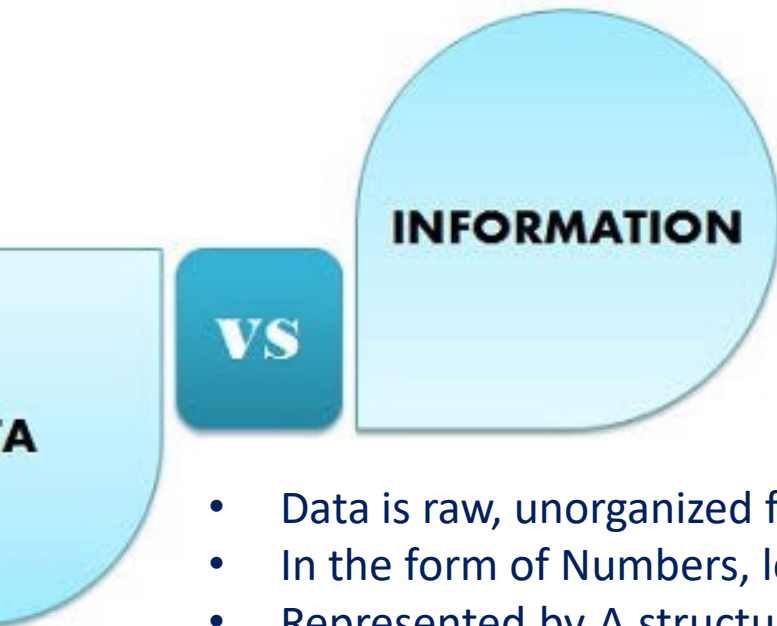


BIOLOGICAL DATABASES



- Data is raw, unorganized facts that need to be processed.
- In the form of Numbers, letters, or a set of characters.
- Represented by A structure, such as tabular data, data tree, a data graph, etc.
- Describes qualitative or Quantitative variables that can be used to make ideas or conclusions



Data

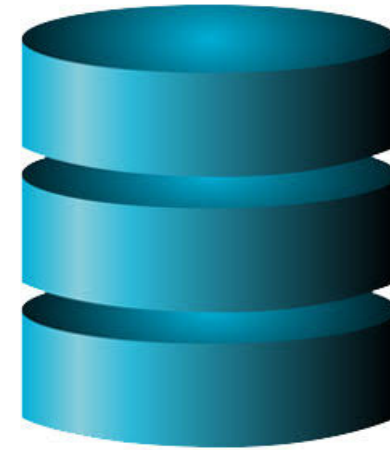


- When data is processed, organized, structured or presented in a given context so as to make it useful, it is called information.
- Collected in the form of ideas and inferences.
- Describes a group of data which carries news and meaning.
- Represented in Language, ideas, and thoughts based on the data.

What is database??



A database is a collection of data in an organized manner, which is accessible in various ways.




Biological databases

- **Biological databases** are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis.
- They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetic.
- Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

