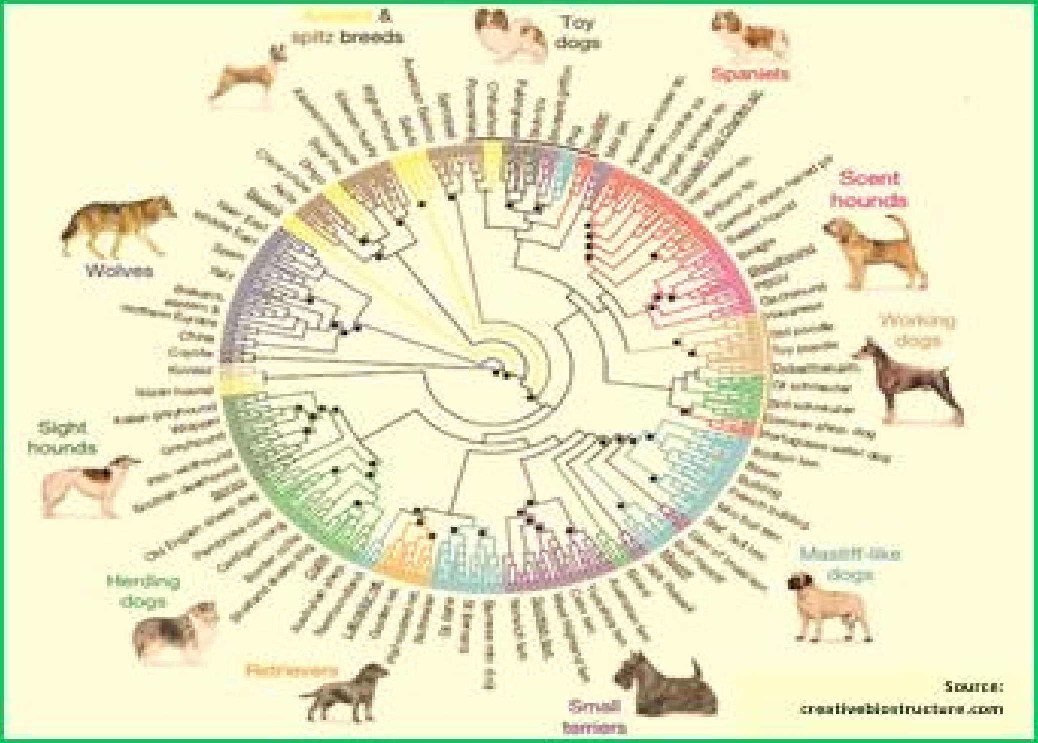
## e-TRAINING

## ON

# Biocomputational Interventions to Analyse Canine & Livestock Genome

## October 6-9, 2020

## e-COMPENDIUM



Source: creativebiostructure.com

*Organizers*

Prof. Ranjna Cheema, Head Department of Bioinformatics
Prof. R. S. Sethi, Head, Department of Biotechnology

*Organized by*

Project Monitoring Unit – DBT-GADVASU-Canine Research center &
Department of Bioinformatics, College of Animal Biotechnology
Guru Angad Dev Veterinary & Animal Sciences University
Ludhiana- 141004, Punjab (India)

*Sponsored by*

DBT Funded Network Research Project:
Parentage determination and cytogenetic profiling in dogs

e-Training

on

# BIOCOMPUTATIONAL INTERVENTIONS TO ANALYSE CANINE & LIVESTOCK GENOMES

Sponsored by

DBT funded network research project ñParentage determination and cytogenetic profiling in dogsò

October 6th to 9th, 2020

# e-COMPENDIUM OF LECTURES

*Co-ordinated, Edited & Compiled by*

**C. S. Mukhopadhyay**
**Anuj Tyagi**
**P. P. Dubey**

**Organized by**
**Project Monitoring Unit-DBT-GADVASU-Canine Research Center & Department of Bioinformatics,**
**College of Animal Biotechnology**
**Guru Angad Dev Veterinary and Animal Sciences University**
**Ludhiana-141004 (Punjab) INDIA**

# Biological Databases

*Anuj Tyagi*
*College of Fisheries*
*Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana 141004, Punjab, India*

## What is a database?

In simple terms, a database is defined as an organized collection of data or information that is electronically stored and accessible from a computer system. The organized nature of the database makes it easy to access, manage, periodically update, and rapidly search the required data/information from a suitable computer system [1].

## Biological databases and their importance

Among various types of databases, the ones constituting the datasets relevant to biological sciences such as molecular biology and bioinformatics are called biological databases. In the current scenario, the importance of biological databases can be understood from the following points [2]:

1. Due to rapidly advancing molecular biology, proteomics, and low-cost high-throughput genome sequencing technologies, huge amounts of biological information such as raw sequencing datasets, proteomes, etc. are being generated at a very rapid rate. Thus, the storage and handling of this staggering information are the major challenges of the current genomics era.
2. In addition to generation, data analysis and drawing of meaningful conclusions are also important parts of any scientific research. This often requires data sharing within the diverse scientific community. In this context, the biological database enables the scientists to access and retrieve the biologically relevant data including the raw data, genome sequences, analyzed datasets, and annotations in easily manageable/organized formats.
3. Biological databases also allow data indexing as well as help remove the data redundancy.
4. At present, biological databases have become the central component of bioinformatics. Through the various data mining tools, all biological information can be easily accessed; thus saving time, resources, and efforts.

## Components of biological database

Similar to other databases, a biological database also has certain basic components (Fig. 1). These are:

a. **Entity** - An entity refers to the thing we want to store in a database. Eg. DNA sequences, Genes, Bibliographic references, etc.

b. **Fields** - The properties of an entity are called fields. Eg. Gene name, gene sequence, mutation (if any), etc.

