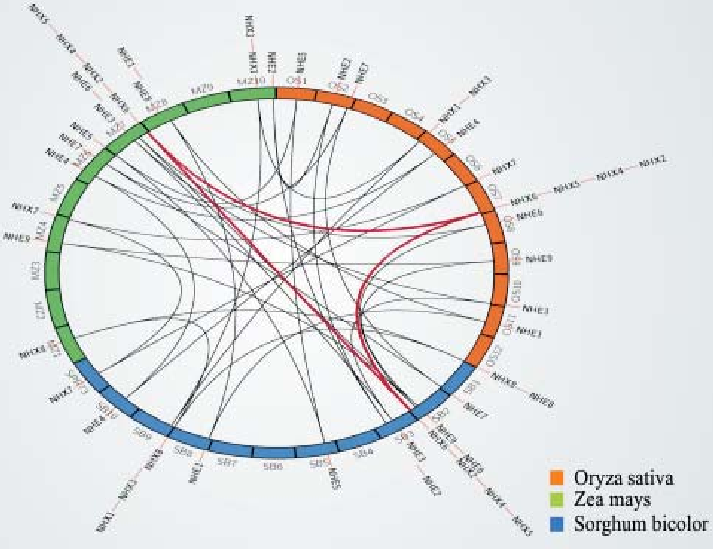
Principles of

Biological Databases

- Prof. P.B. Kavi Kishor
- Mr. L.N. Chavali



Himalaya Publishing House

Principles of Biological Databases

Editors

Prof. P.B. Kavi Kishor
Professor of Genetics,
Osmania University,
Hyderabad.

L.N. Chavali
Technology Consultant.

Himalaya Publishing House

MUMBAI • NEW DELHI • NAGPUR • BENGALURU • HYDERABAD • CHENNAI • PUNE • LUCKNOW • AHMEDABAD
• ERNAKULAM • BHUBANESWAR • INDORE • KOLKATA • GUWAHATI

First Edition : 2013

Published by : Mrs. Meena Pandey for **Himalaya Publishing House Pvt. Ltd.,**
“Ramdoot”, Dr. Bhalerao Marg, Girgaon, **Mumbai - 400 004.**
Phone: 022-23860170/23863863, Fax: 022-23877178
E-mail: himpub@vsnl.com; Website: www.himpub.com

Branch Offices :

New Delhi : “Pooja Apartments”, 4-B, Murari Lal Street, Ansari Road, Darya Ganj,
New Delhi - 110 002. Phone: 011-23270392, 23278631; Fax: 011-23256286

Nagpur : Kundanlal Chandak Industrial Estate, Ghat Road, Nagpur - 440 018.
Phone: 0712-2738731, 3296733; Telefax: 0712-2721215

Bengaluru : No. 16/1 (Old 12/1), 1st Floor, Next to Hotel Highlands, Madhava Nagar,
Race Course Road, Bengaluru - 560 001.
Phone: 080-32919385; Telefax: 080-22286611

Hyderabad : No. 3-4-184, Lingampally, Besides Raghavendra Swamy Matham, Kachiguda,
Hyderabad - 500 027. Phone: 040-27560041, 27550139; Mobile: 09390905282

Chennai : No. 8/2, Madley 2nd Street, Ground Floor, T. Nagar, Chennai - 600 017.
Phone: 044-28144004/28144005; Mobile: 09345345051

Pune : First Floor, "Laksha" Apartment, No. 527, Mehunpura, Shaniwarpeth
(Near Prabhat Theatre), Pune - 411 030. Phone: 020-24496323/24496333;
Mobile: 09370579333

Lucknow : Jai Baba Bhavan, Church Road, Near Manas Complex and Dr. Awasthi Clinic,
Aliganj, Lucknow - 226 024 (U.P.). Phone: 0522-2339329, 4068914;
Mobile: 09307501550

Ahmedabad : 114, “SHAIL”, 1st Floor, Opp. Madhu Sudan House, C.G. Road, Navrang Pura,
Ahmedabad - 380 009. Phone: 079-26560126; Mobile: 09377088847

Ernakulam : 39/176 (New No: 60/251) 1st Floor, Karikkamuri Road, Ernakulam,
Kochi - 682011, Phone: 0484-2378012, 2378016; Mobile: 09344199799

Bhubaneswar : 5 Station Square, Bhubaneswar - 751 001 (Odisha).
Phone: 0674-2532129, Mobile: 09338746007