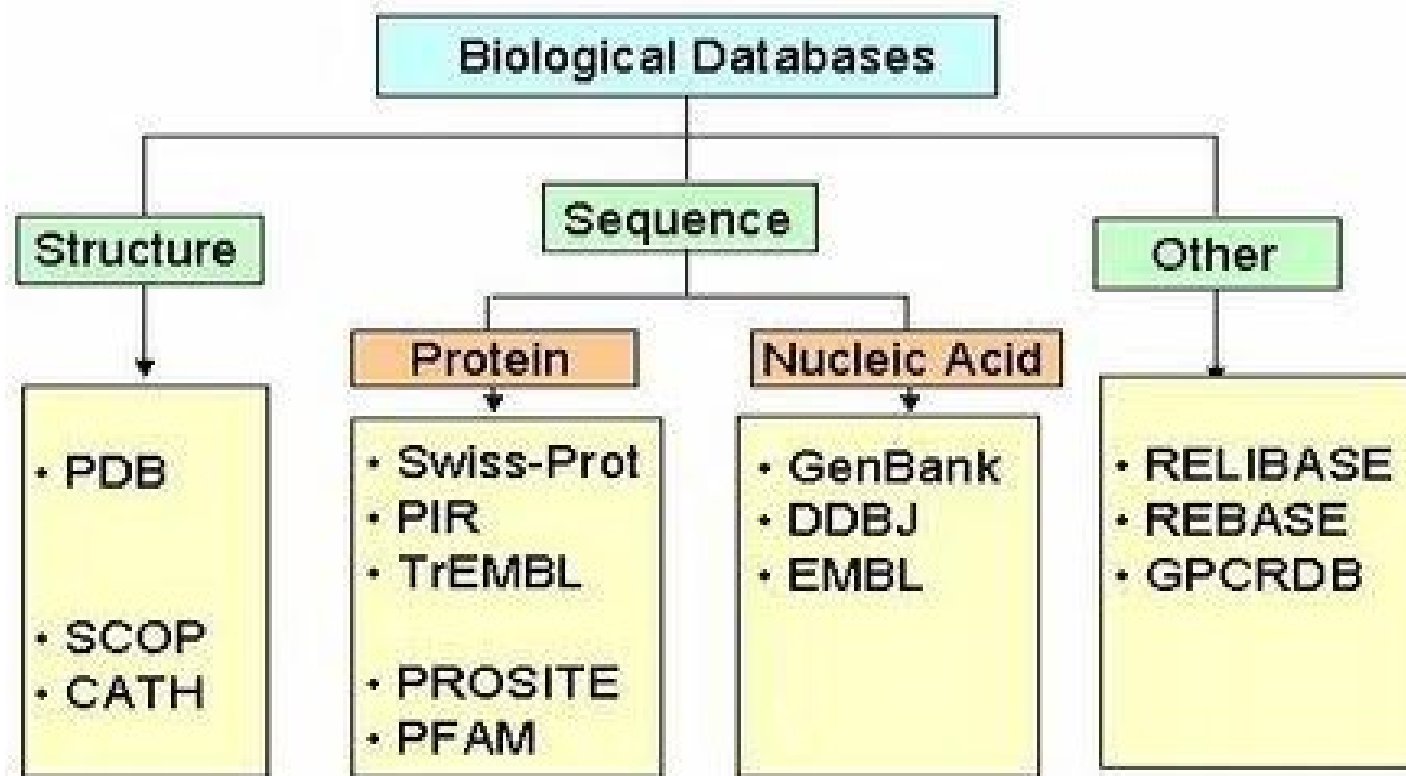


Objectives of Biological databases

- ❖ Recognize various data formats, and know what their primary use.
 - ❖ Know, understand and utilize all types of sequence identifiers.
 - ❖ Know and understand various feature types present in the GenBank flat files.
 - ❖ Know and understand the various GenBank divisions.
- 

Classification of Biological databases

Biological databases can be broadly classified into sequence, structure and functional databases.



Provide a computational support and a user-friendly interface to a researcher for a meaningful analysis of biological data.

Types of Biological databases

Primary (archival)

- ✓ Original submissions by experimentalists
- ✓ Content controlled by the submitter

Ex: GenBank/EMBL/DDBJ

UniProt

PDB

Medline (PubMed)

Secondary (curated)

- ✓ Derived from primary data.
- ✓ Content controlled by third party

Ex: RefSeq

Taxon

UniProt

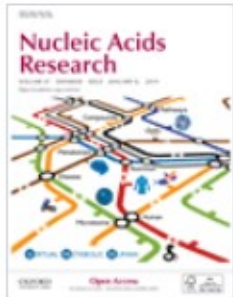
OMIM

Composite/specialized

- ✓ Focused on a particular research interest/organism
- ✓ Usually highly curated
- ✓ Combination of both primary and secondary information

Ex: EST, STS, SNP database, TomatoGDB, TAIR etc.

NAR articles on Biological databases



Volume 47, Issue D1

08 January 2019

[Cover image](#)

ISSN 0305-1048

EISSN 1362-4962

Editorial

[Front Matter](#)

Nucleic acid sequence,
structure, and regulation

Volume 47, Issue D1, 08 January 2019

Database issue

Page 1 of 2

[1](#) [2](#) [Next](#)

EDITORIAL

The 26th annual Nucleic Acids Research database issue and
Molecular Biology Database Collection 

Daniel J Rigden, Xosé M Fernández

Nucleic Acids Res, Volume 47, Issue D1, 08 January 2019, Pages D1–D7,

<https://doi.org/10.1093/nar/gky1267>

[Abstract ▼](#) [View article](#)

An important resource for finding biological databases is a special yearly issue of the journal *Nucleic Acids Research* (NAR).