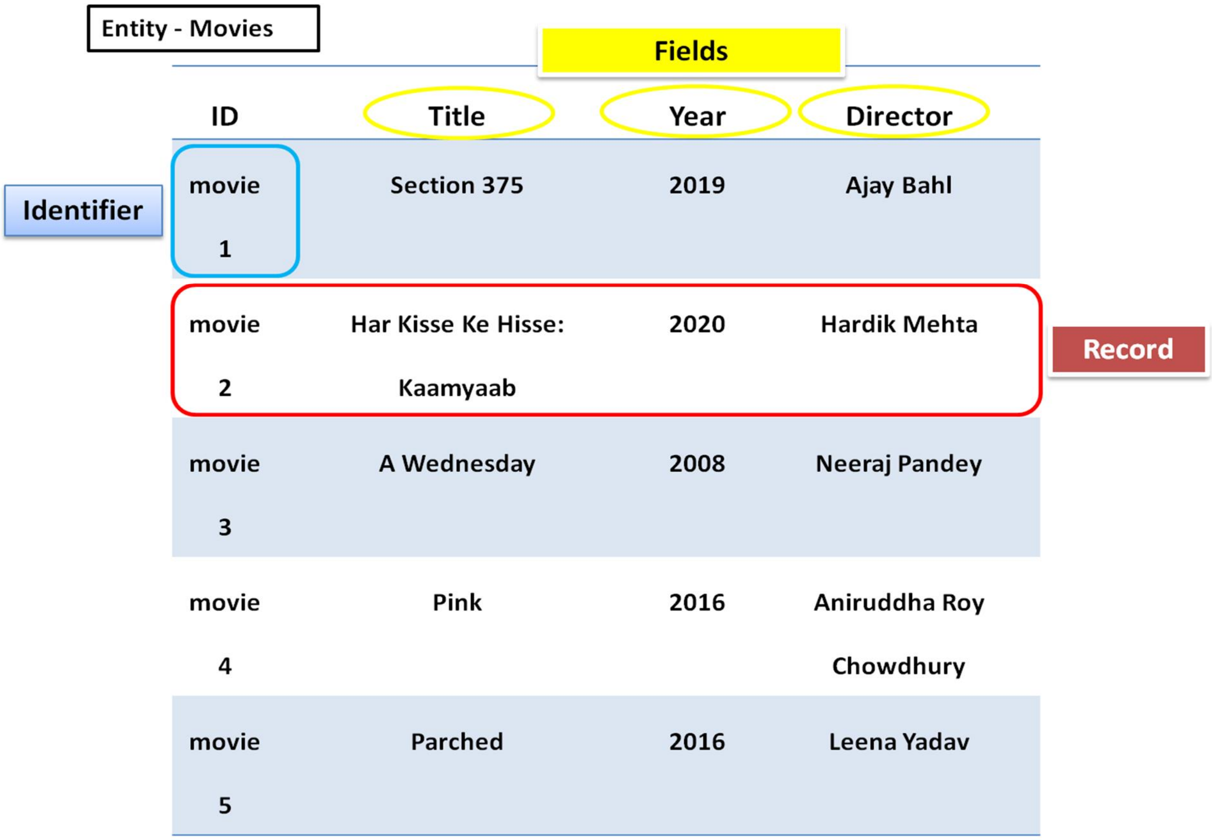
c. **Records** - A record typical refers to a combination of all the fields for a given entity. For eg. Record for gene BRCA1 in GenBank

d. **Identifier** - The unique name which identifies a record.

In the case of a simple database, a single file contains multiple records. Among these records, each one can have the same set of information (fields) along with a unique identifier.

Various components of a database could be easily understood from the below-mentioned example of a database of "Selected movies of Indian Cinema". In the below mentioned Fig. 1:

- The entities stored are movies.
- The records are each row of the table including the movie name.
- The field refers to the columns of the table i.e., Title, Year, Director
- The unique identifiers are movie1, movie2, etc.



| ID | Title | Year | Director |
|---|---|---|---|
| movie 1 | Section 375 | 2019 | Ajay Bahl |
| movie 2 | Har Kisse Ke Hisse: Kaamyaab | 2020 | Hardik Mehta |
| movie 3 | A Wednesday | 2008 | Neeraj Pandey |
| movie 4 | Pink | 2016 | Aniruddha Roy Chowdhury |
| movie 5 | Parched | 2016 | Leena Yadav |

**Fig. 1: Example of a typical tabular database with each row containing a separate record along with distinct fields/attributes in the columns.**

### Types of biological databases

Based on their content, the biological databases can be classified into the following types [3]:

#### a. Primary databases

Primary databases, also known as the archival databases, basically contain experimentally derived datasets such as nucleotide and protein sequences as well as the structural information of macromolecules. This basic information can be accompanied by functional annotation, bibliographies, and links to other databases. The data to the primary database is directly submitted by researchers. Once submitted, the data is assigned an accession number, which is permanent and becomes a part of the scientific record [2]. The followings are examples of primary databases:

i. **Primary nucleotide sequence databases** - The European Nucleotide Archive (ENA), The National Center for Biotechnology Information GenBank (NCBI GenBank), and The DNA Data Bank of Japan (DDBJ), etc.

ii. **Microarray/Functional genomics databases** - Gene Expression Omnibus (GEO) and Array Express Archives etc.

iii. **Protein sequences and structure databases** - Swiss-Prot and Protein Information Resource (PIR) for protein sequences, Protein Databank (PDB) for protein structure.

#### b. Secondary databases

Secondary databases store the information derived from the analysis of primary datasets. Secondary databases contain highly curated information derived from complex computational as well as manual analysis of primary resources and scientific literature. These databases often store information about conserved domain structure/sequences, signal sequences, and active site residues [2].

i. **Protein families, domains and structure databases** - InterPro, PROSITE, SCOP, CATH and NCBI Conserved Domain Database (CDD)

ii. **Protein sequences and functional information databases** - UniProt Knowledgebase (UniProtKB)

iii. **Nucleotide (Genes/Genomes) sequence and annotation databases** - NCBI UniGene, The European Bioinformatics Institute (EBI) Genomes (EBI Genomes), and Ensembl, etc.

#### c. Specialized databases

These databases cater to the needs of specific research interests. Eg. Ribosomal Project Database (RDP), HIV sequence database, The Saccharomyces Genome Database (SGD), Mouse Genome Database (MGD), and Antibiotic Resistance Genes Database (ARDB), etc.

**d.** Sometimes a database can function as both a primary and secondary database. For example, primary peptide sequences can be directly submitted to Uniprot. Besides, Uniprot is also able to infer protein sequences from primary nucleotide sequences, also it can have the automated and manual annotations derived from TrEMBL and SwissProt, respectively.