Indore : Kesardeep Avenue Extension, 73, Narayan Bagh, Flat No. 302, IIIrd Floor,
Near Humpty Dumpty School, Indore - 452 007 (M.P.). Mobile: 09301386468

Kolkata : 108/4, Beliaghata Main Road, Near ID Hospital, Opp. SBI Bank,
Kolkata - 700 010, Phone: 033-32449649, Mobile: 09883055590, 07439040301

Guwahati : House No. 15, Behind Pragjyotish College, Near Sharma Printing Press,
P.O. Bharalumukh, Guwahati - 781009, (Assam).
Mobile: 09883055590, 09883055536

DTP by : HPH, Editorial Office, Bhandup **(Rupeshri)**

Printed at : A to Z Printers, New Delhi. On behalf of HPH.

Preface

This book starts right from the basics with databases and Structured Query Language (SQL). Prior database or SQL knowledge is not necessary, as this book covers everything from database design to creating your first database and understanding how the SQL language is used with databases. You will need to follow its instructions for creating the book's gene database, as this is used for all the examples in SQL.

The objective of Section I, Chapters 1 to 8 is to understand the basic concepts and practices that can be used during design and development of biological database systems. The course contents put more emphasis on Relational Database Management Systems (RDBMS) as it is an accepted database standard for biological systems and present it in a more detailed manner. Entity Relation Modeling and schema representation have been covered extensively through a case study along with third form normalization with a case study. At the end of the Chapter 8, Centralized Database approach has been described briefly with an illustration. Most of the biological systems present in today's word are centralized.

The Chapters 9 to 14 will focus on Structured Query Language, or SQL as it is usually abbreviated. SQL is useful to create the biological database and inserting and extracting biological data. Therefore, going through Chapters 1 to 8 is essential to understand the theory and concepts behind database systems. The SQL examples in Chapters 10 and 11 comply with the modern SQL standards set by organizations such as the American National Standards Institute (ANSI) and the International Standards Organization (ISO). The SQL queries in this chapter are supported by most modern database systems.

The test data that has been used in this book is meant for demonstration purpose only. All genes in the test data may be translated into proteins, however only 8 of them have been shown as proteins. The ER diagram, logical design and physical design may not truly reflect the real time scenario, however, can be modified to accommodate such scenarios. The maximum field lengths in the table design are assumed to be 50 for the sake of simplicity. The gene, organism, genus and species are shown in italics displayed as created in the database.

PL/SQL, DBA and transaction management and performance tuning, etc., are outside the scope of this book. However, Chapters 10 and 11 explains about trigger mechanism, various roles in database and modes of transactions in DBMS. SQL queries that are related to some mathematical functions and string functions, but left and right outer joins (both left and right) have not been addressed in Section I. Therefore, the reader is requested to refer to SQL manuals for any missing content. The database is created and tested using Oracle 8i, 9i and 10g versions. In the appendix, a cursory view of Oracle basics and PL/SQL has been provided as a reference for those who are new to these concepts.

Chapters 14 to 21 explain briefly the basic concepts in data warehouse, approaches to data warehouse building, steps to build a warehouse including dimensional modeling and at the end of the section a case study in plant bioinformatics has been presented with an example to understand the steps involved in warehouse process for building data marts for analysis.

In the second part of the book, many important biological databases and their uses in biology has been described. Since too large number of biological databases exist in literature, only a selected few (those that are often referred by biologists) are described in this book. Types of databases, models of databases, primary nucleic acid and protein databases, secondary protein databases, composite sequence databases, meta-databases, genomic, proteomic and other databases have been described in detail. Search engines for literature have been added for gaining access to the published literature in Journals and Books. Major genome projects (about 19), genomic databases of human, animals, fungi, microorganisms, plant crop genome databases have been described in brief. Finally, organellar and pathway databases have been added.

We hope that this introductory book exclusively on principles of biological databases would serve the basic needs of the beginners like undergraduate and graduate students in Biological Sciences, Biochemistry, Bioinformatics, Pharmacy, Biotechnology and research scholars too.

Prof. P.B. Kavi Kishor
L.N. Chavali

Acknowledgments

There are many individuals whom we like to thank, without whose efforts this volume would not have become a reality. First and foremost, we are grateful to Late Shri. Bh. Krishna Rao, Database Architect, and to Dr. P. Nataraj Sekhar for reviewing the contents of the book. We gratefully acknowledge our parents for being a constant source of inspiration and encouragement.

