

Chapter 2

Biological databases and their application

Parva Kumar Sharma¹ and Inderjit Singh Yadav²

¹Indian Council of Agricultural Research—Indian Agricultural Statistics Research Institute, New Delhi, India, ²School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India

2.1 Introduction

Advancements in the field of Bioinformatics in the last couple of decades have made it possible to generate and analyze the enormous amount of biological data. These data could be either in the form of sequences such as nucleic acids (DNA and RNA) and proteins or as three-dimensional structures of nucleic acids and proteins (Ragunath, Venkatesan, & Rajmohan, 2009). Particularly, in the last decade the steep surge in the sequencing of a large number of nonmodel genomes through the next-generation sequencing (NGS) platform has led to a large accumulation of sequencing data. However, the race between the data generation and its analysis is not in concord. Data analysis is much slower than data generation owing to its requirement of specialized personal (data analyst). Hence, until the data wait for its analysis, it has to be stored in an organized manner along with the already analyzed data. This is essential so that later on the data are easily accessible for the research community.

Originally, the concept of the database has come from the field of computer science, where new advanced computer software is being developed for easy access and retrieval of data. These software, known as Database Management Systems (DBMS), help in organizing, searching, accessing, and analyzing the data by interacting with the end users, various applications, and the database itself. The DBMS also helps in establishing hidden connections among the database records based on their instructions. Based on the manner the data are stored, a data structure is established so that the data in the database can be easily searched and different records in the database can be combined in the final output. Hence, depending on the type of data structures, the DBMS can be divided into two classes: relational database management systems that eventually make relational databases and object-oriented database management systems that make object-oriented databases.

2.1.1 Relational databases

A relational database contains a set of tables that hold the data. As a normal table, each table in the database is made up of rows and columns and is known as a relation. Rows represent values of the field while column represents the field. An attribute is a common feature that is present in all the tables and the columns are indexed according to this. This attribute is used for cross-referencing and any query executed in a relational database selects linked data from different relational tables and combines and displays the data at a unified platform. Hence, information can be searched quickly from a relational database. Structured Query Language is the most common programming language, which is used to make a relational database. It can also be used for database querying. Tables in a relational database should be smartly created to avoid redundancy of the data.

2.1.2 Object-oriented databases

Relational databases have a major drawback in that the complex hierarchical relationships between the data are not described by the tables. To overcome this problem, object-oriented databases have been developed. These databases store data as objects. In object-oriented programming, an object is a unit that contains both data and a code that acts on the data. The structure of the database is such that there is a set of pointers that link the objects telling their relationships. When a user searches the database, the query navigates through the objects with the help of the pointers.

Object-oriented databases are created using languages like C++. Object-oriented databases provide more flexibility in terms of defining the hierarchical relationships between the data. This helps in defining a complex relationship more easily in terms of programming. However, even this type of database suffers from some drawbacks like lack of mathematical basis as in the case of the relational database. Hence, some current databases use both relational and object-oriented databases, jointly known as the object-relational database management system.

The core idea has been taken from the relational database model of computer science, where a relationship exists between the database entries. The entries include gene/protein sequences (Sharma, Vrat, Kumud, Thakur, & Rajinder, 2011), structural information, textual description, citations, and many more. All the above-mentioned entries can be showed as plain text, tables, figures, etc. It is handy to have cross-references between different databases so that one can make use of different databases at the same time and can be able to correlate different data analysis outcomes (Karthick Raja, Selvaraj, & Muthukumar, 2008; Liu, Liu, Liu, Ding, & Yang, 2009). In bioinformatics, a separate branch dedicates itself to the development and implementation of such tools, which enables efficient storage, access, and management of the different types of data (Bourne, 2005).

Definition wise “A biological database is an organized collection of data, usually assembled using software particularly designed to organize, query, update and retrieve the data.”

A biological database serves three main purposes.

2.1.2.1 *It stores the biological data in computer-readable form*

Computers have become an intrinsic part of biological data analysis. Every biological information nowadays is available on computers; hence, having biological data stored in computer-readable form is a prerequisite.

2.1.2.2 *The stored data should be accessed efficiently*

The database search algorithm should be designed in such a way that only the relevant information be displayed on search.

2.1.2.3 *Biological data should be available to the research community in a single place*

Collecting information from the published literature is difficult and time-consuming and hence it is very helpful if a researcher finds all the relevant information on a single page.

The history of biological databases started with the sequencing of Insulin in 1956. The first database was created shortly after that in the form of a book and was named “Atlas of Protein Sequences and Structures” (Dayhoff, 1965) containing all the protein sequences known at that time. Later in the mid-sixties Yeast tRNA was sequenced consisting of 77 bases, which led to the foundations for the nucleic acid sequence database. Also, at this point, the three-dimensional structures of the protein were being studied and the initially solved structures were collected in a today well-known “Protein Data Bank” (PDB) in 1972 (Berman et al., 2000). Initially, the protein databases were maintained by individual laboratories and hence there was no common platform where researchers from all over the world could submit or retrieve protein sequences. Hence, in 1986 SWISS-PROT came into existence that provided a unified database where one can submit and find all the protein sequences. Over time, the biological databases grew larger and larger and the modern databases also include advanced query options and integrated tools that can directly analyze the stored data (Hoskeri, Krishna, & Amruthavalli, 2010).

All the biological databases can be primarily classified into five major categories *viz.* primary, composite, secondary, structural, and specialized databases (Fig. 2.1). Raw information in the form of sequences and structures are stored in a primary database. Examples include GenBank & DNA Data Bank of Japan (DDBJ) for Genome sequences, Swiss-Prot and Protein Information Resources (PIR) for protein sequences, and the PDB for protein structures (Singh et al., 2010). Composite databases are those combining various databases at a single platform and hence save the users’ time in searching inside individual databases. Each composite database has a search algorithm. An example of a composite database is UniProt.