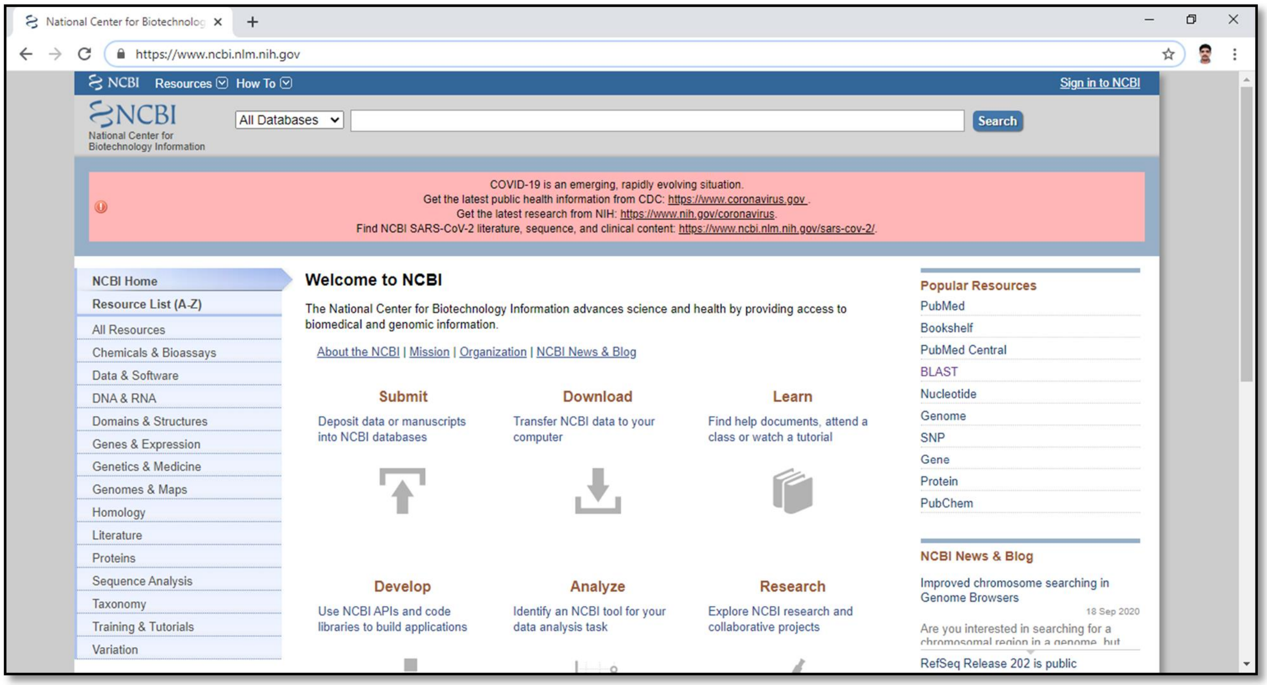
### Important biological databases

Based on their scope, some of the important biological databases have been described below:

### A. Nucleotide sequence databases

### a. The National Center for Biotechnology Information GenBank (NCBI GenBank)

NCBI is a part of the National Library of Medicine (NLM) under the U.S. National Institute of Health (NIH). Since its inception, NCBI has been playing a leading role in bioinformatics by providing online access to biological datasets/information and computational resources/tools to the millions of researchers across the world [4]. Accessible online at https://www.ncbi.nlm.nih.gov/ NCBI hosts approximately 40 online biological databases (Fig. 2), and GenBank is one among them.



COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Improved chromosome searching in Genome Browsers
18 Sep 2020
Are you interested in searching for a chromosomal region in a genome, but

RefSeq Release 202 is public

**Fig. 2: The NCBI Interface at https://www.ncbi.nlm.nih.gov/**

GenBank is a publically available collection of nucleotide sequences, their protein sequences along with annotations [4]. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on daily basis. Being a primary database, the GenBank accepts sequence submission directly from individual scientists and laboratories. It also accepts batch submissions from large-scale sequencing projects. Several types of datasets such as single mRNA or gene sequence, draft, and complete genome assemblies, transcriptome assemblies, and third party annotations are accepted. GenBank continues to grow at an exponential rate, doubling every 18 months. Some of the unique features of Genbank are:

- Being a free public repository, any researcher can submit the sequences to GenBank without incurring any financial cost.
- For the same gene or genome, multiple sequences of varying quality are available in GenBank. Essentially, anything submitted to GenBank is stored.
- To represent the various modification done by the author, a sequence can have several versions.
- Once a record is submitted to GenBank, it is assigned to a specific division based on the source taxonomy or sequencing strategy used to obtain the data. There are 12 taxonomic divisions and 8 functional divisions (Table 1) [4].
- Each GenBank record containing a sequence is assigned a unique identifier called an accession number. The accession number is permanent, and it stays the same throughout the life of the record. Only changes may occur in the sequence version but not in the accession number. For eg. ACCESSION AF000001, VERSION AF000001.5

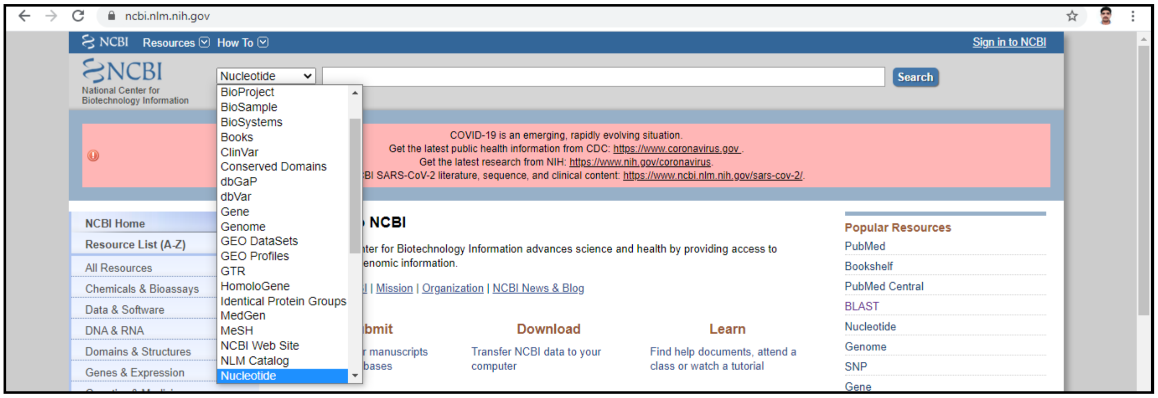
**Table 1: Taxonomic and functional divisions in GenBank**

| **Taxonomic Divisions** | | **Functional Divisions** | |
|---|---|---|---|
| **Division** | **Description** | **Division** | **Description** |
| **SYN** | Synthetic | **TSA** | Transcriptome shotgun data |
| **PHG** | Phages | **WGS** | Whole-genome shotgun data |
| **ENV** | Environmental samples | **PAT** | Patented sequences |
| **VRL** | Viruses | **GSS** | Genome survey sequences |
| **BCT** | Bacteria | **EST** | Expressed sequence tags |
| **PLN** | Plants | **HTG** | High-throughput genomic |
| **MAM** | Other mammals | **STS** | Sequence tagged sites |
| **VRT** | Other vertebrates | **HTC** | High-throughput cDNA |
| **PRI** | Primates | | |
| **UNA** | Unannotated | | |
| **ROD** | Rodents | | |
| **INV** | Invertebrates | | |

## Retrieving the data from GenBank

Almost 2.7 million users daily access the various databases of NCBI. Entrez search and retrieval engine is used for retrieval of data from these resources including the GenBank sequence records. The Entrez system can be accessed from the search box at the top of the NCBI page. The search box also has the option for the dropdown menu for the selection of individual bases (Fig. 2). The Entrez system comprises 39 molecular and literature databases (Fig. 3).

