tables as long as the tables share common data elements.

### B. Database and File Access[4, 11]

### Direct

A method of storage where each storage position has a unique address and can be individually accessed in approximately the same period of time without having to search through other storage positions.

### Query Language

A high level, human like language provided by a database management system that enables users to easily extract data and information from a database.

### Report Generator

A feature of database management systems packages which allows an end user to quickly specify a report format for the display of information retrieved from a data base.

## 3. Types of databases[4]

A. Analytical: A database of data extracted from operational and external databases to provide data tailored to online analytical processing, decision support, and executive information systems

B. End user: External image data warehouse: a central source of data that has been extracted from various organizational databases and standardized and integrated for use throughout the organization.

## II. BIOLOGICAL DATA BASE MANAGEMENT SYSTEM[11,12,13]

Life sciences are a case in point where biological databases have become essential to keep track of various informational bout experimentation and analysis. However, considerable amounts of biological data are still stored in flat files and spreadsheets and do not use DBMSs. This is mainly due to current database systems lacking key functionalities needed for biological data like efficient and native support for annotations, provenance, and data dependencies. Furthermore, biological databases often rely on community-based curation; evolve with rapidly changing semantics, and lack absolute authority to verify the correctness of information. Thus, the characteristics of annotations that need to be attached to the base data cannot be completely foreseen at design time. We are building bdbms, an extensible prototype database engine to support key functionalities needed by biological databases. These functionalities are implemented inside post- greSQL. In this demo, we focus on the following features:

annotation and provenance management, (2) local dependency tracking, and (3) update authorization. Bdbms makes fundamental advances in the use of biological databases through new native and transparent support mechanisms at the database system level. Bdbms treats annotations as first class objects. Bdbms allows adding annotations at multiple granularities, i.e., table, tuple, column, and cell levels, archiving and restoring annotations, and querying the data based on the annotation values.

We extended SQL into A-SQL to support the processing and querying of annotations. A-SQL allows annotations to be seamlessly propagated with query answers with minimal user intervention. Bdbms also includes a systematic approach for tracking dependencies among database items. When a database item is modified, bdbms tracks and annotates any other item that is affected by this modification and needs to be re-verified. This feature is particularly desirable in biological databases because many dependencies cannot be computed using coded functions, e.g., stored procedures and functions. Content-based authorization, i.e., the authorization is based not only on the identity of the user but also on the content of the data, is another feature that is integrated into bdbms. We demonstrate the main features of bdbms using the Purdue Ionomics Information Management System (PiiMS) a web-based system that collects and manages high throughput elemental profiling data and associated metadata on the experimental treatment, sample preparation, and instrument settings necessary to interpret results of mass spectrometry analyses in ionomics. PiiMS supports the entire process of planting, growing, harvesting, drying, and analyzing plants.

## III. SYSTEM OVERVIEW[5, 4, 7, 9, 14]

We give an overview of the main components of bdbms that we will demonstrate. We briefly describe bdbms' functionalities along with the extended SQL language (A-SQL).

### A. Annotation Management[5, 7]

Biologists use annotations as an important mean to communicate and share information about the base data generated from experiments and analyses. Annotations may represent comments about the data inside the database, the source of the data, references to published literature, or the setup and running of experiments. Despite their importance, annotations are not systematically supported as first-class objects inside current DBMSs.

### B. Local Dependency Tracking[5, 7]

It is often the case that data in biological databases are dependent on or derived from other data. The challenge is that most of these derivations cannot be simply modeled using coded functions, e.g., stored procedures or database triggers.  as a function. If a gene sequence is modified, the corresponding protein sequence(s) may become invalid. Similarly, we may store information about chemical reactions, e.g., substrates and reaction parameters. If any of These information is modified, then products of the reaction may become invalid. It is thus important to automatically track such dependencies and maintain the consistency of the data without burdening the users with extra checks. Bdbms enables the modeling of dependencies and derivations using the new concept of Procedural Dependency, an extension to Functional Dependencies. Procedural dependency allows to specify the dependency module or procedure and its characteristics, e.g., executable by the database or not, and invertible or not. For example, the following rule states that the protein sequence depends on the gene sequence through the lab experiment E that is neither executable by the database nor invertible:

Gene. Sequence

Protein Sequence

Such rules allows bdbms to track which items can be automatically re-computed and which items need to be marked as out-dated whenever a change occurs in the database. As a result, bdbms provides two important functionalities: (1) reporting out-dated data that needs to be re-evaluated, and (2) annotating query answers to highlight any out-dated data that's part of the results.

### C. Content-based Authorization[9]

In current DBMSs, users get permission to execute certain operations based on their identity, i.e., grant/revoke access model. Biological databases are usually a community- based and shared effort which may not fit with this model. For example, if only the lab administrator can modify the database, then (s) he becomes a bottleneck. Also, if all lab members can modify the data without revision, the credibility and authenticity of the data may be compromised.