It is a great pleasure to work with Mr. S. Anil Kumar and Ms. P. Hima Kumari, Research Scholars, Department of Genetics, Osmania University, Hyderabad who have helped us in many ways and for verifying database tables with test data and for providing the necessary examples, inputs and feedback.

We like to thank Prof. J. Sobhanaditya, Department of Biochemistry, Osmania University for his steadfast patience for going through the entire book. We would like to specially thank Dr. S. Prashanth, Founder Director, Bioclues, Hyderabad for his help in editing process and for writing part of the Biological Databases. We have developed a special bond with Dr. Prashanth both professionally and personally for the past few months, who has provided us necessary enthusiasm to work on such projects.

Finally, we like to thank M/s. Himalaya Publishing House Pvt. Ltd., Mumbai for bringing a shape to this book and for publishing it on time.

Prof. P.B. Kavi Kishor
L.N. Chavali

Contents

Section I: Database Principles for Biologists

Section II: Biological Databases

Detailed Contents

Section I: Database Principles for Biologists

Section II: Biological Databases

Section I:

Database Principles for Biologists

1 INTRODUCTION

A database is an integrated collection of automated data objects related to one another in the support of a common purpose. The data are the known facts and concepts that are created or recorded in computer readable form and have an implicit meaning. The terms data and information are often used interchangeably and in the wrong context. Information is derived from data.

Each eukaryotic cell contains various cell organelles such as nucleus, mitochondria, chloroplasts etc. These organelles contain DNA which in turn has nucleotide sequences. The analogy to the database is the composition of cell-organelle. The cell represents the database, the organelles represent objects such as files or tables, the DNA in the organelle represent the records, and the nucleotide sequences inside DNA represent the data elements. The database software comes in two parts (a) the application programs which process the user data and (b) a general purpose system that manages the data definition, organization, storage and retrieval. The packaged software that manages storage and retrieval of data is called database management system or DBMS. DBMS is for persistent, consistent and application independent storage and management of data, i.e., DBMS facilitates creation and maintenance of a computerized database. The DBMS is a logical entity that sits between the application programs and the database. It helps in management of data. It responds to the requests from the application program. The Figure 1 below briefly illustrates the relationship.

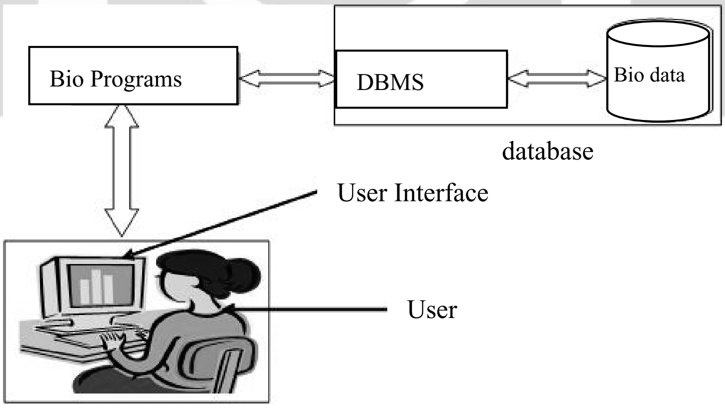


Fig. 1 - Managing the Database

The system that comprises of DBMS, data and sometimes including the application is called database system. In other words, the system that comprises of different objects such as tables and programs to manage the data are called database systems. The physical location and implementation of the database is transparent to application programs. Some of the popular DBMS include:

- Oracle
- DB2
- MS SQL
- MS Access
- Ingress
- PostgreSQL
- MySQL

The main reason to study databases is because of increasing shift from computation based systems to information based systems. The databases are increasing in diversity and volume as many applications deal with large volumes of data and hence information. Ex: genome project, digital libraries. Database helps to

1. Organize the information
2. Serve as a place to put data in logical manner so that retrieval of data on demand becomes easier.
3. Act as a resource for other databases and tools.

Database tables as shown in Table 1 describe some of the records of nucleotide sequences.