Entrez queries can be single words, short phrases, sentences, database identifiers, gene symbols, or names é  just about anything. Often simple searches can result in overwhelming numbers of results or even no results at all. Several built-in Entrez features can help in creating more effective queries. These include Boolean operators (AND, OR, NOT), query translation, and fielded searching using any of the indexed fields available for the database.



COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
BI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

NCBI

ter for Biotechnology Information advances science and health by providing access to enomic information.

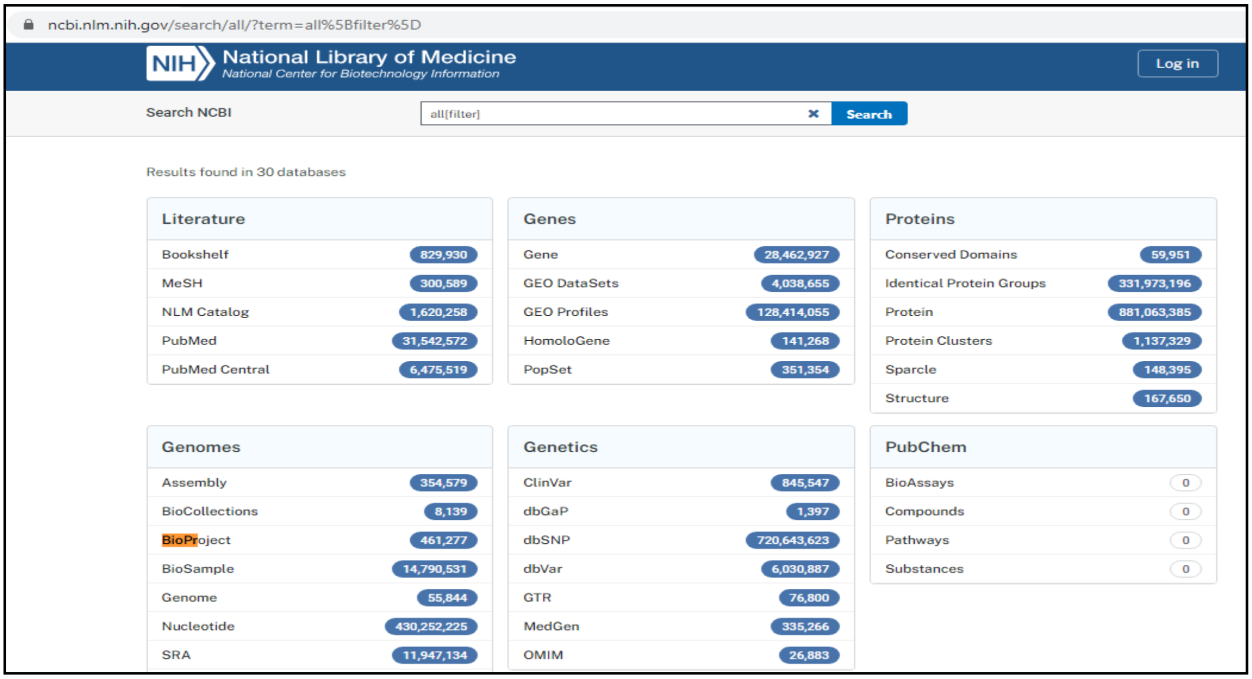
| Mission | Organization | NCBI News & Blog

r manuscripts bases

Transfer NCBI data to your computer

Find help documents, attend a class or watch a tutorial

**Fig. 2: Accessing the Entrez search box with a dropdown database selection menu**


ncbi.nlm.nih.gov/search/all/?term=all%5Bfilter%5D
NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
Search NCBI
all[filter]
Search

Results found in 30 databases

| Literature | |
|---|---|
| Bookshelf | 829,930 |
| MeSH | 300,589 |
| NLM Catalog | 1,620,258 |
| PubMed | 31,542,572 |
| PubMed Central | 6,475,519 |

| Genes | |
|---|---|
| Gene | 28,462,927 |
| GEO DataSets | 4,038,655 |
| GEO Profiles | 128,414,055 |
| HomoloGene | 141,268 |
| PopSet | 351,354 |

| Proteins | |
|---|---|
| Conserved Domains | 59,951 |
| Identical Protein Groups | 331,973,196 |
| Protein | 881,063,385 |
| Protein Clusters | 1,137,329 |
| Sparcle | 148,395 |
| Structure | 167,650 |

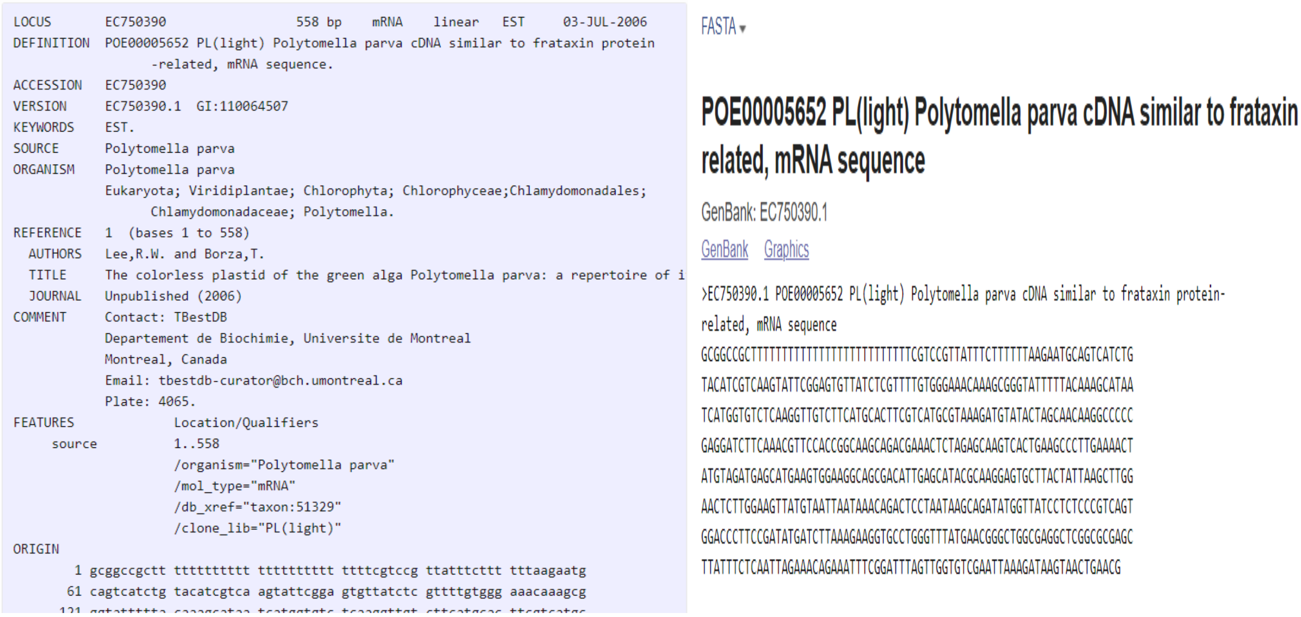
| Genomes | |
|---|---|
| Assembly | 354,579 |
| BioCollections | 8,139 |
| BioProject | 461,277 |
| BioSample | 14,790,531 |
| Genome | 55,844 |
| Nucleotide | 430,252,225 |
| SRA | 11,947,134 |

