Biotechnology:

Concepts, Methodologies, Tools, and Applications

Information Resources Management Association
USA

IGI Global
DISSEMINATOR OF KNOWLEDGE

Published in the United States of America by
IGI Global
Medical Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Information Resources Management Association, issuing body.
Title: Biotechnology : concepts, methodologies, tools, and applications / Information Resources Management Association, editor.
Other titles: Biotechnology (Information Resources Management Association)
Description: Hershey PA : Medical Information Science Reference, [2019] | Includes bibliographical references.
Identifiers: LCCN 2018057821| ISBN 9781522589037 (hardcover) | ISBN 9781522589044 (ebook)
Subjects: | MESH: Biotechnology
Classification: LCC R858 | NLM TP 248.2 | DDC 610.285--dc23 LC record available at https://lccn.loc.gov/2018057821

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

The views expressed in this book are those of the authors, but not necessarily of the publisher.

For electronic access to this publication, please contact: eresources@igi-global.com.

Table of Contents

Section 2
Development and Design Methodologies

Volume II

Section 3
Tools and Technologies

Volume III

Section 4
Utilization and Applications

Section 5
Organizational and Social Implications

Volume IV

Section 6
Critical Issues and Challenges

Chapter 4
Bioinformatics Database Resources

Icxa Khandelwal
Jaypee University of Information Technology, India

Aditi Sharma
Jaypee University of Information Technology, India

Pavan Kumar Agrawal
G. B. Pant Engineering College, India

Rahul Shrivastava
Jaypee University of Information Technology, India

ABSTRACT

Various biological databases are available online, which are classified based on various criteria for ease of access and use. All such bioinformatics database resources have been discussed in brief in this book chapter. The major focus is on most commonly used biological/bioinformatics databases. The authors provide an overview of the information provided and analysis done by each database, information retrieval system and formats available, along with utility of the database to its users. Most widely used databases have been covered in detail so as to enhance readers' understanding. This chapter will serve as a guide to those who are new to the field of bioinformatics database resources, or wish to have consolidated information on various bioinformatics databases available.

INTRODUCTION

The National Center for Biotechnology Information (NCBI) defines bioinformatics as: "the field of science in which biology, computer science, and information technology merge into a single discipline". Bioinformatics can be considered an amalgam of three sub-disciplines:

1. Development of new algorithms as well as statistics so that the relationship between the elements of huge datasets can be determined.

DOI: 10.4018/978-1-5225-8903-7.ch004

2. Analysis as well as interpretation of biological data i.e. various types of sequences and structures.
3. Development of tools and software to ensure efficient access as well as management of biological data (Toomula, 2011).

The bioinformatics database resources focus primarily on the third sub-discipline of bioinformatics. A database can be defined as a computerized and organized storehouse of related information that provides a standardized way for searching, inserting and updating data. The data stored in these databases is persistent and organized. Database Management System (DBMS) is a software application that deals with the user, other applications, and the database itself in order to perform analysis and capture data in a systematic manner.

Bioinformatics databases or biological databases are storehouses of biological information. They can be defined as libraries containing data collected from scientific experiments, published literature and computational analysis. It provides users an interface to facilitate easy and efficient recording, storing, analyzing and retrieval of biological data through application of computer software. Biological data comes in several different formats like text, sequence data, structure, links, etc. and these needs to be taken into account while creating the databases.

There are various criteria on the basis of which the databases can be classified. On the basis of structure, databases can be classified as a text file, flat file, object-oriented and relational databases. On the basis of information, they can be classified as general and specialized databases. Most commonly, they are classified on the basis of the type of data stored in primary, secondary and composite databases (Kumar, 2005).

CLASSIFICATION OF DATABASES

Type 1

Databases can be classified on the basis of structure as Abstract Syntax Notation (ASN.1), Flat files, Object oriented databases, Relational databases, and XML. Table 1 provides a comparison of various types of databases on the basis of structure

- **ASN.1:** This format comprises of a syntax and description of how a particular data type can be represented physically in a data stream or sequential file (Buneman, Davidson, Hart, Overton, & Wong, 1995). This format has been adopted by NCBI for the representation of sequential data. It is one of the major file formats in GenBank (Cooray, 2012).
- **Flat Files:** This implementation is based on only one table, which incorporates the complete data i.e. all the attributes for each variable. Each row of the table specifies a different record. Specified delimiters are used to differentiate among records. Maintenance of data stored is a major drawback of this type of databases. Integration of two or more databases is difficult due to redundancy in data and variation in the format used.
- **Object Oriented Databases:** Object oriented databases can handle complex data types and can be easily integrated with Object Oriented Programming Languages (OOPL) (Codd, 1970). They can be defined as a collection of objects. Objects represent an instance of an entity and comprise attributes as well as methods (Hasegawa, 2008).

- **Relational Databases:** Relational database systems can be defined as a collection of relations or tables. In a relational database, the data is organized in the form of a table where each row contains a record and each column specific an attribute of the record. The ordering of tuples, attributes or values within a tuple do not make any impact on the relation. The data is subjected to various constraints for validation (Cooray, 2012).
- **XML:** XML can be defined as an advanced flat file format. It provides greater support for representation of complex nested data structures. It contains data definitions and supports new definitions and tags upon requirement. The major advantages of this type are fast accessibility, reliability, and scalability. (Cooray, 2012)

Type 2

The databases can be classified into three categories on the basis of the information stored. They are Primary, Secondary and Composite databases.

- **Primary Databases:** Primary databases contain data that is derived experimentally. They usually store information related to the sequences or structures of biological components(Singh, Gupta, Nischal, Khattri, & Nath, 2010)(IASRI, (N.D.)). They can be further divided into protein or nucleotide databases which can be further divided as sequence or structure databases. The most commonly used primary databases are: DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database, GenBank, and Protein Data Bank (PDB) (Toomula, 2011).
- **Secondary Databases:** Secondary databases contain the data that is obtained through the analysis or treatment of data present in primary databases. For instance, it can contain conserved protein sequence, signature sequence active site residues of protein families which are obtained from multiple sequence alignment of related proteins, etc. (Varsale, Wadnerkar, Mandage, & Jadhavrao, 2010; Sahoo, Rani, Dikhit, Ansari, & Das, 2009). These databases can be further classified as metabolic pathways database, protein family database, etc. The most common examples are Class Architecture Topology Homology (CATH), Kyoto Encyclopedia of Genes and Genomics (KEGG), Protein Families (Pfam) and Structural Classification of Proteins (SCOP).
- **Composite Databases*:*** Composite databases are collections of several (usually more than two) primary database resources. This helps in the lessening the tedious task of searching through mul-

Table 1. Comparison between different databases on the basis of structure

Type	Merits	Demerits	Examples
ASN.1	Implementation ease, Standardized	Not easy to integrate	CDD, GenBank, OMIM
Flat File	User-friendly	Not easy to access, integrate and validate	EMBL, DDBJ
Object Oriented	Implementation ease, Supports abstract data types	Factorization of document, Integration enhancement	MITOMAP
Relational	Implementation ease, Reliable, Scalable	Reduced performance when a large number of join operations are used	GDB, SMART
XML	Fast response, Flexible, Better accessibility	Less mature as compared to DBMS	GO, SwissProt

Cooray, 2012.

tiple databases referring to the same data. The approach used, for instance, the search algorithm employed, differs considerably in every composite database. For example DrugBank offers details on drug and their targets, BioGraph incorporates assorted knowledge of biomedical science and Bio Model is a storehouse of computational models of the biological developments, etc. There are many composite databases which provide users with various tools and software for analysis of data. NCBI being a composite database has stored a lot of sequence of nucleotide and protein within its server and thereby suffers from high redundancy in the data deposited (IASRI, (N.D.).