A companion database of NAR- Online Molecular Biology Database Collection lists 1,380 online databases

Sequence databases

Primary DNA databases

DDBJ/EMBL/GenBank

Primary protein databases

GenPept/TrEMBL

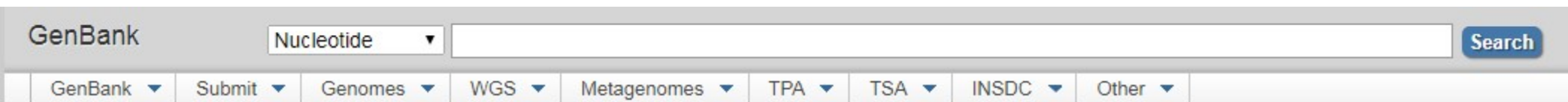
Curated DB

RefSeq (Genomic, mRNA and protein)

Swiss-Prot & PIR -> UniProt (protein)

GenBank

- ✓ **GenBank** is the NIH genetic sequence database of all publicly available DNA and derived protein sequences, with annotations describing the biological information.



GenBank Nucleotide Search

GenBank ▾ Submit ▾ Genomes ▾ WGS ▾ Metagenomes ▾ TPA ▾ TSA ▾ INSDC ▾ Other ▾

GenBank Overview

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). 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 a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

GenBank Resources

- [GenBank Home](#)
- [Submission Types](#)
- [Submission Tools](#)
- [Search GenBank](#)
- [Update GenBank Records](#)

EMBL-EBI

EMBL-EBI

The home for big data in biology

25 Years of EMBL-EBI

In Focus: Exploring innovation in the life sciences >

Our unique Search service helps you explore dozens of biological data resources.
More about EBI Search >

Find a tool for your data analysis.
Find a tool >

Share your scientific data with the world.
Deposit data >

Find a gene, protein or chemical

Example searches: blast keratin bf1

We are EMBL-EBI

Data resources

Research

The **EMBL-EBI** is a hub for **bioinformatics** research and services, developing and maintaining a large number of scientific databases, which are free of charge.

<https://www.ebi.ac.uk/>

DDBJ

The **DNA Data Bank of Japan (DDBJ)** is a [biological database](#) that collects DNA sequences. It is located at the [National Institute of Genetics](#) (NIG) in the [Shizuoka prefecture](#) of Japan.