A secondary database contains the derived information from the primary database (Shanthi, Ramanathan, & Sethumadhavan, 2009). Derived information is information that is the outcome of the analysis of the data of the primary database. Some examples of the derived information are active site residues, conserved sequences, and signature sequences in the case of protein sequences (Varsale, Wadnerkar, Mandage, & Jadhavrao, 2010). A secondary database for PDB contains PDB data in an organized manner depending on the protein properties, for example, based on their secondary structure, such as alpha protein, beta proteins, turns, and helices (Dawson & Kawai, 2009; Vaseeharan & Valli, 2011). Some databases also store information based on domain knowledge. Other examples are PROSITE

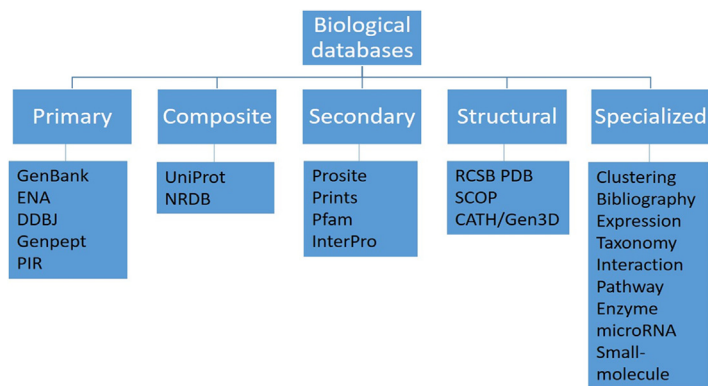


FIGURE 2.1 Biological database classification based on the data and major databases.

(Swiss Institute of Bioinformatics), Structural Classification of Proteins (SCOP) (Cambridge University), eMOTIF (Stanford University), and CATH (University College of London).

2.2 Sequence databases

2.2.1 GenBank

The GenBank (Benson et al., 2013) is the nucleotide sequence database managed by the National Institute of Health. The present version of GenBank (240, October 2020) contains 219,055,207 sequences. GenBank is an annotated collection of all publicly available DNA sequences. GenBank is a collaborative partner of the International Nucleotide Sequence Database Collaboration (INSDC), which is a collaboration of DDBJ (Ogasawara, Kodama, Mashima, Kosuge, & Fujisawa, 2019), the European Nucleotide Archive (ENA), and GenBank at National Center for Biotechnology Information (NCBI). Data are exchanged between these organizations daily. A GenBank release occurs every 2 months and is available from the FTP site.

Data can be submitted to the GeneBank either by Bankit or by tbl2asn, which is a command-line program that automates the creation of sequence records for submission to GenBank. It is used primarily for the submission of complete genomes and large batches of sequences. New sequences must have been published in any scientific journal before the submission of sequence data into the databases of INSDC. The GenBank database provides the most up-to-date and comprehensive DNA sequence information that helps the researchers to analyze either their data or they can download publicly available data and perform analysis of their choice and can contribute to knowledge discovery. Some important and most commonly used databases are listed in Table 2.1.

2.2.2 European Nucleotide Archive

The ENA is a publicly available database that provides annotated nucleotide (DNA and RNA) sequences (<https://www.ebi.ac.uk/ena/browser/home>). Apart from sequence information it also provides other subsidiary information, such as experimental procedures, sequence assembly details, and also the raw unassembled data. All this information is placed in different databases for easy accessibility. ENA is comprised of three databases *viz.* EMBL-Bank, Sequence Read Archive (SRA), and Trace Archive. EMBL-Bank contains the assembled data for sequences, where the submitted data are assembled by the submitter itself into one contig/scaffold. The submitter also provides the annotations of the assembled sequences, particularly for the genic regions.

The next is SRA, which is a platform for the submission of raw reads produced by the NGS platforms (Illumina, PacBio, Nanopore, etc.). Recently SRA also started accepting alignment information of raw reads on a reference genome. SRA is also a part of the INSDC. The third and last database is the Trace Archive. It contains raw reads that are unassembled and are generated through capillary sequencing technology. All the data in ENA can be submitted through the Webin portal submission form or EMBL-Bank flat files for complete genomes. The data in ENA are structured into different classes and taxonomic divisions. ENA is produced and maintained by the European Bioinformatics Institute and is a member of the INSDC.

TABLE 2.1 Major biological database along with their brief description and weblink.

Database name	Description	Weblink	References
NCBI	Provides a variety of databases and tools	http://www.ncbi.nlm.nih.gov/	NCBI
GenBank	Nucleotide sequences	http://www.ncbi.nlm.nih.gov/genbank/	Benson et al. (2013)
ENA-EMBL	Nucleotide sequences	https://www.ebi.ac.uk/ena/browser/home	ENA
DDBJ	Nucleotide sequences	http://www.ddbj.nig.ac.jp/	Ogasawara et al. (2019)
GenPept	Protein sequences	https://healthdata.gov/dataset/genpept	NCBI
PIR	Protein sequences	http://pir.georgetown.edu/	Wu et al. (2003)
UniProt	UniProt consortium maintains the UniProtKB, UniRef, and UniParc	https://www.uniprot.org/	UniProt Consortium (2018)
UniProtKB	Annotated protein data	https://www.uniprot.org/uniprot/	UniProt Consortium (2018)
UniRef	Clustered set of sequences from the UniProtKB	https://www.uniprot.org/uniref/	UniProt Consortium (2018)
UniParc	Publicly available protein sequences	https://www.uniprot.org/uniparc/	UniProt Consortium (2018)
NRDB	Protein sequences	https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/	NCBI
PROSITE	Protein families and domains	http://www.expasy.ch/prosite/	Hulo et al. (2008)
PRINTS	Protein fingerprints	http://130.88.97.239/PRINTS/index.php	Attwood et al. (2003)
Pfam	Protein domains	http://pfam.sanger.ac.uk/	Mistry et al. (2021)
InterPro	Integrated database for protein domains and families	http://www.ebi.ac.uk/interpro/	Mitchell et al. (2019)
RCSB PDB	3D protein structures	http://www.rcsb.org/pdb/home/home.do	Berman et al. (2000)
SCOP	Protein structural classification	http://scop.mrc-lmb.cam.ac.uk/	Andreeva, Kulesha, Gough, and Murzin (2020)
CATH	Classify proteins into evolutionary families	https://www.cathdb.info	Greene et al. (2007)

DDBJ, DNA Data Bank of Japan; *ENA*, European Nucleotide Archive; *NCBI*, National Center for Biotechnology Information; *NRDB*, nonredundant database; *PDB*, Protein Data Bank; *PIR*, Protein Information Resources; *RCSB*, Research Collaboratory for Structural Bioinformatics; *SCOP*, Structural Classification of Proteins; *UniParc*, UniProt Archive; *UniProtKB*, UniProt Knowledgebase; *UniRef*, UniProt Reference Cluster.

2.2.3 DNA Database of Japan

DDBJ (Ogasawara et al., 2019) is another primary nucleotide sequence database and is operated by the Center for Information Biology of Japan. It is also a member of the INSDC and hence at a given point of time has the same data as EMBL and GeneBank. Sharing data can enhance reproducibility and also allow for comparing various datasets. It also helps in a wide coverage of different types of end users.