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI)

It was developed at the National Institutes of Health (NIH) in 1988 for advancement (Wheeler et al., 2007) of science as well as health as it provides access to a large amount of biomedical and genomic information (www.ncbi.nlm.nih.gov/home/about/mission.shtml). It maintains a large scale of databases and bioinformatics tools as well as services. One of the most popular databases is GenBank. It is a nucleic acid sequence database and its data is acknowledged by the scientists all around the world. Another popular and major database is PubMed, a bibliographic database for the biomedical literature. All the databases are available online at official website of NCBI through the Entrez search engine (Wheeler et al., 2007).

Mission

The aim is to find novel techniques and methodologies for dealing with huge and complex data and provide better accessibility to analytical and computational tools.

Organization

The various branches of NCBI are Computational Biology Branch (CBB), Information Engineering Branch (IEB) and Information Resource Branch (IRB).

Options on Homepage

There are various options can be viewed and explored on the homepage of NCBI's website. They are mentioned in Table 2.

Resources

The resources that are present on this site can be divided into two major categories: databases and tools, which are then further divided as follows:

1. **Databases:**
    a. **General:**
        i. Entrez,
        ii. PubMed and PubMed Central,
        iii. Taxonomy, and

Table 2. Various options along with their descriptions present on NCBI's homepage

Option	Description
Submit	Deposition of data
Download	Downloading data from NCBI
Learn	Users can learn about various tools and databases through documents or tutorials
Develop	New applications can be built by using Application Programming Interfaces and code libraries of NCBI
Analyse	Choose an appropriate NCBI tool for a specific data analysis task
Research	Research and collaborative projects of NCBI can be explored

    iv. Protein.
  b. **Gene Level Sequences:**
    i. Gene,
    ii. GenBank,
    iii. Unigene,
    iv. Homologene, and
    v. Reference Sequences.
  c. **Genomic Analysis:**
    i. Entrez Genome.
  d. **Analysis of Gene Expression:**
    i. Gene Expression Omnibus.
  e. **Analysis of Phenotypes:**
    i. Online Mendelian Inheritance in Man, and
    ii. Online Mendelian Inheritance in Animals.
  f. **Molecular Structure and Proteomics:**
    i. Structure Databases,
    ii. Molecular Modeling Database, and
    iii. PubChem.
  g. Hiv-1/Human Protein Interaction Database.

2. **Tools:**
  a. BLAST.
  b. **Gene-Level Analysis:**
    i. Open Reading Frame Finder.
  c. **Genomic Analysis:**
    i. Map Viewer, and
    ii. Model Maker,
    iii. Evidence Viewer.
  d. **Analysis of Gene Expression:**
    i. Genset, and
    ii. Probe.
  e. **Tools Supporting Proteomics Blast Link (Blink):**
    i. Open Mass Spectrometry Search Algorithm (Wheeler et al., 2007).

Entrez Global Query Cross-Database Search System

It is an integrated database search and retrieval system which is widely used (Maglott D., 2005) as it enables text searching using Boolean expressions (Schäffer et al., 2001). The data is integrated from a wide variety of sources and databases to create a uniform information model. Hence it is used for both indexing as well as retrieval (H., 2014). It aids in the availability of extensive links within and between database records (Wheeler et al., 2007). It allows users the combined access to sequential, structural and taxonomic data. Graphical representation of chromosome maps as well as the sequence is also provided. It can also aid the user in obtaining the sequences, structures or references that are related to the query entered. The users can also store their private configuration options using 'My NCBI' (Wheeler et al., 2007).

PubMed and PubMed Central

NCBI National Library of Medicine (NLM) created the PubMed database which is also the part of NCBI Entrez retrieval system. Providing users with an ease in accessing abstracts as well as references from biomedical and life sciences journals was the primary reason behind its creation. To add on, there are links provided for accessing the complete journal articles (Lindberg, 2000). The primary data source for PubMed is MEDLINE database (C. K. J. J. C. M. J., 2000). PubMed Central (PMC) enables users to access freely all the articles it contains (B. J. S. E., 2002).

Taxonomy

Taxonomy database can be accessed using the Taxonomy Browser for either viewing the taxonomic position or retrieving data depending upon the requirement of the user (Wheeler et al., 2007).

Protein

This database stores individual protein sequences in a textual format including FASTA and XML. The most common sources from which these sequences have been obtained are GenBank, NCBI Reference Sequence (RefSeq) project, PDB and SWISS-Prot/UniProtKB. Sets of similar and identical proteins determined by BLAST is also provided for each sequence (S. E., 2013).

GenBank

It is located in the USA. NCBI since 1992 has provided access to GenBank DNA sequence database through NCBI gateway server and hence is accessible freely(Schuler, Epstein, Ohkawa, & Kans, 1996). The three nucleotide sequence databases GenBank, European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ) coordinate among themselves so that all three of them are updated with the latest findings (Pruitt, Tatusova, & Maglott, 2005).

A detailed structure of a nucleotide sequence file format in this database includes the following:

1. **Locus:** This can be defined as a title given by GenBank itself to name the sequence entry. It includes the following:

   a. **Locus Name:** Similar to accession number for the sequence.
   b. **Sequence Length:** Tells the number of bases existing in the sequence.
   c. **Molecule-Type:** Identifies the type of nucleic acid sequence. The various types are mRNA (which is present as cDNA), rRNA, snRNA, and DNA.
   d. **GB Division:** Postulates class of the data according to classification criteria of GenBank.
   e. **Modification Date:** The date on which the record was modified.
2. **Definition:** This denotes the name of the nucleotide sequence.
3. **Accession:** This covers accession number, accession version, and GI number. Accession number can be defined as the unique identifier associated with each nucleotide sequence present in the database. If more than one record is created for a particular sequence then it will have the same accession number but all records will have different versions associated with that accession number.
4. **Keyword:** Defined words that were used to index the entries.
5. **The Source:** This describes organism from which sequences have been obtained. The accepted common name is mentioned first and then the scientific name is mentioned. In the end, the taxonomic lineage according to GenBank is specified.
6. **The Citation:** Includes the journal from which with the sequence was derived as initially the sequences were obtained only from published literature.
7. **Features:** These consist of the information derived from the sequence such as biological source, coding region, exon, intron, promoters, alternate splice patterns, mutations, etc.
8. **Sequence:** Contains the following:
   a. Count of presence of each nucleotide in the sequence,
   b. Whole nucleotide sequence,
   c. Beginning of sequence is determined by keyword “ORIGIN”, and
   d. End is marked as “\\”.

There are many techniques for retrieving and searching data from GenBank. The sequence identifiers can be searched in GenBank along with Entrez Nucleotide. Another approach is using BLAST search and then aligning nucleotide sequences to the query sequence. The last method is to search the appropriate link and then download nucleotide sequences. It intends to offer and reassure access of the most updated data of the nucleotide (www.ncbi.nlm.nih.gov/genbank/).

In order to maintain the confidentiality, GenBank on request, reserves announcement of new submissions for a definite interval of time. If sequences of the human genome are deposited to GenBank, it is mandatory not to include any personal information that can anyhow lead to the revelation of the identity of the individual.

Gene

It is involved in the characterization and organization of gene information about genes. Each gene record can be identified through a unique GeneID (Maglott D., 2005). Organism-specific XML files can be created by applying the organism filter (Maglott, Ostell, Pruitt, & Tatusova, 2007).