| Genetics | |
|---|---|
| ClinVar | 845,547 |
| dbGaP | 1,397 |
| dbSNP | 720,643,623 |
| dbVar | 6,030,887 |
| GTR | 76,800 |
| MedGen | 335,266 |
| OMIM | 26,883 |

| PubChem | |
|---|---|
| BioAssays | 0 |
| Compounds | 0 |
| Pathways | 0 |
| Substances | 0 |

**Fig. 3: The Entrez Global Query results page showing the results of a search for all records in the databases (all[Filter]). The search was performed from the Entrez search box on NCBI main page.**

Sequencing data from GenBank can be retrieved in several formats (Fig. 4A & B).



```
LOCUS       EC750390                 558 bp    mRNA    linear   EST     03-JUL-2006
DEFINITION  POE00005652 PL(light) Polytomella parva cDNA similar to frataxin protein
            -related, mRNA sequence.
ACCESSION   EC750390
VERSION     EC750390.1  GI:110064507
KEYWORDS    EST.
SOURCE      Polytomella parva
ORGANISM    Polytomella parva
            Eukaryota; Viridiplantae; Chlorophyta; Chlorophyceae;Chlamydomonadales;
            Chlamydomonadaceae; Polytomella.
REFERENCE   1  (bases 1 to 558)
  AUTHORS   Lee,R.W. and Borza,T.
  TITLE     The colorless plastid of the green alga Polytomella parva: a repertoire of i
  JOURNAL   Unpublished (2006)
COMMENT     Contact: TBestDB
            Departement de Biochimie, Universite de Montreal
            Montreal, Canada
            Email: tbestdb-curator@bch.umontreal.ca
            Plate: 4065.
FEATURES             Location/Qualifiers
     source          1..558
                     /organism="Polytomella parva"
                     /mol_type="mRNA"
                     /db_xref="taxon:51329"
                     /clone_lib="PL(light)"
ORIGIN
        1 gcggccgctt tttttttttt tttttttttt ttttcgtccg ttatttcttt tttaagaatg
       61 cagtcatctg tacatcgtca agtattcgga gtgttatctc gttttgtggg aaacaaagcg
      121 ggtattttta caaagcataa tcatggtgtc tcaaggttgt cttcatgcac ttcgtcatgc
```

FASTA ▾

**POE00005652 PL(light) Polytomella parva cDNA similar to frataxin related, mRNA sequence**

GenBank: EC750390.1

GenBank Graphics

```
>EC750390.1 POE00005652 PL(light) Polytomella parva cDNA similar to frataxin protein-
related, mRNA sequence
GCGGCCGCTTTTTTTTTTTTTTTTTTTTTTTTTTCGTCCGTTATTTCTTTTTTAAGAATGCAGTCATCTG
TACATCGTCAAGTATTCGGAGTGTTATCTCGTTTTGTGGGAAACAAAGCGGGTATTTTTACAAAGCATAA
TCATGGTGTCTCAAGGTTGTCTTCATGCACTTCGTCATGCGTAAAGATGTATACTAGCAACAAGGCCCCC
GAGGATCTTCAAACGTTCCACCGGCAAGCAGACGAAACTCTAGAGCAAGTCACTGAAGCCCTTGAAAACT
ATGTAGATGAGCATGAAGTGGAAGGCAGCGACATTGAGCATACGCAAGGAGTGCTTACTATTAAGCTTGG
AACTCTTGGAAGTTATGTAATTAATAAACAGACTCCTAATAAGCAGATATGGTTATCCTCTCCCGTCAGT
GGACCCTTCCGATATGATCTTAAAGAAGGTGCCTGGGTTTATGAACGGGCTGGCGAGGCTCGGCGCGAGC
TTATTTCTCAATTAGAAACAGAAATTTCGGATTTAGTTGGTGTCGAATTAAAGATAAGTAACTGAACG
```

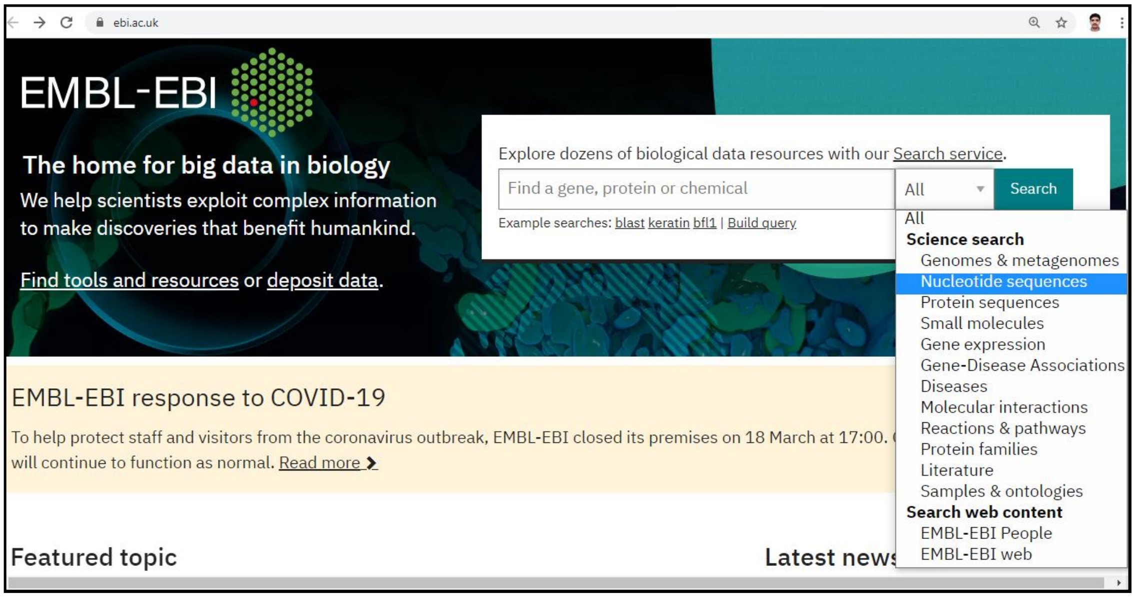
**Fig. 4: Nucleotide sequence data retrieved in GenBank (A) and FASTA (B) file formats.**