2.2.4 GenPept

GenPept is the GenBank gene product data bank and is produced by the NCBI. This database contains protein sequences that are obtained from the translations of protein-coding sequences (CDS) present in the nucleotide sequence database (GeneBank, EMBL, and DDBJ). This database has no sequences, which are derived from any amino acid sequencing and also lack any type of annotation. This database can be helpful for those users, which are working on novel proteins, and any type of prior annotation is not available for that protein. So, the user can search this database and can find a homologous sequence.

2.2.5 Protein information resources

PIR (Wu et al., 2003) evolved from National Biomedical Research Foundation protein sequence database, which was originally developed by Margaret Dayhoff, who published it as “Atlas of protein sequence and structure.” This database

contains the protein name, its classification, and organism of origin. This database also has information on protein function and other general protein characteristics. Proteins that are of biological interest are also indicated in this database. This database has the following distinct features:

- The data are organized into protein family and protein superfamily.
- The database is updated regularly.
- Data are curated, nonredundant, and extensive.
- It includes data from all taxonomic classes, such as eukaryotes, prokaryotes, archaea, and viruses.

PIR database has subsidiary databases that provide other related information for proteins. One is Protein Ontology (PRO) providing the relationships between the proteins in terms of their ontological representations. Each PRO entry denotes a particular class of protein, such as any modified form, any orthologous isoform, or protein complexes. Another subsidiary is the PTMnet providing information on the protein Posttranslational Modifications (PTM) at a systems biology level. This database connects multiple resources, such as other databases, different bioinformatics tools, and tools for data mining, text mining into a resource that can look into the knowledge gaps and can help to find new PTM networks. Another important database is Protein Literature, Information and Knowledge (ProLINK), which is a literature database. This combines information through text mining from various sources and provides a unified picture of the information available for a protein query.

2.3 Composite database

2.3.1 Universal Protein Resource (UniProt)

UniProt was created in 2002 with a collaboration between the European Bioinformatics Institute (EMBL-EBI) (EMBL, Dubitzky, Wolkenhauer, Cho, & Yokota, 2013; UniProt Consortium, 2018), the Swiss Institute of Bioinformatics (SIB) and the PIR. UniProt is a large database that provides protein sequences and their respective annotations comprehensively.

The UniProt consortium formed three databases for different uses:

1. The UniProt Knowledgebase (UniProtKB);
2. The UniProt Reference Cluster (UniRef) databases; and
3. The UniProt Archive (UniParc).

2.3.1.1 UniProtKB

UniProtKB is the main database of UniProt. This database includes maximum annotation information available on proteins, such as their primary sequence, protein function description, their taxonomic classification, and also citation information. Other information if available can be added along with the entry. UniProtKB is divided into two components *viz.* Swiss-Prot and TrEMBL (Boeckmann et al., 2003). Swiss-Prot database is a manually curated and annotated protein sequence database, which is created by Amos Bairoch in 1986 during his Ph.D. work. Later on, it was developed and maintained by the SIB and the European Bioinformatics Institute. Swiss-Prot provides manually curated highly reliable protein sequences that are highly annotated, minimum redundancy, and integrated with other databases. It contains hundreds of thousands of proteins with maximum available information. Swiss-Prot is a distinct database with the following features: annotation, minimal redundancy, and integration with other databases.

In terms of annotation, two classes of data are available for a sequence in Swiss-Prot. The first one is the core data, including protein sequence, protein description, taxonomic data, and citation information. The second one is the annotation, including the protein function, any posttranslational modifications available, secondary structure, domains and sites, and the quaternary structure. Swiss-Prot has merged all the above data to minimize the redundancy in the data. Also, Swiss-Prot has integrated its data with other databases by providing cross-references with those databases like GeneBank/EMBL/DDBL in case of nucleotide sequences or with protein structure databases like PDB.

Due to high-throughput techniques, a large number of sequences are being entered into the database and hence the database specialist cannot annotate at the same pace. To overcome this issue, a Swiss-Prot supplement has been developed, which is Translated EMBL (TrEMBL database). It contains all the translated nucleic acids into proteins from nucleotide sequence databases that are yet to be included in Swiss-Prot. It also contains protein sequences extracted from the literature and protein sequences submitted directly by the user community. It is subdivided into two sections.

- (i) SP-TrEMBL: It contains a sequence that will be eventually incorporated into the Swiss-Prot.

- (ii) REM-TrEMBL: It contains those sequences that will not be incorporated into the Swiss-Prot. These include immunoglobulins and T-cell receptors, synthetic sequences, patent application sequences, and fragments of less than eight amino acids. Also, there is a weekly update to TrEMBL called TrEMBL new.

TrEMBL is automatically annotated and hence their quality is not as high as Swiss-Prot. Both Swiss-Prot and TrEMBL can be accessed via the same Swiss-Prot database.

2.3.1.2 UniRef

UniRef is a nonredundant sequence database for fast sequence similarity searches (Suzek, Huang, McGarvey, Mazumder, & Wu, 2007). UniRef has two main aims: the first one is to enable sequence merging in UniProt and the second one is to search similar sequences more rapidly and informatively (Apweiler et al., 2004).

UniRef has three sequence searching criteria, that is, searching sequences that are 100% similar, $\geq 90\%$, and $\geq 50\%$ similar, all of which are performed by UniRef100, UniRef90, and UniRef50, respectively.

2.3.1.3 UniParc

The UniParc is also a nonredundant database (NRDB) that contains almost all of the publicly available protein sequences in the world. It avoids redundancy from multiple sources by storing a unique sequence only once.

2.3.2 Nonredundant database

The NRDB is the NCBI protein database and is used for the BLAST searches. It contains unique (nonidentical) protein sequences that are collected from GenBank CDS translations, PDB, Swiss-Prot, PIR, and Protein Research Foundation. NRDB is comprehensive and is updated regularly but with the limitation that it is very large. The current release of the NR database (January 31, 2021) is more than 120 Gb in size.

2.4 Secondary database

2.4.1 PROSITE

Swiss Institute of Bioinformatics hosts an important secondary database named PROSITE (Hulo et al., 2008). This database describes proteins in terms of biologically significant sites and signature patterns. In PROSITE, based on these patterns, proteins are divided into families and domains. The idea behind this is that many proteins share the same families or domains based on their sequence similarity and hence can be grouped. It is believed that these grouped proteins have the same functions and share a common ancestor. It has been observed while studying the protein families that some sequences are more conserved in a family. These conserved regions are important in terms of protein functionality, for example, presence of active sites or maintaining the three-dimensional structure of the protein. By closely looking at these conserved regions it is possible to find a signature pattern that is unique to a particular family and can represent that family with the rest of the proteins. These signatures can reliably and rapidly help in the annotation of new protein sequences. This can also help in assigning a new protein sequence to a protein family or function. The database derives the patterns and motifs from the multiple sequence alignment of protein sequences which are saved as regular expressions which is a particular pattern of the sequences of amino acids. In Prosite, the amino acids in a regular expression are represented by the one-letter code and are separated by hyphens.