UniGene

UniGene can be defined as a software that partitions the sequences present in GenBank into sets of non-redundant gene- oriented clusters. These collections have been employed in the creation of unique sequences for microarray fabrication in order to comprehend gene expression study on a large scale (Schuler, 1997).

Entrez Genome

Entrez Genome (Tatusova, Karsch-Mizrachi, & Ostell, 1999) provides access to a large number of complete microbial genomic sequences and an enormous number of viral genomic sequences. There are also a huge amount of reference sequences for eukaryotic organelles. It is supplemented by Entrez Genome Project database. This database provides the status of various on-going annotation, assembly and sequencing projects (Wheeler et al., 2007).

Gene Expression Omnibus (GEO)

It can be defined as a data repository as well as retrieval system for various types of high-throughput molecular abundance data. It accepts array comparative genomic hybridization (aCGH) data, chromatin immune- precipitation on array (ChIP-chip) data, gene expression data and SNP array data (Barrett et al., 2007).

Online Mendelian Inheritance in Man (OMIM)

It contains a catalog of human genes as well as genetic disorders (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005). The data comprises of disease phenotypes, gene polymorphism, genes, map locations and patterns of inheritance (Wheeler et al., 2007).

Online Mendelian Inheritance in Animals (OMIA)

It is a database that contains the genes, inherited disorders as well as traits in fauna species excluding human and mouse. It also has the links to the records that are relevant in OMIM, PubMed and Gene databases (Wheeler et al., 2007).

PubChem

PubChem can be considered as a molecular library containing relational databases which were created using Microsoft SQL servers. The major focus is on the biological, chemical and structural properties of small molecules so that they can be used as diagnostic and therapeutic agents. All of the deposited data can be accessed freely by the users. It comprises of three sub-databases which are: PCSubstance, PCCompound, and PCBioAssay which contain substance information, compound structures and bioactivity data of compounds respectively (Wheeler et al., 2007).

Basic Alignment Local Search Tool (BLAST)

NCBI developed BLAST, a powerful tool for comparing sequences from various organisms (T., 2002). It can be defined as an algorithm to determine the similarity between biological sequences (Altschul, Gish, Miller, Myers, & Lipman, 1990). Gapped alignments having links to the database are provided in the final result (Wheeler et al., 2007). It has been reported that this tool has the capacity to search entire DNA database in less than 15 seconds. (ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf). The input is in the FASTA or Genbank format and output can be displayed in various formats like HTML, XML formatting and plain text (T., 2002). The score of each alignment is assigned an Expectation Value (E-value) which is a measure of statistical significance (Wheeler et al., 2007).

Open Reading Frame Finder

Six-frame translation of nucleotide sequence is performed by this tool. The result is location of each ORF within a specified size range (Wheeler et al., 2007).

EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL)

It is a molecule-based biology research foundation, maintained by 21 member states and formed in the year 1974 (www.embl.fr/aboutus/general_information/organisation/member_states/index.html). It stores and makes available raw nucleotide sequences. It is situated in UK (IASRI, (N.D.)). European Bioinformatics Institute (EBI) maintains EMBL nucleotide sequence database (Garg, Pundhir, Prakash, & Kumar, 2008).

Mission of EBI

The various aims of the organization are as follows:

- To provide freely available data and bioinformatics services to all facets of the scientific community.
- To contribute to the advancement of biology through basic investigator-driven research.
- To provide advanced bioinformatics training to scientists at all levels.
- To help disseminate cutting-edge technologies to industry.
- To coordinate biological data provision throughout Europe (www.ebi.ac.uk/).

European Nucleotide Archive (ENA)

Free as well as unrestricted information access on DNA and RNA sequences is provided by ENA. This archive is created using three databases which are Sequence Read Archive, Trace Archive and EMBL Nucleotide Sequence Database (www.ebi.ac.uk/ena). The information in ENA can be extracted manually or programmatically and resultant files can be obtained in various formats like XML, HTML, FASTA and FASTQ. (O. J., 2002) Using accession numbers and other specific text queries the users can obtain individual archives (Leinonen et al., 2010).

EMBL Nucleotide Sequence Database

It contains the high level genome assembly data of sequences and their functional annotation (Stoesser et al., 2003; Amid et al., 2011). The data is store in flat file format.

Data Classes

The different data classes of sequences are mentioned in Table 3.

Resources at EMBL-EBI

The EBI website provides link to access many services like various biological tools and databases. Some of the most common resources are listed below in the Appendix.

DNA DATA BANK OF JAPAN (DDBJ)

This biological database resource belongs to National Institute of Genetics (NIG) in Japan. DDBJ is the only nucleotide sequence data bank currently present in Asia. Although DDBJ essentially has Japanese researchers as contributors but it also accepts the data from researchers of other countries. It is an associate of the International Nucleotide Sequence Database Collaboration (INSDC). The major driving force behind DDBJ operations is the advancement of the quality of INSD as the nucleotide sequence accounts organism development more directly than other biological constituents.

Tasks

Key tasks of DDBJ Center are as follows:

1. Construction and operation of INSDC which offers nucleotide and amino acid sequence data along with the patent request.
2. Provides searching and analysis of biological data.
3. Training course and journal.

DDBJ Flat File Format

The data submitted in DDBJ is managed and retrieved according to the DDBJ format (flat file). The flat file includes the sequence and the information of who submitted the data, references, source organisms, and information about the feature, etc (www.ddbj.nig.ac.jp/ddbjingtop-e.html).

Table 3. Summary of data classes

Data Class	Definition	Example
EST	Raw expressed sequence tags without sequence quality information	FASTA Flat file HTML XML
WGS	Genomic contigs	FASTA Flat file HTML XML
GSS	Genome survey sequence; single pass, single direction sequence	FASTA Flat file HTML XML
HTC	High throughput assembled transcriptomic sequence and optional annotation	FASTA Flat file HTML XML
HTG	High throughput assembled genomic sequence and optional annotation	FASTA Flat file HTML XML
STD	Assembled and annotated sequences	FASTA Flat file HTML XML
CON	Scaffolds build from genomic or transcriptomic contigs	FASTA Flat file HTML XML
STS	Sequence tagged site	FASTA Flat file HTML XML
PAT	Patent sequences	FASTA Flat file HTML XML
TSA	Transcriptomic contigs	FASTA Flat file HTML XML
CDS	Coding sequences	FASTA Flat file HTML XML

www.ebi.ac.uk/ena/submit/sequence-format.

PROTEIN DATA BANK (PDB)

PDB is a universal free archive for structural data of biological macromolecules. It was established in 1971 at Brookhaven National Laboratory under the governance of Walter Hamilton and initially contained only 7 protein structures. There were two major reasons which initiated the formation of PDB. The first one was increasing assembly of datasets related to protein structure. Another reason was the availability of Brookhaven Raster Display (BRAD) that envisages the protein structures in 3-D (Meyer, 1997). It is now conserved by the Research Collaboratory for Structural Bioinformatics (RCSB). It contains three-dimensional structures of proteins, nucleic acid fragments, RNA molecules, large peptides and complex structures of proteins and nucleic acids (Berman et al., 2000).

Data Storage and Acquisition

The data for each structure is stored in a distinct file and hence the data is stored in flat file arrangement. The major source for the three-dimensional structure of proteins includes cryo-electron microscopy, molecular modeling, NMR experimentations and X-ray crystallography trials (IASRI, (N.D.)). Each structure present within PDB is given four character alphanumeric characters. The protein structure files may be viewed using the open resource. The RCSB PDB website covers an extensive list of both free and marketable molecule conception programs that includes Jmol, Pymol, and Rasmol and web browser plugins (Protein_Data_Bank, N.D.).