Bdbms provides a monitoring system, termed Content-based Authorization, where the authorization is based on the identity of the user as well as the content of the modified data. The database administrator can turn the content-based approval feature ON or OFF for a certain table using the two following The content-based approval mechanism maintains a log of all update operations, i.e., INSERT, UPDATE, and DELETE. All non-approved updates will be visible with an annotation mentioning they were not approved yet. The logs are then revised by authorized users to approve/disapprove the Operations. If an operation is disapproved, bdbms executes an inverse operation that negates the effect of the original operation. This inverse operation is automatically generated and stored when the original operation is executed.

## RELATED TO FUTURE PROSPECT

### Public Databases for Molecular Biology

### I. Primary sequence databases[13]

The International Nucleotide Sequence Database (INSD) consists of the following databases.

DDBJ (DNA Data Bank of Japan)

EMBL Nucleotide DB (European Molecular Biology Laboratory)

GenBank (National Center for Biotechnology Information)

UniProtKB (Universal Protein Resource Knowledgebase)

The four largest databases are Gene Bank, (the U.S.'s collection of various biological data), EMBL, (Europe's collection of nucleotide sequence data), DDBJ, (DNA Data Bank of Japan), and UniProt, (Universal Protein Resource). Gene Bank is a service provided by NCBI, which stores sequence data and "biological sequence related data." EMBL is a service provided by EBI, the European Bioinformatics Institute, and provides a collection of nucleotide sequence data, as its name suggests. DDBJ is a nucleotide database. UniProt is a high-quality and comprehensive universal protein resource. It provides translations of sequences from EMBL, Gene Bank, and DDBJ, in its UniProt Knowledgebase (UniProtKB).

These databanks represent the current knowledge about the sequences of all organisms. They interchange the stored information and are the source for many other databases.

Note that Gene Bank, EMBL, and DDBJ work very closely with one-another, and as a result what one can find from one of these sources they can find from any of the other two and vice-versa.

### II. Meta-databases[12]

Strictly speaking a Meta database can be considered a database of databases, rather than any one integration project or technology. They collect data from different sources and usually make them available in new and more convenient form, or with an emphasis on a particular disease or organism.

Entrez (National Center for Biotechnology Information).

euGenes (Indiana University).

Gene Cards (Weizmann Inst.).

SOURCE (Stanford University).

mGen containing four of the world biggest databases GenBank, Refseq, EMBL and DDBJ - easy and simple program friendly gene extraction.

Bioinformatic Harvester (Karlsruhe Institute of Technology) - Integrating 26 major protein/gene resources.

MetaBase (KOBIC) - A user contributed database of biological databases.

ConsensusPathDB - A molecular functional interaction database, integrating information from 12 other databases.

### **III.** Genome **Databases**[7, 3]

These databases collect organism genome sequences, annotate and analyze them, and provide public access. Some add curation of experimental literature to improve computed annotations. These databases may hold many species genomes, or a single model organism genome.

CAMERA Resource for microbial genomics and met genomics.

Corn, the Maize Genetics and Genomics Database.

Ensemble provides automatic annotation databases for human, mouse, other vertebrate and eukaryote genomes.

ERIC (Enterpathogen Resource Integration Center) Curated database containing annotated genome data for five enteropathogens - Escherichia coli, Shigella, Salmonella, Yersinia enterocolitica, and Y. pestis.

Flybase, genome of the model organism Drosophila melanogaster.

MGI Mouse Genome (Jackson Lab.).

JGI Genomes of the DOE-Joint Genome Institute provides databases of many eukaryote and microbial genomes.

National Microbial Pathogen Data Resource. A manually curated database of annotated genome data for the pathogens Campylobacter, Chlamydia, Chlamydophila, Haemophilus, Listeria, Mycoplasma, Neisseria, Staphylococcus, Streptococcus, Treponema, Ureaplasma, and Vibrio.

Saccharomyces Genome Database, genome of the yeast model organism.

Viral Bioinformatics Resource Center Curated database containing annotated genome data for eleven virus families.

The SEED platform for microbial genome analysis includes all complete microbial genomes, and most partial genomes. The platform is used to annotate microbial genomes using subsystems.

Xenbase, genome of the model organism Xenopus tropicalis and Xenopus laevis .

Wormbase, genome of the model organism Caenorhabditis elegans.

Zebrafish Information Network, genome of this fish model organism.

TAIR, the Arabidopsis Information Resource.

UCSC Malaria Genome Browser, genome of malaria causing species (Plasmodium falciparumata and others).

RGD Rat Genome Database: Genomic and phenotype data for Rattus norvegicus.