A typical example of PROSITE regular expression is [DERGHYI]-[PARVA]-[INDER]-{ANWE}-x(2)-P. Square brackets indicate that a position can contain more than one residues. A lowercase letter “x” means the particular position can be filled by any amino acid. Full brackets mean that the preceding amino acid or the pattern is repeating that number of times. Curly brackets indicate that the following amino acids are not present. ProRule is a section of PROSITE that describes manually created rules that can be used for automatic annotation of the proteins in the UniProtKB/SwissProt format based on the PROSITE motifs (mainly profiles). For database querying PROSITE server offers various options, such as keyword searching and sequence searching. Through ScanProsite, three databases, that is, Swissprot, TrEMBL, and PDB, can be searched for the presence of a user-defined pattern.

2.4.2 PRINTS

The PRINTS database (Attwood et al., 2003) uses the concept of fingerprints to classify sequences. Fingerprints are the outcome of multiple sequence alignment and consist of sequence motifs. In PRINTS, two types of fingerprints are

represented. The first one is simple fingerprints that are essentially single motif and the second one is composite fingerprints that are more complex and contain multiple motifs. In PRINTS, most of the fingerprints are the composite type because discrimination power is greater for multicomponent searches, and results are consequently easier to interpret.

The conceptual base is that proteins contain several sequence motifs that are functionally active and conserved in most of the proteins. Hence, using the concept of fingerprint the sensitivity of the sequence analysis increases as even in the absence of one motif one can affiliate a protein to a particular family of proteins. Through PRINTS, a researcher can also check for cross-references in the related database and hence can access some other important information. PRINTS database also contains information about each fingerprint and its biological function. Again, the database searching can be performed via keywords or sequences.

2.4.3 Pfam

The Pfam database (Mistry et al., 2021) classifies protein families based on domains. Definition wise, the domain is a functional region of a protein and a single protein can consist of one or more domains. Sequence alignments are at the base of Pfam. Hidden Markov models (HMM) are automatically constructed using high quality and manually tested alignments. More sequences kept on adding automatically in the alignment in the Swiss-Prot database. The new alignments should be interesting in terms of function and structures and should contain sequences related to evolution. However, the alignments are not fully manual and are partially automatic and due to that, no evolutionary relationship may be predicted with each other. Therefore the results of a search on the Pfam database should be checked carefully. Pfam also contains protein clans. Clans are groups of related entries based on sequence similarity, structure similarity, or profile HMM.

2.4.4 InterPro

InterPro (Mitchell et al., 2019) classifies proteins into families by predicting domains and important sites and hence provides functional analysis. InterPro combines key databases into a full database of signatures. InterPro merges many protein databases, such as TrEMBL, Pfam, Swissprot, PRINTS, PROSITE, ProDom, Smart, and TIGRFAMs (TIGR), and provides a simple but efficient query system for these databases. The results page pools the output of each database and provides it at a single page. This enables the researcher to look into results from every database, and it is possible to quickly compare the results taking into consideration of the pros and cons of the individual databases. The query page is quite intuitive and aids for an easy text and sequence search.

2.5 Structural databases

2.5.1 Research collaboratory for structural bioinformatics protein data bank

The PDB is a structure database that contains the three-dimensional crystal structure of macromolecules that are experimentally determined (Berman et al., 2000). These experimental methods are X-ray crystallography and NMR spectroscopy and nowadays cryo-electron microscopy is also used. Originally founded in 1971 at Brookhaven National Laboratory, it contains only seven entries. In 2020 the number of structures reached 171,916. These structures include mainly proteins, but nucleic acid structures, such as DNA and RNA and some protein–nucleic acid complexes, are also present.

A small number of other macromolecules, such as polysaccharides and glycopeptides, are also present. PDB only contains experimentally determined entries; however, for *in silico* models, a separate section has been made and called PDB models. Recently added, one can also search for drugs and ligands. Visualization tools are also present through which one can visualize a structure in three-dimensional in the web browser itself. PDB offers extension search options, such as searching PDB ID, Keywords, or direct sequences. One can also perform a BLAST search directly from PDB. A database record provides general information about the submission plus the coordinates of the three-dimensional crystal structure.

PDB is a very important database when it comes to the areas of structural biology. Structures in PDB have wide applications. They can be used for various studies including identification of new protein structures via *in silico* approaches or can be used for protein–nucleic acid interaction studies. Some databases, such as SCOP and CATH, use PDB data to classify proteins into classes or domains.

2.5.2 SCOP

The SCOP database clusters different proteins that performs a similar biological function and are evolutionarily related to a common structural organization. This common structural organization could be the full protein or only in the active center region. Therefore one can predict the function of a protein that is not known with respect to its structural structure by comparing it with that of the known proteins. SCOP provides such kind of forecasts. SCOP hierarchically classifies proteins of known structures (Andreeva et al., 2020).

It classifies protein into three major groups that are families, superfamilies, and folds. Families define proteins that have a clear evolutionary relationship with each other and are limited by a sequence identity constraint of at least 30% of the total length of proteins. However, if a connection is established due to a similar function and structure, then it is possible that a protein that is below the threshold can be assigned to a family. Proteins that have very low sequence identity and have some relationships due to structural and functional similarities are put in superfamilies. Proteins having the same secondary structural arrangements are grouped into folds. It does not matter that the similarity of the proteins is based on physicochemical principles.

2.5.3 CATH/Gene3D

The CATH database was created in the mid-1990s by Prof Christine Orengo (Greene et al., 2007). Her group was interested in developing the algorithms for classifying proteins into different evolutionary families. They used sequence and structural information for this classification and eventually classified the protein structures into four categories: Class (C), Architecture (A), Topology (T), and Homologous Superfamily (H), which took the shape of a database. Protein three-dimensional structures are taken from the PDB and their individual chains are classified into domains. These domains are then classified into the above-mentioned four categories. Proteins are assigned in each category automatically; however, sometimes manual additions can be done. The Class category takes into account the proportion of secondary structural elements without taking into account their agreements or links. Four protein classes have been defined by CATH: mainly alpha (proteins with an abundance of helices), mainly beta (sheets), both helices and sheets (alpha-beta), and finally proteins with very few secondary structural elements.