Atomic Coordinate Entry Format Description

The various sections of the PDB file are:

1. Title Section,
2. Primary Structure Section,
3. Heterogen Section,
4. Secondary Structure Section,
5. Connectivity Annotation Section,
6. Miscellaneous Features Section,
7. Crystallographic and Coordinate Transformation Section,
8. Coordinate Section,
9. Connectivity Section, and
10. Bookkeeping Section.

Title Section

It contains the elements which describe the experiment and biological macromolecules present in the record entered. The elements of this section along with their description are mentioned below:

- **HEADER:** It contains an idCode field which is used for unique identification, a classification, and date when the coordinates were deposited to the PDB archive for a PDB record.

Coordinate Section

It stores a collection of atomic coordinates as well as the MODEL and ENDMDL records. The elements of this section along with their description are mentioned below:

- **MODEL:** It provides a specification of the model serial number.
- **ATOM:** It provides atomic coordinates for standard amino acids and nucleotides.
- **ANISOU:** It has anisotropic temperature factors.
- **TER:** It determines the end of a list of ATOM/HETATM records for a chain.
- **HETATM:** It is used for the representation of non-polymer or other "non-standard" chemical coordinates.
- **ENDMDL:** It is paired with a corresponding MODEL element to generate individual structures (Berman et al., 2003).

Connectivity Section

It contains information on atomic connectivity. Element of this section along with its description is mentioned below:

- **CONECT:** It determines connectivity between atoms for which coordinates have been supplied (Berman et al., 2003).

Bookkeeping Section

It has final information about the file itself. Element of this section along with its description is mentioned below:

- **MASTER:** It contains the number of lines in the coordinate file for selected record types.
- **END:** It denotes the end of the PDB file (Berman et al., 2003).

UNIPROT

It is a database of freely accessible protein sequences which contains high-quality data and functional information for the proteins. Many of the records have been obtained from genome sequencing projects. The information regarding the biological function of the protein has been extracted from the research literature. European Bioinformatics Institute (EBI), Swiss Institute of Bioinformatics (SIB) and Protein Information Resource (PIR) constitute the UniProt consortium. Each one of them is deeply engaged in protein database maintenance and annotation (Consortium, 2011)(Eck & Dayhoff, 1966). It includes four core databases: UniProtKB, UniParc, UniRef, and UniMes. It is funded by grants from European Commission, National Human Genome Research Institute, NIH, NCI-caBIG, the Department of Defense and Swiss Federal Government (Consortium, 2010)

UniProtKB

UniProt Knowledgebase (UniProtKB) is a protein database that is partially curated by experts. It includes three databases: Swiss-Prot, TrEMBL, and PIR-PSD. The former one contains reviewed and manually annotated records whereas the latter one comprises the un-reviewed and automatically annotated entries (Consortium, 2010).

Swiss-Prot

It can be defined as a manually curated protein sequence database with high annotation level. It was created by Amos Bairoch in 1986, developed by the Swiss Institute of Bioinformatics and subsequently further enhanced by Rolf Apweiler at EBI (Bairoch 2000; Bairoch & Apweiler, 1996; Altairac, 2006). It is well known for its high annotation level, a low degree of redundancy, standardized nomenclature usage, and links to specialized databases (O'Donovan et al., 2002).

To provide a set of relevant information for a particular protein is the core aim of this database. It also aggregates data obtained from scientific literature and bio curator-evaluated computational analysis. To remain up-to-date the annotations are often reviewed periodically (Apweiler, Bairoch, Wu, et al., 2004). It has Sequence Retrieval System (SRS) that helps in searching through relevant databases like for Translated EMBL (TrEMBL) on the same site.

Each data in SwissProt that belongs to a protein sequence is considered to have a separate core data and annotation. Former includes protein sequences, related references, bibliography and taxonomy of the organism from where the sequence has been extracted. Latter includes function(s) performed by the protein, post-translational modification, functional sites, structural domain sites, secondary structure features, likenesses to other proteins and diseases that may be caused due to a mutation in diverse strains (IASRI, (N.D.)).

Table 4 describes the codes which are used in a record of Swiss-Prot.

TrEMBL

Since the sequence data were being generated at a very high rate with respect to the ability of the SwissProt in performing the annotation hence TrEMBL Nucleotide Sequence Data Library was created so as to facilitate computer-annotated data for those proteins which could not be entered in Swiss-Prot (Apweiler, Bairoch, & Wu, 2004). The records present in this database are automatically annotated and have been analyzed computationally with high quality. Automatic processing and insertion of the translation of annotated coding sequences present in the three major nucleotide sequence database are done through this database. It also takes into account the sequences from PDB, Ensembl, RefSeq and CCDS (Consortium, 2011).

PIR: International Protein Sequence Database (PIR - PSD)

PIR was developed initially at National Biomedical Research Foundation (NBRF) in the year 1984. PIR, Munich Information Centre for Protein Sequences (MIPS) and Japan International Protein Information Database (JIPID) collaborated together to form PIR-PSD(IASRI, (N.D.)) which can be defined as an integrated public bioinformatics resource to support genomic and proteomic research. It also helps in the

Table 4. Codes along with the full-name and description as used in Swiss-Prot

Code	Full- Name	Description
ID	Identification	It is a unique identifier that related with each entry and it appears in the beginning of a record.
AC	Accession Numbers	The name of data record might change but the AC cannot be changed. Also, if there are more than one accession numbers it suggests that the record was fabricated by integration with other records.
DT	Date	It comprises date consistent with the data entry formation, alteration date of sequence and annotation respectively.
DE	Description	It contains the general details of the sequence.
GN	Gene name	It contains the name of the encoding gene(s).
OS, OG, OC	Organism name, organelle, organism classification	These have the details of name and taxonomy of the organism.
RN, RP, RX, RA, RT, RL	Reference number, Position, comments, cross-reference, authors, title, and location	These have the bibliographic information.
CC	Comments	It has the free text comments.
DR	Database cross reference	It provides cross-reference to other databases.
KW	Keywords	It contains a list of keywords that can be used in indexes.
FT	Features tables	It defines regions or sites of concern in the sequence.
SQ	Sequence headers	It directs beginning of sequence statistics and gives a short summary of its contents.

IASRI, (N.D.).

identification and interpretation of protein sequences. Since 2002, it has become a part of UniProtKB (Wu & Nebert, 2004). This database contains non – redundant data, is annotated by experts, is comprehensive in nature and uses object-oriented DBMS (Database Management System). Classification of the sequences of protein on the basis of super-family is the unique characteristic of this database. The classification criteria also takes into account the homology domain as well as sequence motifs (IASRI, (N.D.)).

The curation status of PIRSF Database can be categorized as uncurated, preliminary and full/Full (with description). PIRSF Membership can be classified as follows: full (F), associate (A) and seed (S). Full is used for proteins which share complete sequence similarity as well as common domain architecture. Associate is used for those members whose sizes are greater than the family length range. Seed is used when it is required to create family specific full-length and domain HMMs using already present full members. PIRSF Family Level are divided into the Homeomorphic family (HFam), Subfamily (SubFam) and Superfamily (SuperFam).

The members which are categorized as HFam family are homologous as well as homeomorphic. The category of subfamily delineates protein clusters within a homeomorphic family which have specialized functions and/or variable domain architecture. Superfamily level brings together a number of distantly related families and orphan proteins that share one or more domains (http://pir.georgetown.edu/pirwww/support/help.shtml).