### b. EMBL: European Molecular Biology Laboratory

Similar to GenBank, the EMBL database (http://www.ebi.ac.uk/embl/index.html) maintained at European Bioinformatics Institute (EBI) is part of the European Nucleotide Archive (ENA) aimed at constructing a comprehensive catalog of the world's nucleotide sequencing information. (Fig 5). EMBL exchanges the new and updated data on daily basis with GenBank and DDBJ [5]. Primary data is accepted from individual researchers/laboratories and large sequencing centers such as Sanger Centre etc. All types of sequencing datasets such as gene, genomes, ESTs, transcriptome and WGS, etc. are accepted. All the submissions are assigned unique and permanent accession numbers. Protein translations from CDS and annotation are also part of the EMBL database. The database has been integrated with Sequence Retrieval System (SRS) at EBI for easy search key-based online retrieval of sequence data by the scientific community.

### c. DNA Data Bank of Japan (DDBJ)

DDBJ (https://www.ddbj.nig.ac.jp/) is the only nucleotide sequence database located in Asia. It was established in 1986 by the Center for Information Biology (CIB) under the National Institute of Genetics (NIG) of Japan in collaboration with NCBI in the USA and EBI in Europe. As part of the International Nucleotide Sequence Database (INSD; http://www.insdc.org) consortium, these three databases play a crucial role in ensuring the integrity of information shared and submitted (Fig. 6). DDBJ also contains protein sequences and structures along with nucleotide/genomic resources [6]. Approximately 75% of direct nucleotide submission data to DDBJ comes from the Japanese researchers followed by Iran (4.7%) and India (3.7%). Similar to other databases, DDBJ also has a simple GUI enabling access to sequence submission, analysis, search, FTP download, and supercomputing facility (Fig. 6). However, access to the supercomputing facility is restricted to researchers from Japan only. In comparison to GenBank and ENA, the share of DDBJ in overall sequence submission is relatively less (Fig 7 A & B).



The home for big data in biology

We help scientists exploit complex information to make discoveries that benefit humankind.

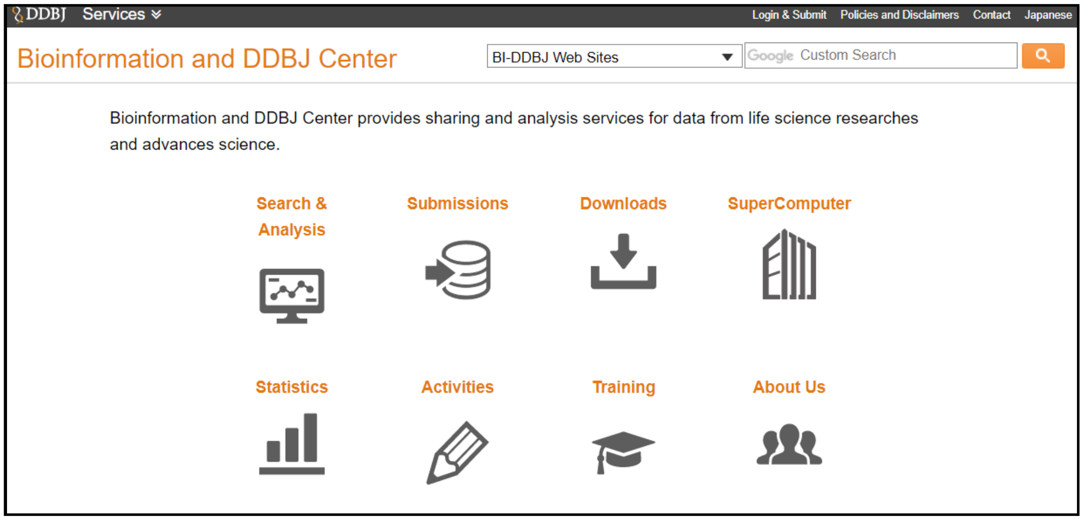
Find tools and resources or deposit data.

Explore dozens of biological data resources with our Search service.

EMBL-EBI response to COVID-19

To help protect staff and visitors from the coronavirus outbreak, EMBL-EBI closed its premises on 18 March at 17:00. ( will continue to function as normal. Read more ❯

**Fig. 5: EMBL interface showing the search box and access to various database types**



Bioinformation and DDBJ Center

Bioinformation and DDBJ Center provides sharing and analysis services for data from life science researches and advances science.

**Fig. 6: DDBJ interface showing access to various sequence submission, search, analysis, ftp download and supercomputing facility options.**