Architecture based on the orientations of the secondary structures describes the shape of the domain while ignoring the connections between the secondary structures and is manually organized using a simple description of secondary structure arrangements like a 3-layer sandwich or barrel. The Topology category describes the protein form and interconnection of secondary elements. In Topology, algorithms that use empirically derived parameters are used to define the domains, which are then classified accordingly. At last, the proteins are assigned to the Homologous category if there is enough support that domains are evolutionary related, that is, proteins having a common origin. Sequences are compared to calculate the sequence similarity and after that structures are also compared based on the topological category. An extra category called sequence families is also defined where the domains are said to have similar functions owing to their high sequence identity (> 35%).

Gene3D is a subsidiary resource of CATH where proteins lacking structural information and having only sequence information are put in related domains and eventually put into the homologous categories. Protein sequences from UniProtKB and Ensembl are taken and CATH HMM is used to predict their domains and final assignment of homologous categories. CATH database is very useful in the identification of new/existing domains in proteins. Researchers frequently want to know about their newly found proteins that whether they have novel domains or not which they found very easy using CATH.

2.6 Specialized database

2.6.1 Clustering databases

It is a commonly known fact that similar sequences will often if not always tend to give similar functions. Finding similar sequences from the different databases is a time-consuming and tiring process. Clustering databases help us in this regard. They can reduce our time and efforts by precomputing sequence similarities and hence group similar sequences. The database of Clusters of Orthologous Genes/Groups (COGs) takes data in the form of protein sequences from completely sequenced genomes from different taxonomic lineages and then tries to find the pattern of the sequence

similarity which then make clusters (Tatusov et al., 2001). Each cluster in the database contains orthologous proteins or a set of paralogous sequences from at least three lineages. Orthologous sequences have the same function, and one member could be able to annotate the other members of the COG. Some specialized databases, their description, and application are represented in Table 2.2.

TABLE 2.2 Information of specialized databases and their applications.

Database name	Description/application	Weblink	References
Cluster of Orthologous Groups	Phylogenetic classification of proteins encoded in complete genomes	https://www.ncbi.nlm.nih.gov/research/cog	Tatusov et al. (2001)
PubMed	Literature	https://pubmed.ncbi.nlm.nih.gov/	NCBI
BIOSIS	Literature	https://www.ebsco.com/products/research-databases/biosis-previews	BIOSIS
ArrayExpress	Microarray data	https://www.ebi.ac.uk/arrayexpress/	Parkinson et al., 2007
BioStudies	Microarray data	https://www.ebi.ac.uk/biostudies/	Sarkans et al. (2018)
Gene Expression Omnibus	Microarray, NGS, and other forms of high-throughput functional genomics data	https://www.ncbi.nlm.nih.gov/geo/	Clough and Barrett (2016)
Taxonomy (NCBI)	Taxonomic classification of species	http://www.ncbi.nlm.nih.gov/taxonomy/taxonomyhomehtml/	NCBI
Taxonomy (UniProtKB)	Taxonomic classification of species	https://www.uniprot.org/taxonomy/	UniProt Consortium (2018)
IntAct	Interactions between nucleotides or proteins	https://www.ebi.ac.uk/intact/	Kerrien et al. (2007)
DIP	Protein interactions	https://dip.doe-mbi.ucla.edu/dip/Main.cgi	Salwinski et al. (2004)
KEGG	Pathways database	http://www.genome.jp/kegg/	Kanehisa and Goto (2000)
IntEnz	Interacting enzymes	https://www.ebi.ac.uk/intenz/	Fleischmann et al. (2004)
NC-IUBMB	Enzyme nomenclature	https://iubmb.org/	Slater (2005)
BRENDA	Enzyme database	Enzyme Database—BRENDA (brenda-enzymes.org)	Barthelmes, Ebeling, Chang, Schomburg, and Schomburg (2007)
ENZYME	Enzyme database	https://enzyme.expasy.org/	Bairoch (2000)
miRNA database	miRNA sequences and annotation	http://www.mirbase.org/	Kozomara, Birgaoanu, and Griffiths-Jones (2019)
PubChem	Small chemical substances	https://pubchem.ncbi.nlm.nih.gov	Kim et al. (2019)
ChemDB	Small molecule database	http://cdb.ics.uci.edu/	Chen, Swamidass, Dou, Bruand, and Baldi (2005)
DrugBank	Database of drugs and drug targets	https://www.drugbank.com/	Wishart et al. (2018)
VIOLIN	Vaccine data analysis and target prediction	http://www.violinet.org/	He et al. (2014)
Antijen	Binding data for drug and target	http://www.ddg-pharmfac.net/antijen/Antijen/antijenhhomepage.htm	Toseland et al. (2005)

BRENDA, Braunschweig Enzyme Database; *DIP*, Database of Interacting Proteins; *KEGG*, Kyoto Encyclopedia of Genes and Genomes; *NCBI*, National Center for Biotechnology Information; *NC-IUBMB*, Nomenclature Committee of the International Union of Biochemistry and Molecular Biology; *NGS*, next-generation sequencing; *UniProtKB*, UniProt Knowledgebase; *VIOLIN*, Vaccine Investigation and Online Information Network.

2.6.2 Bibliographic databases

Searching literature is a time-consuming and tiring process. There are thousands of journals and going through each of them is not possible. Here, bibliographic database becomes handy and saves time and resources. The bibliographic database provides a common platform for information taken from various sources, such as a journal, books, conference reports, and many others. Some of these databases are specialized in biology and medicine like PubMed. It is the largest database in the field of life sciences with over 30 million citations. PubMed can be searched through the interfaces available both at NCBI and EMBL-EBI.

BIOSIS Previews is produced by the Web of Science group. This database covers research areas like preclinical and experimental, animal studies, methodology, and many more. This also includes BIOSIS indexing and the MeSH term for diseases. It contains about 350,000 references from about 5000 journals.

2.6.3 Expression databases

Gene expression is the key to biological systems. Understanding the gene expression will help us in understanding the biological systems. Microarray experiments are at the heart of gene expression analysis, which provides a quick and reliable method for generating a huge amount of gene expression data. This enormous data are stored in an organized manner in an expression database. ArrayExpress is a public expression database for storing transcription data (Parkinson et al., 2007). It uses Minimum Information About a Microarray Experiment format, which is the standard format created by the MGED society for annotation and data storage. Recently started (October 2020), all the ArrayExpress data are being transferred to the BioStudies database (Sarkans et al., 2018). The BioStudies database describes the biological studies along with their links to the other databases, such as EMBL-EBI or elsewhere. BioStudies, via a simple format, can take a wide range of studies.