There are various rules for writing a protein pattern and they are as follows:

1. Usually, capital letters are used for denoting amino acid residues and "-" is inserted in between two amino acids.
2. To provide multiple amino acids as a choice in a particular position "[…]" can be used.
3. To exclude a set of amino acids "{…}" is used.
4. If a particular position has "x" then it means that any amino acid can be inserted there.
5. If an amino acid occurs more than once consecutively then this can be denoted using "(n)", where n is a number of times that amino acid has occurred.
6. On similar grounds, as the previous rule, "(n1,n2)" are employed for multiple or variable positions.
7. To match a pattern at N or C terminus, the symbol ">" is required at the beginning or end respectively.

UniParc

UniProt Archive (UniParc) stores proteins sequences from publicly available protein sequence database in a non-redundant manner (Apweiler, Bairoch, Wu, et al., 2004) and it is updated on a regular basis. Since proteins may exist in several databases and there are high chances that a single sequence is present multiple times in the same database. Hence to avoid redundancy of data, each unique sequence is presented only once in this database. The identical sequences are merged even if they belong to different species. A unique identifier called UPI is given to each sequence which enables the identification of a same protein from various source databases. The protein sequences present in this database are without any annotation. Database cross-references are provided in order to facilitate the retrieval of more detailed information from the source databases.

UniRef

Clustered sets of protein sequences from UniProtKB and selected UniParc records are comprised in UniProt Reference Clusters (UniRef).(Suzek, Huang, McGarvey, Mazumder, & Wu, 2007) The UniRef100 database is involved in combining the identical sequences and sequence fragments (from any species) into a single UniRef entry. The accession numbers of all the merged records, sequence of a representative protein and links to the corresponding UniProtKB and UniParc records are mentioned. (Consortium, 2010)

UniMES

UniProt Metagenomic and Environmental Sequences (UniMES) database has been created for environmental and metagenomic data (Consortium, 2010). In order to improve the original data through more analysis, proteins that have already been predicted are merged using InterPro. UniMES is the source containing data from Global Ocean Sampling Expedition (GOS) (Yooseph et al., 2007). UniProt Reference Clusters or UniProt Knowledgebase do not contain data of environmental sample of this database.

STRUCTURAL CLASSIFICATION OF PROTEINS (SCOP)

SCOP database was started by Alexey Murzin in the year 1994 at Centre of Protein Engineering and was later on established at Laboratory of Molecular Biology (Chandonia et al., 2007) in Cambridge University, England (Hubbard, Ailey, Brenner, Murzin, & Chothia, 1999). It is an open source database and hence is accessible freely (Subramanian, Muthurajan, & Ayyanar, 2008). The structural domains of proteins were manually classified using the criteria of likeliness of sequences as well as structures in order to create this database (Conte et al., 2000; Andreeva et al., 2004). Its foremost function is to categorize 3D structures of proteins in a hierarchical pattern of structural levels which include family, super-family, fold and class so as to determine the evolutionary association between proteins. (IASRI, (N.D.))

SCOP Levels

PDB contains freely available 3D protein structures which are used by SCOP for classification. The protein domain is the main element of classification of structure in SCOP (Hubbard et al., 1999). The various levels are described in Table 5.

The root of this hierarchical classification is class which has the following major types:

1. All α proteins class which contains domains comprising α-helices.
2. All β proteins class which contains domains comprising β-sheets.
3. α and β proteins class which contains individual units as β-α-β and β-sheets are usually parallel to each other.
4. α or β proteins class which contains separated α and β regions and β-sheets are arranged in an anti-parallel fashion.
5. Multi-domain proteins class which contains those folds comprise at-least two domains of different classes.
6. Membrane and cell surface proteins as well as peptides class which contains proteins excluding the ones present in the immune system.
7. Small proteins class which contains proteins that have a metal ligand, heme and/or disulfide bridges.

It can be inferred from the above table that families are more closely related than super-families. The criteria for placing various domains within a fold into the same family is that either they share a minimum of 30% sequence similarity or they share the same function with 15% sequence similarity. To support the classification of domains into super-families and families BLAST is used (Andreeva et al., 2004; Hubbard et al., 1999).

Table 5. Different levels of SCOP database

Level	Description
Class	Types of folds.
Fold	Various domain shapes contained in one single class.
Superfamily	Domains having at-least a distant common ancestor and belonging to a particular fold are clustered together.
Family	Domains sharing a more recent common ancestor in a super-family are grouped into families.
Protein Domain	Clustered domains in families.
Species	Gathering of protein domains in accordance with species.
Domain	Either an entire protein or a part of protein.

Hubbard et al., 1999.

Methodology of CATH

The methodology adopted by this database is as follows:

1. Separation of proteins into various domains
2. Created domains are then automatically sorted into classes.
3. Clustering is performed on the basis of sequence similarity upon the created classes.
4. Groups generated in step 3 constitute the H levels of the classification.
5. Topology level is created by structural comparison of H levels.
6. Architecture level is not assigned computationally.
7. At the end, Class Level classification is performed based on the following 4 criteria:
    a. Content of secondary structure,
    b. Contacts in secondary structure,
    c. Alteration scores of secondary structure, and
    d. Parallel strands percentage(Sillitoe et al., 2015).

Search Options Provided by CATH

The database can be searched by one of the following ways:

1. Entering a text, ID or keyword,
2. Uploading a protein sequence in FASTA format,
3. Uploading a PDB structure, and
4. Browsing the hierarchy (Sillitoe et al., 2015).

Data Files Provided by CATH

The various types of data files provided by CATH along with their description are mentioned in Table 8.

Using the Online Database

The classic investigation in this database begins by tracing a single PDB structure to its function and homologues. Let us assume that a user is interested in domains that can be found in a particular PDB chain. The user has the PDBID of the desired protein. The steps to be performed are as follows:

1. In the top right corner of home page there are links to key functions performed and the 'Quick Search' box. The user can enter the PDBID into this box and then click on Enter button.
2. The result page then obtained will have a list of all records that match with the query –term. Chain, Domains, Node and PDB are the return types obtained.
3. The domain in which user is interested will have a CATH Domain code which can be defined as a PDB Chain Identifier extension.
4. If on results page, user clicks on PDB record then structures and sequences for the chains corresponding to that record can be viewed. There is also a tabular representation of corresponding chains and domains.

Table 8. Type and description of data files provided by CATH

Type of File	Description of File
CathCathedral library	It is the library containing graphs of secondary structure of domain.
CathDomainDescriptionFile	It contains complete information of various domains in CATH.
CathDomainList	It contains a list of CATH domains which have been assigned already.
CathDomainPdb	It is a library containing PDB files which have been chopped for representative CATH domains.
CathDomall	It stores for each PDB Chain domain boundaries in “domall” format.
CathHmm	It contains HMM library file for CATH domains.
CathHmm (+unclassified)	It contains HMM library file for all CATH domains, including those which have not been classified till now.
CathNames	It contains a list of manually assigned names of CATH classification nodes.
CathUnclassifiedList	It contains a list of CATH domains that have not been classified till now.
Chains	It contains a list of PDB chains in CATH.
Domain Sequences (ATOM)	It is a FASTA sequence database created through ATOM records in PDB for all CATH domains.
Domain Sequences (COMBS)	It is a FASTA sequence database created through COMBS sequence data for all CATH domains.
Representative Domain Sequences (ATOM)	It is a clustered FASTA sequence database for CATH domains.
Representative Domain Sequences (COMBS)	It is a clustered FASTA sequence database for CATH domains.
Representatives	It has a list of CATH representative domains.

Sillitoe et al., 2015.