Table 1 - Sample Tables with Data

GENE	NAME	LOCATION	ORGANISM	SEQUENCE
	NHX1	L1p2	*Sorghum*	ATCGGCTAATCG
	AKT1	14q32	*Oryza*	CGGCAGGACC

ORGANISM	NAME	FAMILY NAME	GENUS	SPECIES
	Sorghum	Poaceace	*Sorghum*	*bicolor*
	Oryza	Poaceace	*Oryza*	*Sativa*

FUNCTION	NAME	DESCRIPTION
	NHX1	Provides resistance to salt
	AKT1	Provides resistance to salt

Some of the key technological terms of the database are:

Metadata – Description of the database that is stored in a catalog i.e., data of data. DBMS uses metadata information for storing and retrieval of information.

Data Independence – Insulation between the data and application programs.

Data Abstraction – A way to hide data storage using a data model.

Data View - Each user may see a different view of the database and the data that is needed by the user.

There are two approaches to data management:

1. File based Systems
2. Database Systems

FILE BASED SYSTEMS

A flat-file structure is good only for extremely simple databases. A flat-file structure is not practical for most business applications. A "flat file" is a plain text or mixed text which usually contains one record per line. The attributes in each record are separated by delimiters such as commas or have a fixed length. No structural relationships exist between the records. Data stored in files have a specific format and the programs that use these files depend on knowledge about that format. *Example:* The flat files are widely being used to maintain the configuration data of the systems. The given below example illustrates the usage of flat files in biological applications.

The example of the flat file for bacteria is shown in Table 2

Table 2 – The Flat File Format

Bacteria Name	**Family Name**	**Genus Name**	**Species Name**
Escherichia	Enterobacteriaceae	*Escherichia*	*coli*
Pseudomonas	Pseudomonodaceae	*Pseudomonas*	*syringae*
Agrobacterium	Rhizobiaceae	*Agrobacterium*	*tumefaciens*

The file based systems have the following disadvantages

- Data Redundancy and Inconsistency
- Unanticipated Queries
- Data Isolation/dependency
- Concurrent Access Anomalies
- Security Problems
- Integrity Problems

DATABASE SYSTEMS

A database is a collection of permanently stored data that is:

- **Logically Related -** data relates to other data
- **Shared -** many users may access data

- **Protected -** access to data is controlled
- **Managed -** data has integrity and value

The purpose of a database system is to bridge the gap between information and data – the data stored in memory or on disk must be converted to usable information. A database is a model of a real world system. The contents (sometimes called the extension) of a database represent the state of what is being modeled. Changes in the database represent events occurring in the environment that change the state of what is being modeled. It is appropriate to structure a database to mirror what it is intended to model.

The primary goal of the database is:

1. Minimize data redundancy i.e., duplication of data and store the data in multiple files or tables
2. Ability to change in data structure without making changes in the programs that process the data

Examples of some of the popular commercial databases include:

- Medical Records
- Library Catalogs
- Bank Accounts
- Telephone Directories
- Airline Bookings and so on...

Examples of some of the popular biological databases include:

1. Micro Array
2. Structure Database
3. Mass Spectrometry
4. Specific Organism
5. Functional Annotations
6. Bibliographic database

The database system allows users of the system to Store/Update/Retrieve/organize/protect their data. The database system has users of type such as given below.

End Users – Use the database system to achieve his/her goals

Application Users – Write software programs to allow end users to interface with the database systems

Database Administrator (DBA) – Manage and maintain the database system

Systems programmer – Creates or modifies the database software

Database Benefits

Following are some of the key benefits of the database system.

1. Minimal Data Redundancy
2. Data Consistency
3. Data Integration
4. Data sharing among multiple users
5. Data Persistence
6. Application Development Ease
7. Data Independence
8. Backup and Recovery Services

Data Redundancy

File based applications may have private data specific to each application and all such private data is stored in files. This can lead to considerable redundancy in data storage thereby wasting storage space. Ex: A gene and a protein may store family related data in its respective private files. Table 3 shows the redundancy of gene - NHX$_1$.

Table 3 - Gene NHX$_1$ Redundancy

	GENE NAME	ORGANISM	SPECIES	FAMILY	FUNCTION DESCRIPTION
REPORT	NHX$_1$	Sorghum	bicolor	Poaceace	Provides resistance to salt
	NHX$_1$	Oryza	sativa	Poaceace	Provides resistance to salt

In database system, the goal is to control the redundancy in storage but may not mean to eliminate all redundancy.