### **IV. Genome Browsers**[7]

Genome Browsers enable researchers to visualize and browse entire genomes (most have many complete genomes) with annotated data including gene prediction and structure, proteins, expression, regulation, variation, comparative analysis, etc. Annotated data is usually from multiple diverse sources.

Integrated Microbial Genomes (IMG) system by the DOE-Joint Genome Institute.

UCSC Genome Bioinformatics Genome Browser and Tools (UCSC).

Ensembl The Ensembl Genome Browser (Sanger Institute and EBI).

GBrowse The GMOD GBrowse Project.

Pathway Tools Genome Browser.

X: Map A genome browser that shows Affymetrix Exon Microarray hit locations alongside the gene, transcript and exon data on a Google maps api.

Viral Genome Organizer (VGO) A genome browser providing visualization and analysis tools for annotated whole genomes from the eleven virus families in the VBRC (Viral Bioinformatics Resource Center) databases.

Apollo Genome Annotation Curation Tool A cross-platform, JAVA-based standalone genome viewer with enterprise-level functionality and customizations. The standard for many model organism databases.

SEED viewer for visualizing and interrogating the SEED database of complete microbial genomes

Integrated Genome Browser (IGB) A cross-platform, Java-based desktop genome viewer.

Argo Genome Browser A free and open source standalone Java-based genome browser for visualizing and manually annotating whole genomes.

OMGBrowse An extensible automated genome annotating service. Based on JBrowse.

### **V. Protein sequence databases**[5, 6]

UniProt Universal Protein Resource (UniProt Consortium: EBI, Expasy, PIR).[5]

PIR Protein Information Resource (Georgetown University Medical Center (GUMC).

Swiss-Prot Protein Knowledgebase (Swiss Institute of Bioinformatics).[6]

PEDANT Protein Extraction, Description and ANalysis Tool.

PROSITE Database of Protein Families and Domains.

DIP Database of Interacting Proteins (Univ. of California).

Pfam Protein families database of alignments and HMMs (Sanger Institute).

PRINTS PRINTS is a compendium of protein fingerprints (Manchester University).

ProDom Comprehensive set of Protein Domain Families (INRA/CNRS).

SignalP 3.0 Server for signal peptide prediction (including cleavage site prediction), based on artificial neural networks and HMMs.

SUPERFAMILY Library of HMMs representing superfamilies and database of (superfamily and family) annotations for all completely sequenced organisms.

Annotation Clearing House a project from the National Microbial Pathogen Data Resource.

**VI.** Protein structure **Databases**[7, 8]

Protein Data Bank (PDB) (Research Collaboratory for Structural Bioinformatics (RCSB).[7]

Protein Model Portal (PMP) Meta database that combines several databases of protein structure models (Biozentrum, Basel, Switzerland).[8]

CATH Protein Structure Classification.

SCOP Structural Classification of Proteins.

SWISS-MODEL Server and Repository for Protein Structure Models.

ModBase Database of Comparative Protein Structure Models (Sali Lab, UCSF).

**VII.** Protein-protein interactions

BioGRID A General Repository for Interaction Datasets (Samuel Lunenfeld Research Institute).[9]

STRING: STRING is a database of known and predicted protein-protein interactions. (EMBL)

DIP Database of Interacting Proteins.

BIND Biomolecular Interaction Network Database .

**VIII. Signaling Pathway Databases**

Netpath - A curated resource of signal transduction pathways in humans.

Reactome.

NCI-Nature Pathway Interaction Database.

**IX.** Metabolic pathway **Databases**

BioCyc Database Collection including EcoCyc and MetaCyc.

KEGG PATHWAY Database (Univ. of Kyoto).[10]

MANET database (University of Illinois).[11]

Reactome (Cold Spring Harbor Laboratory, EBI, Gene Ontology Consortium)[12]

**X. Microarray databases**

Main article: Microarray databases.

ArrayExpress (European Bioinformatics Institute).

Gene Expression Omnibus (National Center for Biotechnology Information).

GPX (Scottish Centre for Genomic Technology and Informatics).

maxd (University of Manchester).

Stanford Microarray Database (SMD) (Stanford University).[11]

**XI. Mathematical Model Databases**

Biomodels Database.

CellML.

XII. PCR / Real time PCR primer Databases

PathoOligoDB: A free QPCR oligo database for pathogens .

**XIII. Specialized databases (in alphabetical order).** [4, 5, 8, 10-12]

Antibody Central Antibody information database and search resource.

BIOMOVIE (ETH Zurich) movies related to biology and biotechnology.

CGAP Cancer Genes (National Cancer Institute).

Clone Registry Clone Collections (National Center for Biotechnology Information).

Connectivity map Transcriptional expression data and correlation tools for drugs.

CTD The Comparative Toxicogenomics Database describes chemical-gene-disease interactions

DBGET H.sapiens (Univ. of Kyoto).

DiProDB A database to collect and analysed thermodynamic, structural and other dinucleotide properties.