Gene Expression Omnibus (GEO) is a database that provides free microarray data, NGS data, and other high-throughput functional genomics data submitted by various researchers (Clough & Barrett, 2016). Various applications and web-based interfaces are available, which assist in the querying and downloading of the data. GEO serves three main goals:

- It efficiently stores high-throughput functional genomics data.
- Data can be submitted in a very simple and straight forward manner.
- Easy querying and downloading.

2.6.4 Taxonomy databases

Databases that store information about the taxonomical classification of any organism are known as taxonomy database. NCBI Taxonomy database is the largest taxonomical database that contains the name and classification of every sequence present in INSDC. Information is available for every known, living or extinct and unknown organism. Cross-referencing is present for taxonomy databases from other databases, such as NCBI, EMBL, and UniProt. UniProtKB also has a taxonomy database called “Taxonomy,” which is a manually curated database. They also provide external links, strains of organisms if present, and information about a viral host.

2.6.5 Interaction databases

All biology is interactions. All the biological processes work by interacting with each other. Hence, describing an interaction is very important for understanding biological mechanisms. Databases dedicated to interactions between various biomolecules, such as DNA, RNA, and proteins, model various interactions and tell us how systems and molecules are interrelated. IntAct molecular interaction database is an open-source database and also provides analysis tools for molecular interaction data (Kerrien et al., 2007). All the interactions present in IntAct are taken either from literature curation or by direct submission. Interactions are searchable and can be viewed using an interactive graphical user interface for protein networks.

Database of Interacting Proteins (DIP) stores only those interactions between proteins, which are experimentally determined (Salwinski et al., 2004). It collects data from various sources and presents them as one set of protein–protein interactions. The data in DIP are both manually and automatically curated. Automatic curation is done by computational approaches that are driven by the knowledge acquired by the reliable DIP data for protein–protein interactions. Currently, it contains data for 28,850 proteins having 81,923 interactions.

2.6.6 Pathway databases

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database is a collection of databases related to pathways, chemical reactions, genes, and proteins (Kanehisa & Goto, 2000). Each pathway in KEGG is manually drawn from knowledge taken from published experimental details on metabolism and various other functions of the organisms. A pathway in KEGG contains a network of molecular interactions where genes are linked to the respective gene product (s) (proteins). This helped the researchers to develop a KEGG pathway mapping tool in which we can provide the gene details of a genome, which are then compared to the database and hence provide the putative pathway information about that genome. The databases at KEGG are divided into four categories based on the information present, which are systems, genomic, chemical, and health information. Systems information databases comprise PATHWAY, MODULE, and BRITE. Genomic information contains GENOME, GENES, SSDB, and ORTHOLOGY. In chemical information, four databases are present, that is, COMPOUND, GLYCAN, REACTION, and ENZYME. Health information databases have NETWORK, VARIANT, DRUG, DISEASE, and ENVIRON.

2.6.7 Enzyme databases

Nomenclature Committee (NC) of the International Union of Biochemistry and Molecular Biology (IUBMB) created the Integrated relational Enzyme database (IntEnz) database (Fleischmann et al., 2004). This database combines data from three different enzyme databases, which are the NC-IUBMB Enzyme Classification list (Slater, 2005), the Braunschweig Enzyme Database (BRENDA) of enzyme function (Barthelmes et al., 2007), and the Enzyme Nomenclature database (ENZYME), and incorporates all of their contents (Bairoch, 2000). The ENZYME database contains the data of each enzyme along with their Enzyme Commission number. Other information is also available for each record, such as alternative names, cofactors, catalytic activities, disease information, and also cross-reference with UniProt. BRENDA has a similar kind of records along with a classification by species also.

2.6.8 microRNA database

miRBase is the most popular database for miRNA (miR) (Kozomara, Birgaoanu, & Griffiths-Jones, 2019). It contains published miR sequences and annotations. Data (miR transcripts) are stored in miRBase such that they represent a predicted hairpin loop structure, which is known as mir in the database. Other information regarding the sequence and location of mature miR is also present. A user can search both hairpin and mature sequences through searching platforms.

2.6.9 Small molecule database

NCBI's PubChem (Kim et al., 2019) is a database that stores information about small chemical molecules along with their biological activities. Data present in PubChem are being contributed from several hundred organizations, which may be a governmental organization, such as NCBI Structure, NCBI taxonomy, or any chemical vendor, or even from journals. It comprises three components: PubChem Substance, PubChem Compound, and PubChem BioAssay. PubChem Substance is used for the search of different substances that have been produced by various manufacturers. Also, you can search for substances of unknown composition and substances whose molecular structure is yet not known. PubChem Compound contains about 109 million compounds (November 2020) along with their two-dimensional structures. The graphical user interface as a molecular structure editor is available for querying where one can design a full or partial structure and then search it.

In addition to this, other criteria can be used for searching certain physicochemical parameters, such as molecular weight range, acceptors or donors of hydrogen bonds or a *P*-value range, and several others. The records of the two databases are linked and point toward a third database, PubChem BioAssay, if the particular data are available. Bioactivity results of molecules are stored in PubChem BioAssay and Entrez is used for the query process. There are many applications in the PubChem database due to the presence of many internal and external database linkages. For example, if you have a known enzyme inhibitor, you can find other possible similar inhibitors. Moreover, the identification of small molecules with different structures but similar biological effects can be identified. PubChem is the largest open-source chemistry database and hence is an important database when it comes in terms of drug discovery since it contains active chemical substances from a variety of sources which increases the probability of successful lead identification.

Another small molecule database that can be used for drug designing is ChemDB (Chen et al., 2005). It contains small molecules from various public databases and digital catalogs of more than 150 vendors. All the substances are annotated with properties, such as physicochemical, three-dimensional structure, and solubility. Molecules can be downloaded in various formats. Moreover, chemical reaction information is available in this database. A fuzzy text matching algorithm can be used to search the database, which gives better results. DrugBank is a public database that provides information about drugs and their targets (Wishart et al., 2018). This database combines both drugs and their targets to provide chemical and pharmacological information, such as sequence, structure, and pathways.

2.6.10 Vaccine design database

Vaccine Investigation and Online Information Network (VIOLIN) is an important database for vaccine designing providing curated data with various tools that aids in vaccine design (He et al., 2014). All the vaccine research data available from the humans and lab animals are available in VIOLIN. Data can be downloaded in XML format from VIOLIN.