Data Consistency

Both gene and its ISO-forms have biological functions. Therefore, a function is represented by different variants of the same gene. The functions of all ISO-forms of a gene may be same or different. There may be multiple entries of a given function in the database. If the system is not aware of this duplication, it can lead to 'inconsistency', which means that the entries of the function for the given gene-variants may be incorrect or contradictory information of the function is saved in the same database. A database which is in inconsistent state can only provide conflicting or wrong information. If the redundancy is controlled instead of removing it, then the system should ensure that the database is not inconsistent.

Data Integrity

Data integrity is expressed in terms of quality and the reliability of the data. In a broader sense, data integrity includes the protection of the database from unauthorized access and unauthorized changes. One of the major functions of DBMS is to support the task of bringing only correct and consistent data into the database. Ex: Sequence errors of gene or protein.

Data Sharing

A database system allows several users to access and share the database concurrently as shown in Figure 2. Different users of the system query the database for different data simultaneously. Such concurrent use of data increases the economy of a system.

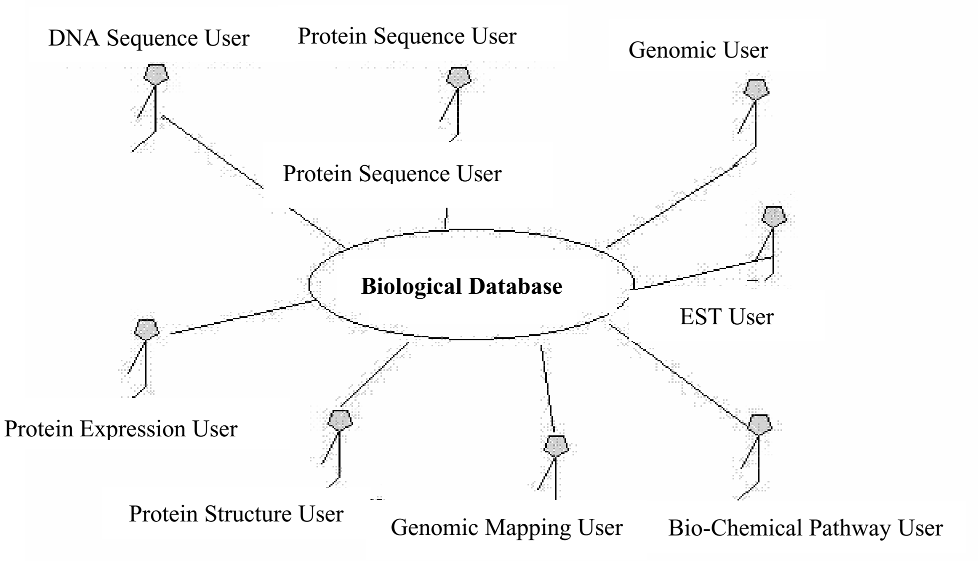


Fig. 2 - Concurrent Use of Data by Different Users

Data Persistence

Data persistence means that in a DBMS all data is maintained as long as it is not deleted explicitly. The life span of data needs to be determined directly or indirectly be the user and must not be dependent on system features. Additionally data once stored in a database must not be lost.

Application Development Ease

Once DB design in place, application development becomes easy due to availability of logical design before application development.

Back-up and Recovery Services

Back-up is a way to archive the information on of the medium. The recovery process helps to safeguard data going awry. The back and recovery plans are part of contingency framework.

Data Independence

There are two categories of data independence as mentioned below.

Logical Data Independence: The capacity to change the conceptual schema without having to change the external schemas and their application programs i.e., the capacity to expand or reduce the database without having to change the external schemas and their application programs.

Physical Data Independence: The capacity to change the internal schema without having to change the conceptual schema. When a schema at a lower level is changed, only the **mappings** between this schema and higher-lever schemas need to be changed in a DBMS that fully supports data independence. The higher-level schemas themselves are unchanged. Hence, the application programs need not be changed since they refer to the external schemas.

Database Limitations

Some of the limitations of database have been listed below.

- May be limited by the complexities of the data model.
- May be limited by special operations that are not supported by DBMS.
- Overhead limitation because of security and recovery procedures.

Database Architecture

As per ANSI/SPARK architecture, the database architecture has 3 views.

1. Internal View
2. Conceptual View
3. External

The diagram as shown in Figure 3 describes the 3 different views of the architecture.