5. When the user goes to the pages corresponding to the query domain then it will be observe that a tab entitled ‘History’ has been inserted. This tab contains the actions taken by curators of this database for assignment of domains.
6. Domain recognition in structures is done using CATHEDRAL.

SSAP is a server provided by this database for pair-wise comparison of two structures (Sillitoe et al., 2015).

PFAM

A database of protein families, Pfam contains annotations as well as multiple sequence alignments generated using hidden Markov models (Finn et al., 2009; Finn et al., 2006; Bateman et al., 2004). There are 4 elements present in all the families or patterns. These are: annotation, seed alignment, HMM profile and full alignment of sequences. Using seed alignment, sequences are bootstrapped into multiple alignments and eventually family (IASRI, (N.D.)).

Features

The user can get the information about known protein structures, multiple alignments, protein domain architectures and species distribution for each family in Pfam. It has been reported that Pfam contains minimum one match for 80% of protein sequences present in UniProt Knowledgebase. Now, the Pfam consortium is involved in the coordination of the annotation of Pfam families via Wikipedia (Finn et al., 2009).

Types

The two sub-types of Pfam database are Pfam-A and Pfam-B. The former one is manually curated. It stores protein sequence alignment and hidden Markov model for each record. Since the records in Pfam-A are unable to consider all known proteins, Pfam-B was created to serve as an automatically generated supplement. A huge number of small protein families obtained from clusters are present in Pfam-B(Heger, Wilton, Sivakumar, & Holm, 2005). When no Pfam-A families are found then Pfam-B families were used earlier but it has been discontinued from release 28.0 (Finn et al., 2009). iPfam (Finn, Marshall, & Bateman, 2005) is a database that was built on domain description provided by Pfam. It determines whether the different proteins described together in PDB are so enough to potentially interact.

Pfam 28.0

This is the latest version of Pfam. This database creates the higher-level clustering of families which are related and these groupings are termed as clans. A clan can be defined as a collection of Pfam-A records that which are related to each other by sequence, structure or profile-HMM similarity. A family in Pfam is now usually retrieved to as a Pfam-A record. Each record includes seed alignment which is curated, profile HMMs and full alignment which has been automatically generated (Finn et al., 2009).

Classification of Pfam Records

Each record in Pfam can be classified into one of four ways which are: Family, Domain, Repeat and Motifs. Family can be defined as a collection of related protein regions. Domain is a structural unit of a protein that can evolve and function independently. Repeat is a term used for a short unit which becomes stable when there are multiple copies but in isolation it is unstable. Motif is a short unit which is present outside globular domains (Finn et al., 2009).

Pfam-A Family Page

This is a page through which a user is able to view Pfam annotation for a protein family. Also, domain architectures can be sighted where information regarding a particular family is developed; it supports alignments in several formats for the family, therefore can be downloaded. To add on, for each family information such as structural details, phylogenetic, the HMM logo and species distribution are made accessible.

Clan

A clan can be defined as a category of protein families that are triggered from the single evolutionary origin. The likeness in tertiary structure or common motif sequence of the protein indicates the evolutionary relationships. Clan alignment is the alignment of the seed alignments of all the families within a particular clan.

Criteria for Categorizing Families Into Clans

There are a several criteria used for classifying families into clans. The golden standard is the usage of structures for making the classification. Profile comparisons such as HHsearch are used in the absence of structures. Sequence that matches two HMMs in the identical region of protein sequence is considered. SCOOP is a method that takes into consideration the common matches when searched, which may thereby specify an association. Such type of information is employed in order to decide about relationship among families.

Site Organization

The major page for accessing information is the family page, as the name suggests it describes the Pfam family records. To navigate to the family pages the users can also enter the Pfam identifier or the accession number in the keyword search box. There are various tabs for specific information such as alignments, curation and models, domain organization or architectures, functional annotation, HMM logo, interactions, species distribution, structure and Trees (Finn et al., 2009).

Keyword Search

The search box in the page header of each page of Pfam website can be used for keyword search. This type of searching can be done by entering various types of queries. Some of the most common ways are: Gene Ontology IDs and terms, HEADER and TITLE fields from PDB entries, InterPro entry abstracts, sequence entry description and species fields in UniProt and text fields in Pfam entries (Finn et al., 2009).

Using Pfam

To determine the domain architecture of the protein of interest, the user needs to search that particular protein sequence against the Pfam library of HMMs. If the protein is not recognized by Pfam then the user shall paste the complete protein sequence in search page. The sequence will be searched against the HMMs present in the database and the matches will be displayed as the result. If there a large number of sequences to be searched then batch upload facility can be used. The user needs to upload a file containing all the sequences in the FASTA format. The results will be emailed back to the user usually 48 hours after submission. If there are a considerably very huge number of sequences then Pfam searches can be performed locally by the user through the use of the 'pfam_scan.pl' script. This technique requires the user to have additional data files from website, HMMER3 software and Pfam HMM libraries. Pfam Alyzer is a tool which can be used by the user to identify the proteins which contain a specific combination of domains and to specify particular species and the evolutionary distances allowed between domains (Finn et al., 2009).

Scores in Pfam

E-Values and Bit-Scores

HMMER3 calculates E-values (expectation values). It can be defined as a count of hits expected in order to have a score which is either equal to or better than the value by chance. E-value is considered good if it is much less than 1. Since E-values are dependent on the size of the database searched, hence a second system is used for retaining Pfam models. This in-house system is based on a bit score and hence does not depend on the size of the database searched for. Bit score gathering (GA) threshold is calculated for each Pfam family and thereby is set manually in such a manner that all sequences scoring either exact or overhead this threshold appear in full alignment.

Sequence vs. Domain Scores

HMMER3 calculates two kinds of scores known as sequence score and domain score. Sequence score provides the score for the complete sequence. Domain sequence provides the score for the domain(s) on that particular sequence. The sequence score can be defined as an aggregate score of a sequence which is aligned to HMM model. In a sequence, a single domain is said to be existing when scores are the same. Therefore, the result of these multiple illustrations of domain enhances confidence thus concluding that sequence belongs to a specific family of protein.

Table 9 provides a list of most commonly used terms in Pfam's website.

PROSITE

Prosite is a protein pattern database which was created in 1988 by Amos Bairoch and belongs to Swiss Institute of Bioinformatics. It includes the basic patterns which are found in incomplete protein sequences, for instance, the specific functional or structural domains. Generally, patterns are found using multiple sequence alignment and then further processed according to the database (IASRI, (N.D.)). Protein motifs as well as patterns in this database are determined as regular expressions. Each record holds two forms of data that include patterns and relative descriptive text. A line commencing with "PA" declares the expression. References, as well as association for all the protein sequences that comprise of the pattern, are also revealed. Documentation files contain the descriptive text which is connected with the accession number using the expression data (IASRI, (N.D.)).

The data entries describe the protein domains, families and functional sites. In addition, they also have the information about the amino acid patterns as well as profiles. After manual curation by a team of Swiss Institute of Bioinformatics, the data is incorporated into Swiss-Prot protein annotation. It is involved in recognizing possible specific functions of newly discovered proteins and analysis of known proteins for previously undetermined action. It also suggests various implements for protein sequence analysis and detection of motif present within the protein sequence. It is part of the ExPASy proteomics analysis servers as well (De Castro et al., 2006; Hulo et al., 2008).

drugs and other health-related constituents are kept in KEGG ENVIRON database. KEGG MEDICUS is a database in health information category and it takes into account the package information of all advertised drugs in Japan (Kanehisa, Goto, Furumichi, Tanabe, & Hirakawa, 2010).