Another vaccine-related database is the AntiJen, which focuses on the kinetic, thermodynamic, cellular, and molecular information (Toseland et al., 2005). AntiJen is a database framework concentrated on the coordination of kinetic, thermodynamic, cellular, and molecular information. Recently AntiJen v2.0 is released, which contains quantitative binding data for peptides binding to TCR–MHC complexes, MHC ligand, B-cell epitope, T-cell epitope, TAP, and immunological protein–protein interactions. It also includes copy numbers, peptide library, and diffusion coefficient. All the data are experimentally derived, and no predicted data are available.

2.7 Database searching and annotation

Searching a database is a very important aspect of the usability of the database. A good search program will help in finding relevant information more quickly and efficiently. Two main search programs, that is, Entrez and Sequence Retrieval System (SRS), are available for NCBI and EBI, respectively.

2.7.1 Entrez

Entrez is used to query the GenBank database and subsequently used for the query of all databases associated with NCBI. Entrez is a very efficient tool as it can be used for both simple and complicated queries. In Entrez, one can use single search terms that can be used for certain database fields and can make use of logical operators (AND, OR, NOT) for combining the search terms. In the case of a single term search, a field-ID is placed after the search term as in search term [field-ID]. For example, if one wants to search a sequence from a particular species whose length is between 3000 and 4000 base pairs, it would require a search pattern, such as species_name [ORGN] AND 3000:4000 [SLEN]. A complete reference on how to use Entrez can be found on the Entrez help page.

2.7.2 The sequence retrieval system

As stated above, there is a common query system for both DDBJ and EMBL which is based on SRS. There are two simple web-based forms for basic queries (DDBJ-SRS and EBI-SRS). However, SRS can also handle complex queries. The SRS was developed at EBI and it helps in managing the primary and secondary biological databases (Etzold, Ulyanov, & Argos, 1996). The basic functioning of the SRS at DDBJ or SBI is the same. In the EBI-SRS, one can also save their searches by the establishment of a permanent project by creating an account. This allows restarting the interrupted searches. The EBI-SRS help page describes in detail how to use the SRS.

2.7.3 Annotation

Though the biological databases provide a tremendous amount of data that looks fascinating at a first glance, there is much more we need to add to this raw data to make it more meaningful. This process is known as an annotation. In this process, experts organize and interpret the available raw data to extract biological inferences. Data generation starts with raw sequencing reads generated by various NGS technologies. These raw reads are assembled using various assembly tools and a genome assembly is generated. Now the subsequent genome annotation starts, which includes identification of genes with coding and noncoding regions, pseudogenes, repeats, regulatory elements, and many more.

The coding regions can be translated into protein sequences, which are of particular interest for three-dimensional structure prediction, posttranslational modification, and active site prediction or can be used for potential drug targets (Ofraan, Punta, Schneider, & Rost, 2005). Last but not the least, system biology combines both genomics and proteomics to look at the organism level. Transcripts' expression level, transcription factors, and interaction networks are the main areas of studies in systems biology.

Some genome annotation pipeline has been developed, which assists in the genome annotations. UCSC genome browser (Haeussler et al., 2019) and tools provided at NCBI (Sayers et al., 2020) are a couple of examples. The ENSEMBLE pipeline also provides automated genome annotations and predicts gene structure (Flicek et al., 2008). Annotations can be done by three methods: manual, similarity-based, and ab initio.

In manual curation, highly trained curators search the literature for any experimental evidence supporting the data and then attach that information with the data. This type of annotation provides the highest accuracy in datasets. However, this type of annotation is a time-consuming process and can cover only a small fraction of available data as experimental evidence is not available for most of the data. Swiss-Prot database of UniProtKB is a manually curated highly accurate database for protein annotations.

The second is the similarity-based methods in which the already annotated data are used to annotate similar sequences that are considered as homologous. Homologous sequences are believed to share a common ancestor and with that, share sequences, structures, and functions. Some very effective sequence comparison methods have been developed, which can identify homologous sequences, such as PSI-BLAST (Altschul et al., 1997) and HMM (Eddy, 1996). These methods make sequence profiles from related sequences and then use these profiles to identify/annotate new protein sequences.

The last method is the ab initio method, which uses predefined rules to predict the feature. These rules are not arbitrarily but are based on training according to previous annotations available. Rosetta method is a popular ab initio method that predicts protein folds by calculating the energy of different conformations of a protein structure (Das et al., 2007). Ab initio methods are also used for the prediction of gene promoters (Goni, Perez, Torrents, & Orozco, 2007) and transcription factors (Kaplan, Friedman, & Margalit, 2005).

2.8 Conclusions

The use of high-performance computational platforms has led researchers to perform research experiments in a considerably less amount of time. Although the experiments take less time, a huge amount of raw data also gets generated. Biological databases provide an easy and efficient infrastructure for storing this data. Based on the type of data stored, a database can be categorized into one of the several categories. Databases also analyze these data to extract many biological meaningful insights. Some database also provides various online and offline tools, which again help in data analysis. This chapter gives an overview of databases, which are essentially become a part of the day-to-day bioinformatics analysis pipeline. It provides the database classification, type of data stored, and tools available for analysis. The growth of biological databases will help in our understanding of biological systems.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402.
- Andreeva, A., Kulesha, E., Gough, J., & Murzin, A. G. (2020). The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, 48(D1), D376–D382.
- Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., . . . Yeh, L. S. (2004). UniProt: The Universal Protein knowledgebase. *Nucleic Acids Research*, 32(Database issue), D115–D119.
- Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., . . . Zygouri, C. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research*, 31(1), 400–402.
- Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, 28(1), 304–305.
- Barthelme, J., Ebeling, C., Chang, A., Schomburg, I., & Schomburg, D. (2007). BRENDA, AMENDA and FRENDA: The enzyme information system in 2007. *Nucleic Acids Research*, 35(Database issue), D511–D514.