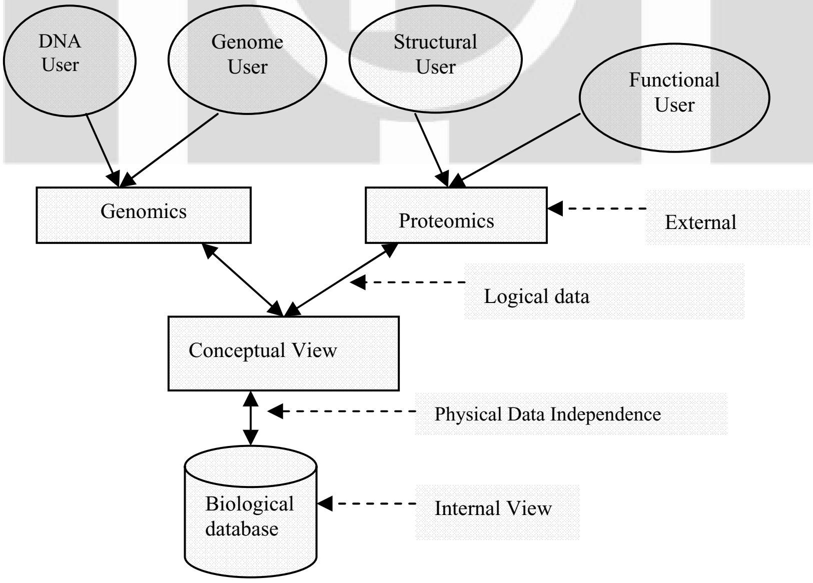


Fig. 3 - Database Architecture

External View: This view is closest to the users and is more concerned with the way the data is viewed by individual users. In this view, the data is presented in useful form and some parts of the data in the data may be hidden. This is level meant for users and application programmers.

Internal View: It deals with physical storage of data like structure of records on disk. At this level, there is a single view of the total database as actually stored. This view is meant for systems designers and database system programmers.

Conceptual View: This view is a level of 'indirection' between internal and external views. It can be thought as a view of community of users. All external views, each of which have their respective representation of the database will be dealt by a single conceptual view. It is concerned with organization of data such as abstractions that are used to remove the unnecessary details of internal level. This view is meant for application programmers and database administrators. Conceptual-level concepts permit us to model the applications independent of any particular data model and the conceptual model is close the way the users perceive the data. They are a convenient mechanism to allow the structure of a database to evolve over time as the environment being modeled, user needs and information requirements change. Conceptual modeling provides a framework for developing a database structure or schema from the top down. It also helps identify and define the entities, attributes of each and the relationships among the entities. The very popular conceptual-level model is the entity-relationship model which is covered in the later part of this chapter. The mappings of the three views during information translation provide data independence. Any changes to the data at the internal level should not affect the schema in the conceptual level. This is called physical independence of data. The changes in the data at the conceptual level should not impact the user views. This is called the logical independence of data.

Structured Data

A fundamental feature of the database approach is that the database system does not only contain the data but also the complete definition and description of these data. These descriptions are basically details about the extent, the structure, the type and the format of all data and, additionally, the relationship between the data. This kind of stored data is called metadata ("data about data"). A simple example how data can be described in a database is shown in Table 4. The table has three columns. The first column is gene, the second column as accession number and third column as protein that determines it as a gene or protein. All the columns in this table are coded as strings.

Table 4 - Sample Gene Table in Database

GENE	ACCNO	PROTEIN
NHX_1	EU482408	N

Biological Database and its Organization

Biological database is the database of sequence. As we know, there are three kinds of biological sequences namely protein, DNA and RNA. The volume of biological data has been growing in recent year and is doubled in size every 15 or 16 months. Because of huge volume of biological data, biological database has greatly developed and became a part of the biologist's everyday toolbox. The bio database is queried everyday 40,000 times on an average. To make biological database work efficiently, good database design and efficient search methods are required.

The data in molecular databases is organized as

1. Curated data
2. Archival data

Curated Data

- Non-redundant data, only one sequence per gene.
- Data contains value added information entered and validated by experts.

Archival Data

- Nothing but repository of information.
- May contain redundant data. For example, archival may have several sequences for the same gene.
- May contain wide range of data. Ex: partial sequences, patent sequences.
- Contains data from genome projects, patent offices and individual scientists.

There are over 600 different biological databases. Protein and Sequence databases are widely referred and used databases in biology.