Edinburgh Mouse Atlas .

GreenPhylDB (A phylogenomic database for plant comparative genomics).

GyDB The Gypsy Database of Mobile Genetic Elements (Universitat de València).

Genome Database for Rosaceae (International Genomics and Genetics Database for Rosaceous crops).

GDB Hum. Genome Db (Human Genome Organization).

HGMD disease-causing mutations (HGMD Human Gene Mutation Database).

HUGO (Official Human Genome Database: HUGO Gene Nomenclature Committee).

HvrBase++ Human and primate mitochondrial DNA.

INTERFEROME The Database of Interferon Regulated Genes.

List with SNP-Databases .

NCBI-UniGene (National Center for Biotechnology Information).

OMIM Inherited Diseases (Online Mendelian Inheritance in Man).

OrthoMaM (A database of Orthologous Mammalian Markers).

p53 The p53 Knowledgebase.

PhenCode linking human mutations with phenotype.

Plasma Proteome Database Human plasma proteins along with their isoforms.

Polygenic Pathways Genes and risk factors implicated in Alzheimer's disease, Bipolar disorder or Schizophrenia.

SHMPD: The Singapore Human Mutation and Polymorphism Database.

SciClyc an Open-access database to shared antibodies, cell cultures, and documents for biomedical research.

XTractorDiscovering Newer Scientific Relations across Pub Med Abstracts. A tool to obtain manually annotated relationships for Proteins, Diseases, Drugs and Biological Processes as they get published in Pub Med.

## Acknowledgement

It is my great privilege, pride and honor in expressing my deepest sense of gratitude to my esteemed inspiring, untiring mentor **Sofiya Verma, Dept.of Pharmaceutics** Shri Ram Institute Of Technology., in presenting this review entitled "**Biological Data Managing: A Opportunity For molecular biology and Pharmaceutical**" Her constant inspiration, memorable guidance, innovative ideas, instigation, at various stages and above all her untiring valuable support has brought out this thesis to an exquisite workmanship and great success.

## REFERENCES

1. Brooks bank C., Camon E., Harris M.A., Magrane M., Martin M., Mulder N., O'Donovan,C., Parkinson H., Tuli M., Apweiler R. *et al.* The European Bioinformatics Institute's data resources. Nucleic Acids Res., (2003), 31, 43–50.
2. Miyazaki S., Sugawara H., Ikeo K., Gojobori T., and Tateno Y. DDBJ in the stream of various biological data. Nucleic Acids Res., (2004), 32, D31–D34.
3. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J. and Wheeler D.L. GenBank: update. Nucleic Acids Res., (2004), 32, D23–D26.
4. Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M. *et al.* The Universal Protein Resource (UniProt). Nucleic Acids Res., (2005), 33, D154–D159.
5. Mulder N.J., Apweiler,R., Attwood,T.K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., Bork P. *et al.* The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res., (2003), 31, 315–318.
6. Golovin A., Oldfield T.J., Tate J.G., Velankar S., Barton G.J., Boutselakis H., Dimitropoulos D., Fillon J., Hussain A., Henrick K. *et al.* E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res., (2004), 32, D211–D216.
7. Clamp M., Andrews D., Barker D., Bevan P., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V. *et al.*: accommodating comparative genomics. Nucleic Acids Res., (2003) 31, 38–42.
8. Fleischmann A., Darsow M., Degtyarenko K., Fleischmann W., Boyce S., Axelsen K.B., Bairoch,A., Schomburg D., Tipton, K.F. and Apweiler,R. IntEnz, the integrated relational enzyme database. Nucleic Acids Res., (2004),32, D434–D437.
9. Hermjakob H., Montecchi-Palazzi L., Lewington C., Mudai S., Kerrien S., Orchard S., Vingron M., Roechert B., Roepstorff P. and Apweiler R. IntAct: an open source molecular interaction database. Nucleic Acids Res., (2004), 32, D452–D455.
10. Lombard V., Camon E.B., Parkinson H.E., Hingamp P., Stoesser G. and Redaschi N. EMBL-Align: a new public nucleotide and amino acid multiple sequence alignment database. Bioinformatics, (2002), 18, 763–764.
11. Zdobnov E.M., Lopez R., Apweiler R. and Etzold T. The EBI SRS server—new features. Bioinformatics, (2002) 18, 1149–1150.
12. Zhdanov E.M., Lopez R., Apweiler R. and Etzold T. The EBI SRS server—recent developments. Bioinformatics, (2002) 18, 368–373.
13. Leinonen R., Nardone F., Oyewole O., Redaschi N. and Stoehr P, The EMBL sequence version archive. Bioinformatics, (2003) 19, 1861–1862.
14. Pearson W.R. Using the FASTA program to search protein and DNA sequence databases. Methods Mol. Biol., 1994 24, 307–331.

***********