- Benson, D. A., Cavanaugh, M., Fau, -, Clark, K., Clark, K., Fau, -, Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, 1362–4962, D36–D42.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1), 365–370.
- Bourne, P. (2005). Will a biological database be different from a biological journal? *PLoS Computational Biology*, 1(3), 179–181.
- Chen, J., Swamidass, S. J., Dou, Y., Bruand, J., & Baldi, P. (2005). ChemDB: A public database of small molecules and related cheminformatics resources. *Bioinformatics (Oxford, England)*, 21(22), 4133–4139.
- Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods in Molecular Biology*, 1418, 93–110.
- Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Baker, D. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins*, 69(8), 118–128.
- Dawson, W., & Kawai, G. (2009). Modeling the chain entropy of biopolymers: Unifying Two different random walk models under one framework. *Journal of Computer Science & Systems Biology*, 02, 001–023.
- Dayhoff, M. O. (1965). *Atlas of protein sequence and structure*. National Biomedical Research Foundation.
- Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, 6(3), 361–365.
- EMBL. (2013). EMBL Genome Database. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho & H. Yokota (Eds.), *Encyclopedia of systems biology* (pp. 653–653). New York, NY: Springer.
- Etzold, T., Ulyanov, A., & Argos, P. (1996). SRS: Information retrieval system for molecular biology data banks. *Methods in Enzymology*, 266, 114–128.
- Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K. B., Apweiler, R. (2004). IntEnz, the integrated relational enzyme database. *Nucleic Acids Research*, 32(Database issue), D434–D437.
- Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Searle, S. (2008). Ensembl 2008. *Nucleic Acids Research*, 36(Database issue), D707–D714.
- Goni, J. R., Perez, A., Torrents, D., & Orozco, M. (2007). Determining promoter location based on DNA structure first-principles calculations. *Genome Biology*, 8(12), R263.
- Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Orengo, C. A. (2007). The CATH domain structure database: New protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Research*, 35(Database issue), D291–D297.
- Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Kent, W. J. (2019). The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research*, 47(D1), D853–D858.
- He, Y., Racz, R., Sayers, S., Lin, Y., Todd, T., Hur, J., Xiang, Z. (2014). Updates on the web-based VIOLIN vaccine database and analysis system. *Nucleic Acids Research*, 42(Database issue), D1124–D1132.
- Hoskeri, J., Krishna, V., & Amruthavalli, C. (2010). Functional annotation of conserved hypothetical proteins in *Rickettsia Massiliae* MTU5. *Journal of Computer Science & Systems Biology*, 03(02), 050–052.
- Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuče, B. A., de Castro, E., Sigrist, C. J. (2008). The 20 years of PROSITE. *Nucleic Acids Research*, 36(Database issue), D245–D249.
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1), 27–30.
- Kaplan, T., Friedman, N., & Margalit, H. (2005). Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Computational Biology*, 1(1), e1.
- Karthick Raja, N., Selvaraj, J., & Muthukumar. (2008). Prediction of three dimensional model and active site analysis of inducible serine protease inhibitor-2 (ISPI-2) in *Galleria mellonella*. *Journal of Computer Science & Systems Biology*, 01, 119–125.
- Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Hermjakob, H. (2007). IntAct—Open source resource for molecular interaction data. *Nucleic Acids Research*, 35(Database issue), D561–D565.
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., . . . Bolton, E. E. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research*, 47(D1), D1102–D1109.
- Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: From microRNA sequences to function. *Nucleic Acids Research*, 47(D1), D155–D162.
- Liu, Z., Liu, Y., Liu, S., Ding, X., & Yang, Y. (2009). Analysis of the sequence of ITS1-5.8S-ITS2 regions of the three species of fructus *Evodiae* in Guizhou Province of China and identification of main ingredients of their medicinal chemistry. *Journal of Computer Science and Systems Biology*, 2, 200–207.
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar Gustavo, A., Sonnhammer, E. L. L., Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1), D412–D419.
- Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Finn, R. D. (2019). InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1), D351–D360.
- Ofran, Y., Punta, M., Schneider, R., & Rost, B. (2005). Beyond annotation transfer by homology: Novel protein-function prediction methods to assist drug discovery. *Drug Discovery Today*, 10(21), 1475–1482.

- Ogasawara, O., Kodama, Y., Mashima, J., Kosuge, T., & Fujisawa, T. (2019). DDBJ Database updates and computational infrastructure enhancement. *Nucleic Acids Research*, *48*(D1), D45–D50.
- Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farnie, A., Brazma, A. (2007). ArrayExpress—A public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*, *35*(Database issue), D747–D750.
- Ragunath, P. K., Venkatesan, P., & Rajmohan, R. (2009). New curriculum design model for bioinformatics postgraduate program using systems biology approach. *Journal of Computer Science & Systems Biology*, *02*, 300–305.
- Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., & Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Research*, *32*(Database issue), D449–D451.
- Sarkans, U., Gostev, M., Athar, A., Behrangi, E., Melnichuk, O., Ali, A., McEntyre, J. (2018). The BioStudies database—One stop shop for all data supporting a life sciences study. *Nucleic Acids Research*, *46*(D1), D1266–D1270.
- Sayers, E. W., Beck, J., Brister, J. R., Bolton, E. E., Canese, K., Comeau, D. C., Ostell, J. (2020). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *48*(D1), D9–D16.
- Shanthi, V., Ramanathan, K., & Sethumadhavan, R. (2009). Role of the cation- π interaction in therapeutic proteins: A comparative study with conventional stabilizing forces. *Journal of Computer Science & Systems Biology*, *2*, 051–068.
- Sharma, N., Vrat, B. S., Kumud, J., Thakur, P. D., & Rajinder, K. (2011). Comparative in silico analysis of partial coat protein gene sequence of zucchini yellow mosaic virus infecting summer squash (*Cucurbita pepo* L.) isolated from India. *Journal of Proteomics & Bioinformatics*, *04*, 068–073.
- Singh, S., Gupta, S., Nischal, A., Khattri, S., Nath, R., Seth, P., & Kumar, P. (2010). Comparative modeling study of the 3-D structure of small delta antigen protein of hepatitis delta virus. *Journal of Computer Science & Systems Biology*, *3*, 001–004.
- Slater, E. C. (2005). This is the IUBMB history—The history of IUB (MB). *IUBMB Life*, *57*(4/5), 203–211.
- Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics (Oxford, England)*, *23*(10), 1282–1288.
- Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Koonin, E. V. (2001). The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, *29*(1), 22–28.
- Toseland, C. P., Clayton, D. J., McSparron, H., Hemsley, S. L., Blythe, M. J., Paine, K., Flower, D. R. (2005). AntiJen: A quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Research*, *1*(1), 4.
- UniProt Consortium, T. (2018). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *46*(5), 2699.
- Varsale, A. R., Wadnerkar, A. S., Mandage, R. H., & Jadhavrao, P. K. (2010). Cheminformatics. *Journal of Proteomics & Bioinformatics*, *3*, 253–259.
- Vaseeharan, B., & Valli, S. J. (2011). In silico homology modeling of prophenoloxidase activating factor serine proteinase gene from the haemocytes of *Fenneropenaeus indicus*. *Journal of Proteomics & Bioinformatics*, *4*, 053–057.
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082.
- Wu, C. H., Yeh, L. S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Barker, W. C. (2003). The protein information resource. *Nucleic Acids Research*, *31*(1), 345–347.