Classification on the Basis of Systems Data Stored

In this category we have three sub-types which are as follows:

1. **KEGG PATHWAY:** Pathway database is one of the fundamental KEGG resources which stores computerized information on the interaction of molecular networks. It is also known as the wiring diagram database. The objective was to enable the interpretation of genome sequence data through this database. It contains KEGG pathway maps which have experimental information on different pathways present in cell or organism. Through a molecular network, the genes in genomic sequence can be linked to gene products in the pathway. It also determines the pathways that are most likely to be encoded in a genome. The pathway maps are classified into the following sections: cellular processes, environmental information processing, genetic information processing, metabolism and organismal systems. Cellular processes deal with death, growth and membrane functions of the cell. Environmental information processing considers transportation through membranes and different mechanisms of signal transduction. Genetic information processing takes into account the transcription, translation, replication and repair mechanisms. Metabolism segment has visually drawn global maps of metabolic pathways. Organismal systems include various systems in an organism like endocrine, immune and nervous system.
2. **KEGG MODULE:** It highlights the functional units present in a pathway map. For instance, it will store the sub-pathways which have been reported to be conserved among specific molecular complexes as well as organisms. The gene sets present in these modules can be linked with various specific metabolic capacities as well as phenotypic features.
3. **KEGG BRITE:** It can be defined as an ontology database that contains a categorized arrangement of various biological entities. It can have many dissimilar relationships which are in contrast with KEGG PATHWAY as it contains only molecular interactions and reactions.

KEGG Search

It accepts an entry identifier in order to save the resultant entry. The identifier may be in the form of a database-dependent prefix followed by a five-digit number. If DBGET mode is implemented, then DBGET search will be made against the entire KEGG database. MEDICUS is a default mode in the KEGG MEDICUS page. A keyword search also implemented in order to get the required information (www.genome.jp/kegg/).

CONCLUSION

The book chapter provides an overview of the most common databases on the information and analysis provided by each database, information retrieval system and formats available, along with utility of the database to its users. The diversity of databases makes it challenging to identify which database should be used to solve a particular problem because database nomenclature is not standardized and data formats are also varying. Hence databases struggle with data redundancy and data inconsistencies. Different strategies have been proposed to prioritize choice of a particular database on the basis of the purpose of usage.

REFERENCES

Altairac, S. (2006). Naissance d'une banque de données: Interview du prof. Amos Bairoch. *Protéines à la Une*.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi:10.1016/S0022-2836(05)80360-2 PMID:2231712

Amid, C., Birney, E., Bower, L., Cerdeño-Tárraga, A., Cheng, Y., Cleland, I., ... Hunter, C. (2011). Major submissions tool developments at the European Nucleotide Archive. *Nucleic Acids Research*, gkr946. PMID:22080548

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., & Murzin, A. G. (2004). SCOP database in 2004: Refinements integrate structure and sequence family data. *Nucleic Acids Research*, *32*(suppl 1), D226–D229. doi:10.1093/nar/gkh039 PMID:14681400

Apweiler, R., Bairoch, A., & Wu, C. H. (2004). Protein sequence databases. *Current Opinion in Chemical Biology*, *8*(1), 76–80. doi:10.1016/j.cbpa.2003.12.004 PMID:15036160

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., ... Magrane, M. (2004). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *32*(suppl 1), D115–D119. doi:10.1093/nar/gkh131 PMID:14681372

Bairoch, A., & Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Research*, *24*(1), 21–25. doi:10.1093/nar/24.1.21 PMID:8594581

Bairoch, A. M. (2000). Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics (Oxford, England)*, *16*(1), 48–64. doi:10.1093/bioinformatics/16.1.48 PMID:10812477

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., ... Edgar, R. (2007). NCBI GEO: Mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Research*, *35*(suppl 1), D760–D765. doi:10.1093/nar/gkl887 PMID:17099226

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., ... Sonnhammer, E. L. (2004). The Pfam protein families database. *Nucleic Acids Research*, *32*(suppl 1), D138–D141. doi:10.1093/nar/gkh121 PMID:14681378

Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, *10*(12), 980–980. doi:10.1038/nsb1203-980 PMID:14634627

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., ... Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*(1), 235–242. doi:10.1093/nar/28.1.235 PMID:10592235

Birney, E., & Clamp, M. (2004). Biological database design and implementation. *Briefings in Bioinformatics*, *5*(5), 31–38. doi:10.1093/bib/5.1.31 PMID:15153304

Buneman, P., Davidson, S. B., Hart, K., Overton, C., & Wong, L. (1995). *A data transformation system for biological data sources*. Academic Press.

Chandonia, J.-M., Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., & Murzin, A. G. (2007). *Data growth and its impact on the SCOP database: new developments*. Berkeley, CA: Ernest Orlando Lawrence Berkeley National Laboratory.

Cochrane, G., Akhtar, R., Aldebert, P., Althorpe, N., Baldwin, A., Bates, K., ... Browne, P. (2008). Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, *36*(suppl 1), D5–D12. doi:10.1093/nar/gkm1018 PMID:18039715

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, *13*(6), 377–387. doi:10.1145/362384.362685

Consortium, U. (2010). The universal protein resource (UniProt) in 2010. *Nucleic Acids Research*, *38*(suppl 1), D142–D148. doi:10.1093/nar/gkp846 PMID:19843607

Consortium, U. (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research*, *39*(suppl 1), D214–D219. doi:10.1093/nar/gkq1020 PMID:21051339

Conte, L. L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., & Chothia, C. (2000). SCOP: A structural classification of proteins database. *Nucleic Acids Research*, *28*(1), 257–259. doi:10.1093/nar/28.1.257 PMID:10592240

Conte, L. L., Brenner, S. E., Hubbard, T. J., Chothia, C., & Murzin, A. G. (2002). SCOP database in 2002: Refinements accommodate structural genomics. *Nucleic Acids Research*, *30*(1), 264–267. doi:10.1093/nar/30.1.264 PMID:11752311

Cooray, D. M. P. N. S. (2012). Molecular biological databases: Evolutionary history, data modeling, implementation and ethical background. *Sri Lanka. Journal of Biomedical Informatics*, *3*(1), 2–11. doi:10.4038ljbmi.v3i1.2489

Day, R., Beck, D. A., Armen, R. S., & Daggett, V. (2003). A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Science*, *12*(10), 2150–2160. doi:10.1110/ps.0306803 PMID:14500873

De Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., ... Hulo, N. (2006). ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research*, *34*(suppl 2), W362–W365. doi:10.1093/nar/gkl124 PMID:16845026

Finn, R. D., Marshall, M., & Bateman, A. (2005). iPfam: Visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics (Oxford, England)*, *21*(3), 410–412. doi:10.1093/bioinformatics/bti011 PMID:15353450

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., ... Durbin, R. (2006). Pfam: Clans, web tools and services. *Nucleic Acids Research*, *34*(suppl 1), D247–D251. doi:10.1093/nar/gkj149 PMID:16381856

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., ... Forslund, K. (2009). The Pfam protein families database. *Nucleic Acids Research*. PMID:19920124

Garg, N., Pundhir, S., Prakash, A., & Kumar, A. (2008). PCR primer design: DREB genes. *J Comput Sci Syst Biol, 1*, 21-40.