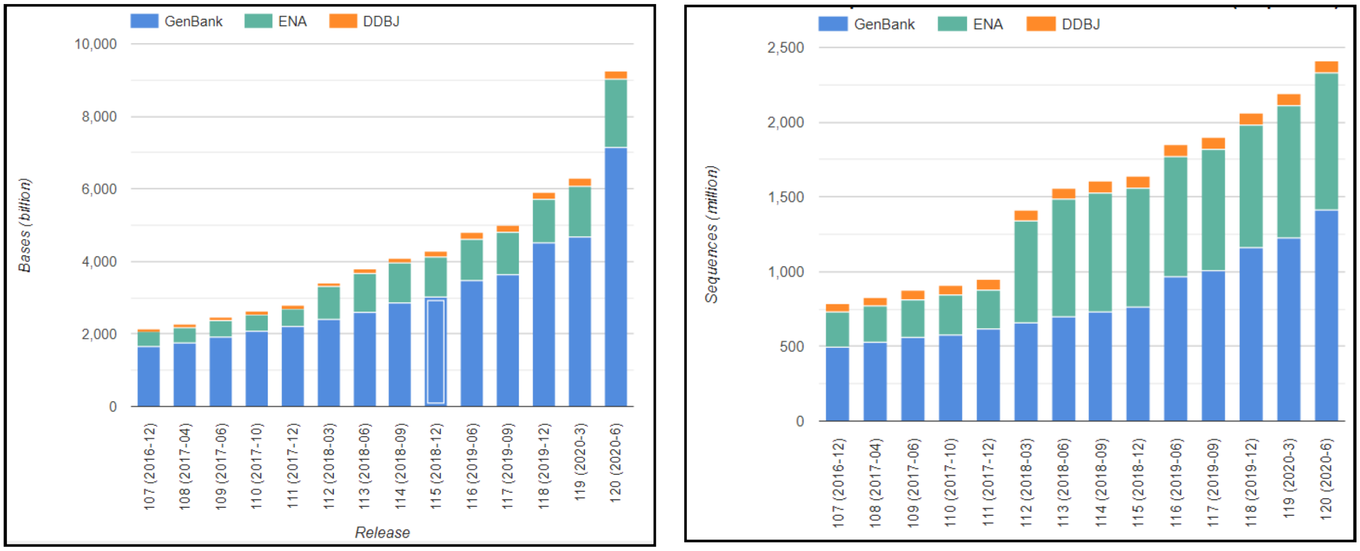


**Fig. 7: Data submission statistics to GenBank, ENA, and DDBJ in terms of bases (A) and sequences (B)**

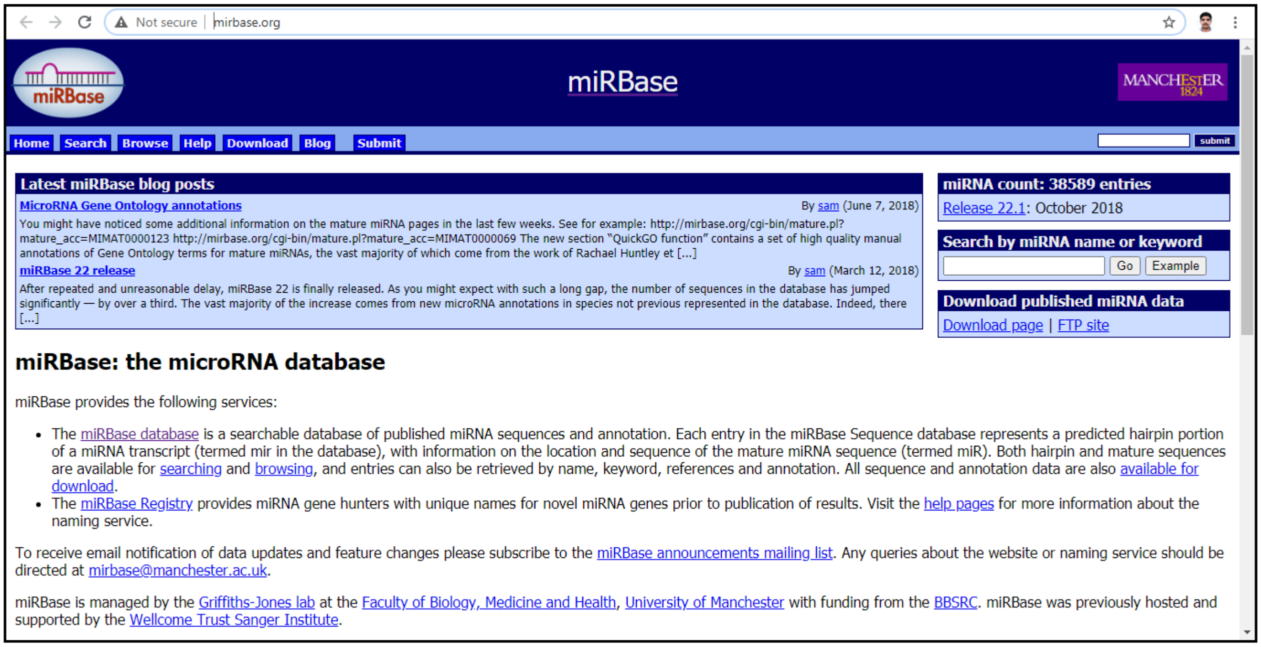
### d. miRBase

miRBase (http://www.mirbase.org/) is the online primary database of microRNA sequences along with their annotation (Fig. 8). It was established in 2002 as a microRNA Registry and later renamed as miRBase. This database is responsible for assigning the gene names to novel microRNAs. Each entry in the miRBase sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references, and annotation. All sequence and annotation data are also available for download. Sequences submitted by researchers are the primary source of data in miRBase. Due to rapid advancements in next-generation sequencing technologies, large numbers of sequences in miRBase come from NGS-based small RNA sequencing experiments [7]. The latest v22.1 of miRBase, released during October 2018, contained 38,589 entries.

## B. Protein databases

### a. Protein databank (PDB)

PDB is the globally accessible primary database of protein structures. It contains a three-dimensional crystallographic structure of protein molecules. In contrast to its name, PDB also stores the structures of other molecules of biological significance such as fragments of nucleic acids, RNA molecules, large peptides, and protein-nucleic acid complexes. The data constituting the PDB is mainly derived from NMR, X-ray crystallography, and molecular modeling. The Protein Data Bank (PDB) was established at Brookhaven National Laboratories (BNL) in 1971 as an archive for biological macromolecular crystal structures. In October 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) took over the management responsibility of the PDB (https://www.rcsb.org/). Data processing at PDB consists of several steps starting from data deposition, annotation, and validation to distribution (Fig. 9). All primary data collected at PDB is assigned a unique PDB identity.