Goto, S., Nishioka, T., & Kanehisa, M. (1999). LIGAND database for enzymes, compounds and reactions. *Nucleic Acids Research*, *27*(1), 377–379. doi:10.1093/nar/27.1.377 PMID:9847234

Hadley, C., & Jones, D. T. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure (London, England)*, *7*(9), 1099–1112. doi:10.1016/S0969-2126(99)80177-4 PMID:10508779

Hamm, G. H., & Cameron, G. N. (1986). The EMBL data library. *Nucleic Acids Research*, *14*(1), 5–9. doi:10.1093/nar/14.1.5 PMID:3945550

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, *33*(suppl 1), D514–D517. doi:10.1093/nar/gki033 PMID:15608251

Hasegawa, H. (2008). *Genome Databases Current Implementation Practices*. Retrieved in October.

Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K. F., Ueda, N., Hamajima, M., ... Kanehisa, M. (2006). KEGG as a glycome informatics resource. *Glycobiology*, *16*(5), 63R–70R. doi:10.1093/glycob/cwj010 PMID:16014746

Heger, A., Wilton, C. A., Sivakumar, A., & Holm, L. (2005). ADDA: A domain database with global coverage of the protein universe. *Nucleic Acids Research*, *33*(suppl 1), D188–D191. doi:10.1093/nar/gki096 PMID:15608174

Hubbard, T. J., Ailey, B., Brenner, S. E., Murzin, A. G., & Chothia, C. (1999). SCOP: A Structural Classification of Proteins database. *Nucleic Acids Research*, *27*(1), 254–256. doi:10.1093/nar/27.1.254 PMID:9847194

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B. A., De Castro, E., ... Sigrist, C. J. (2008). The 20 years of PROSITE. *Nucleic Acids Research*, *36*(suppl 1), D245–D249. doi:10.1093/nar/gkm977 PMID:18003654

Kanehisa, M. (1997). A database for post-genome analysis. *Trends in Genetics*, *9*(13), 375–376. doi:10.1016/S0168-9525(97)01223-7 PMID:9287494

Kanehisa, M. (2013). Chemical and genomic evolution of enzyme-catalyzed reaction networks. *FEBS Letters*, *587*(17), 2731–2737. doi:10.1016/j.febslet.2013.06.026 PMID:23816707

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. doi:10.1093/nar/28.1.27 PMID:10592173

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, *38*(Databasesuppl 1), D355–D360. doi:10.1093/nar/gkp896 PMID:19880382

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Research*, *42*(D1), D199–D205. doi:10.1093/nar/gkt1076 PMID:24214961

Kumar, S. (2005). *Bioinformatics Web*. Retrieved November 2015, from http://www.bioinformaticsweb.net/data.html

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., ... Gibson, R. (2010). The European nucleotide archive. *Nucleic Acids Research*.

Lindberg, D. (2000). Internet access to the National Library of Medicine. *Effective Clinical Practice*, *3*(5), 256. PMID:11185333

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2007). Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Research*, *35*(suppl 1), D26–D31. doi:10.1093/nar/gkl993 PMID:17148475

Mcwilliam, H., Valentin, F., Goujon, M., Li, W., Narayanasamy, M., Martin, J., ... Lopez, R. (2009). Web services at the european bioinformatics institute-2009. *Nucleic Acids Research*, *37*(suppl 2), W6–W10. doi:10.1093/nar/gkp302 PMID:19435877

Meyer, E. E. (1997). The first years of the Protein Data Bank. *Protein Science*, *6*(7), 1591–1597. doi:10.1002/pro.5560060724 PMID:9232661

Morya, V., Dewaker, V., Mecarty, S., & Singh, R. (2010). In silico analysis of metabolic pathways for identification of putative drug targets for Staphylococcus aureus. *J Comput Sci Syst Biol, 3*(3), 62-69.

Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., & Kanehisa, M. (2013). Modular architecture of metabolic pathways revealed by conserved sequences of reactions. *Journal of Chemical Information and Modeling*, *53*(3), 613–622. doi:10.1021/ci3005379 PMID:23384306

ODonovan, C., Martin, M. J., Gattiker, A., Gasteiger, E., Bairoch, A., & Apweiler, R. (2002). High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Briefings in Bioinformatics*, *3*(3), 275–284. doi:10.1093/bib/3.3.275 PMID:12230036

Orengo, C. A., Michie, A., Jones, S., Jones, D. T., Swindells, M., & Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure (London, England)*, *5*(8), 1093–1109. doi:10.1016/S0969-2126(97)00260-8 PMID:9309224

Protein_Data_Bank. (n.d.). *Protein Data Bank Wiki*. Retrieved from https://en.wikipedia.org/wiki/Protein_Data_Bank

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*(suppl 1), D501–D504. doi:10.1093/nar/gki025 PMID:15608248

Rao, V., Das, S., & Umari, E. (2009). Glycomics Data Mining. *J Comput Sci Syst Biol*, *2*, 262–265.

APPENDIX

Table 10. A list of resources that can be accessed and/or are available through EMBL-EBI

Services	Description
ArrayExpress	It contains information on experimentations related to gene expression.
BioModels Database	It contains catalog of computational models which are related to life sciences.
Chemical Entities of Biological Interest (ChEBI)	It can be considered as a database as well as ontology of molecular entities.
Clustal Omega	It performs multiple sequence alignment of nucleotide or protein sequences.
Clustal Phylogeny	It generates a phylogenetic tree based on the result given by ClustalW2 program.
Complex Portal	It stores manually curated biological macromolecular complexes from several model organisms
Ensembl project	It contains the genome databanks for eukaryotes. This is a joint venture with Wellcome Trust Sanger Institute.
Enzyme Portal	It provides detailed information for various enzymes.
Europe PubMed Central	It is a database that provides free access to biological research literature.
European Nucleotide Archive (ENA)	It provides information related to nucleotide sequencing.
Experimental Factor Ontology (EFO)	For various biomedical records, it provides an ontology of experimental variables.
Expression Atlas	This database stores the information regarding expression of genes based on different conditions.
FASTA	It is a protein sequence similarity search tool.
FingerPRINTScan	It determines PRINTS for a protein query sequence.
Gene ontology	It is an ontology of gene functions along with their processes.
GeneWise	It compares a protein sequence with a genomic DNA sequence.
GGSEARCH	It is used for searching sequences that are homologous to the desired query.
HMMER	It is a tool for protein homology search that uses profile hidden Markov models (HMMs).
IntEnz	It is an integrated relational enzyme database.
InterPro	It is a database of classification of proteins.
InterProScan 5	It searches sequences against InterPro's analytic protein signatures.
Kalign	It is a tool for multiple sequence alignment.
LALIGN	It is a tool to identify internal duplications.
MAFFT	It is a tool for multiple sequence alignment.
MEROPS	It is a database of proteolytic enzymes.
MUSCLE	It is a tool for multiple sequence alignment.
MView	It is used to reformat a multiple sequence alignment or transform a sequence similarity search into a multiple sequence alignment.
NCBI BLAST	It is a tool for local similarity search.
Patent databases	It is a database of non-redundant patent sequences.
Pfam	It has been created to assign conserved protein domains and families.
Phobius	It is a tool used for the prediction of signal peptides and trans membrane topology from amino acid sequence of protein.
PICR	It provides mapping between protein identifier name spaces.
Pratt	It is a tool for discovering patterns in unaligned protein sequences.
PROSITE Scan	It is a tool for searching protein query sequence.
Protein Data Bank in Europe	It collects, distributes and organizes 3D structural data on biological macromolecular structures and their complexes.
Proteomics Identifications Database (PRIDE)	It can be defined as a repository of protein expression data which is determined with the help of mass spectrometry
UniProt (The Universal Protein Resource)	It contains the protein sequence and functional annotation data.

Retrieved from: http://www.ebi.ac.uk/services/all.