**Latest miRBase blog posts**

MicroRNA Gene Ontology annotations — By sam (June 7, 2018)

You might have noticed some additional information on the mature miRNA pages in the last few weeks. See for example: http://mirbase.org/cgi-bin/mature.pl?mature_acc=MIMAT0000123 http://mirbase.org/cgi-bin/mature.pl?mature_acc=MIMAT0000069 The new section "QuickGO function" contains a set of high quality manual annotations of Gene Ontology terms for mature miRNAs, the vast majority of which come from the work of Rachael Huntley et […]

miRBase 22 release — By sam (March 12, 2018)

After repeated and unreasonable delay, miRBase 22 is finally released. As you might expect with such a long gap, the number of sequences in the database has jumped significantly — by over a third. The vast majority of the increase comes from new microRNA annotations in species not previous represented in the database. Indeed, there […]

**miRNA count: 38589 entries**

Release 22.1: October 2018

**Search by miRNA name or keyword**

**Download published miRNA data**

Download page | FTP site

**miRBase: the microRNA database**

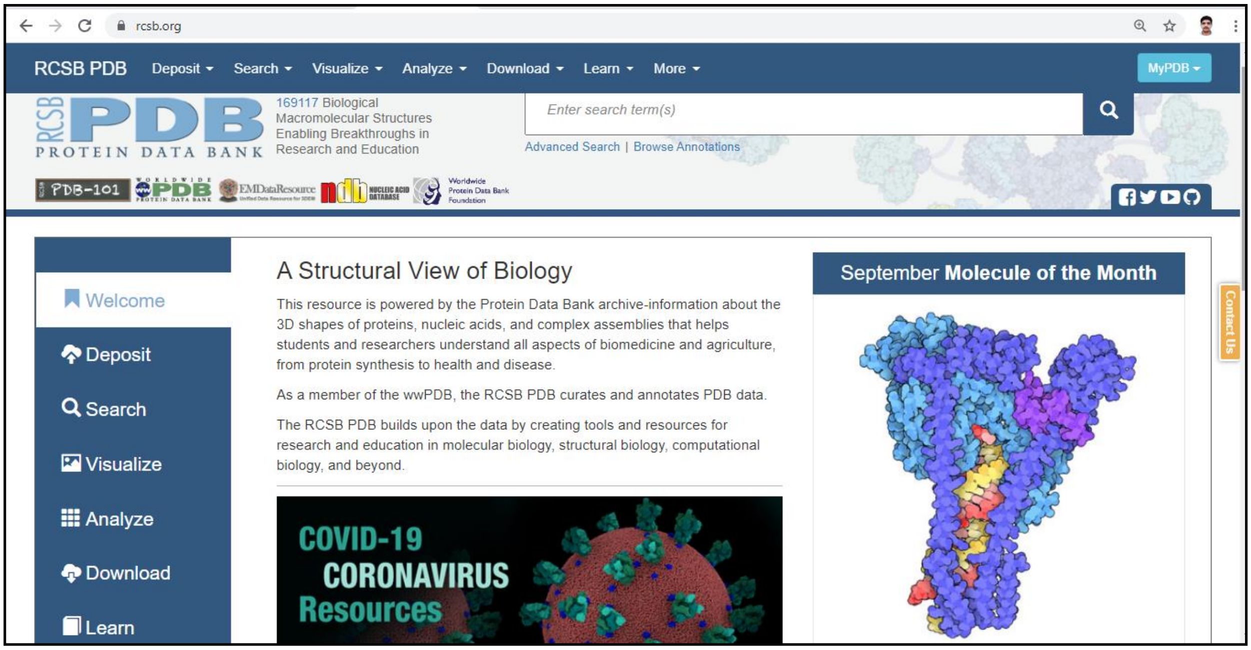
miRBase provides the following services:

- The miRBase database is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download.
- The miRBase Registry provides miRNA gene hunters with unique names for novel miRNA genes prior to publication of results. Visit the help pages for more information about the naming service.

To receive email notification of data updates and feature changes please subscribe to the miRBase announcements mailing list. Any queries about the website or naming service should be directed at mirbase@manchester.ac.uk.

miRBase is managed by the Griffiths-Jones lab at the Faculty of Biology, Medicine and Health, University of Manchester with funding from the BBSRC. miRBase was previously hosted and supported by the Wellcome Trust Sanger Institute.

**Fig. 8: miRBase interface at http://www.mirbase.org/ showing various functionalities**



169117 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

**A Structural View of Biology**

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

**Fig. 9: PDB interface at https://www.rcsb.org/**

### b. Protein Information Resource (PIR)

The Protein Information Resource (PIR) (https://proteininformationresource.org/) is the largest, most comprehensive, annotated protein sequence database in the public domain. The PIR-International Protein Sequence Database (PIR-PSD), works in collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Sequence Database (JIPID). PIR was established in 1984 by the National Biomedical Research Foundation (NBRF), USA. It is a non-redundant database with expert annotations. Protein sequences in the database are classified based on the motifs, homology domains, and superfamily concept. These classification approaches result in a more detailed understanding of sequence-function and structure relationships. The information contained in PIR, Swiss-Prot, and TrEMBL has also been joined to create a central repository and a freely accessible Universal Protein Resource (UniProt) database of protein sequences and functions.

### c. Swiss-Prot

Swiss-Prot is a highly curated protein sequence database with manual annotations and reviews. Swiss-Prot was created at the Department of Medical Biochemistry of the University of Geneva, and it also closely collaborate with EMBL. Each entry in Swiss-Prot is stored as core data and annotation. In the core data, the protein sequence is stored in single amino acid codes along with taxonomy and related references, whereas annotations provide information about protein function, domain structure as well as post-translational protein modifications.

## Conclusion

Biological databases have been playing a crucial role in modern scientific research by acting as a store of rapidly growing datasets and information. These databases also allow easy and rapid sharing of data within the scientific community leading to knowledge and discovery. Large numbers of nucleotide, protein, and specialized databases are presently available. Due to space constraints and very high efforts required, the description of each database is not possible here. However, detailed literature about these databases is now available, and it can be explored as per the interest of the individual researcher.

## References:

1. Cannataro M, Guzzi PH, Tradigo G, Veltri P (2014) Biological databases. In: Kasabov N (ed) Springer Handbook of Bio-/Neuroinformatics. Springer Handbooks. Springer, Berlin, Heidelberg. , Berlin, Heidelberg.
2. Baxevanis AD, Bateman A (2015) The Importance of Biological Databases in Biological Discovery. Curr Protoc Bioinformatics 50:1 1 1-8

3. Zou D, Ma L, Yu J, Zhang Z (2015) Biological databases for human research. Genomics Proteomics Bioinformatics 13:55-63
4. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. Nucleic Acids Res 41:D36-42
5. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Hoad G, Kanz C, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Nardone F, Pastor MP, Plaister S, Sobhany S, Stoehr P, Vaughan R, Wu D, Zhu W, Apweiler R (2007) EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Res 35:D16-20
6. Dobay A, Dobay MP (2013) DDBJ Genome Resources. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H (eds) Encyclopedia of Systems Biology. Springer New York, New York, NY, pp 548-550
7. Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to function. Nucleic Acids Res 47:D155-D162
8. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535-542
9. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235-242
10. Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, Xiao C, Yeh LS, Ledley RS, Janda JF, Pfeiffer F, Mewes HW, Tsugita A, Wu C (2000) The protein information resource (PIR). Nucleic Acids Res 28:41-44
11. Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45-48