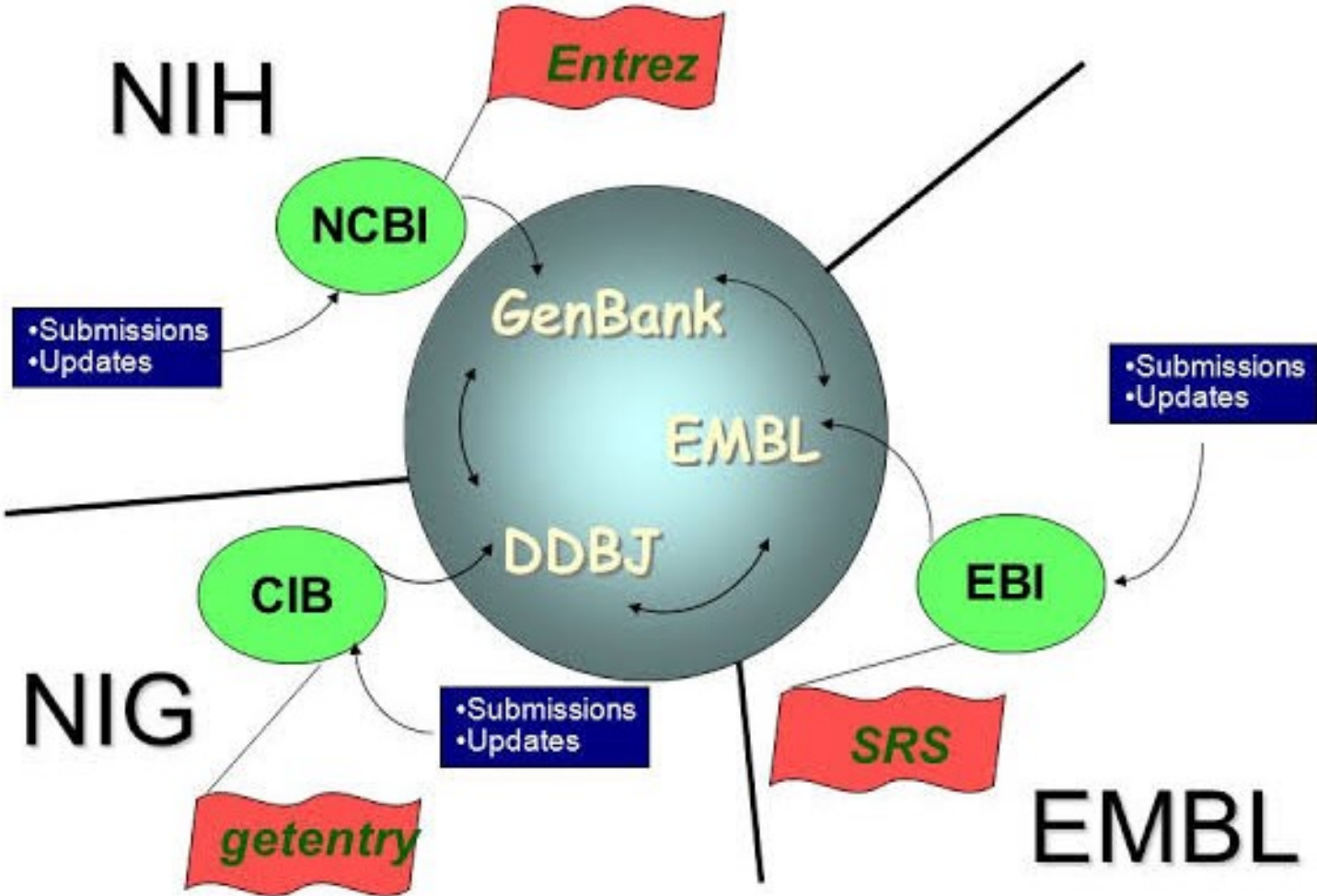


The screenshot shows the DDBJ Services page with a dark header bar containing the DDBJ logo, 'Services' with a dropdown arrow, and navigation links for 'Login & Submit', 'Policies and Disclaimers', 'Contact', and 'Japanese'. The main content area is divided into three columns:

- Search** (magnifying glass icon):
 - getentry
 - ARSA
 - DRA Search
 - TXSearch
 - BLAST
- Analysis** (wrench icon):
 - Vector Screening System
 - ClustalW
 - WABI (Web API for Biology)
 - DDBJ FTP Site
- Databases** (database icon):
 - Annotated/Assembled Sequences (DDBJ)
 - Sequence Read Archive (DRA)
 - Genomic Expression Archive (GEA)
 - BioProject
 - BioSample
 - Japanese Genotype-phenotype Archive (JGA)
 - Submission portal D-way
- NIG SuperComputer** (server rack icon):
 - NIG SuperComputer
- DBCLS Services** (DBCLS logo icon):
 - AOE
 - CRISPRdirect
 - DBCLS SRA
 - Gendoo
 - GGGenome
 - GGRNA
 - RefEx

<https://www.ddbj.nig.ac.jp/index-e.html>

International Nucleotide Sequence Database Collaboration INSDC



GenBank Flat File (GBFF)

```
LOCUS       AJ251330             1653 bp    mRNA    linear    PLN 15-APR-2005
DEFINITION  Oryza sativa mRNA for MAPK4 protein (mapk4 gene).
ACCESSION   AJ251330
VERSION     AJ251330.1
KEYWORDS    mapk4 gene; MAPK4 protein.
SOURCE      Oryza sativa (rice)
ORGANISM    Oryza sativa
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BOP
            clade; Oryzoideae; Oryzaceae; Oryzinae; Oryza.
REFERENCE   1
AUTHORS     Huang,H.J., Fu,S.F., Tai,Y.H., Huang,D.D. and Kuo,T.T.
TITLE       Molecular cloning and expression of a MAP kinase homologue from
            rice
JOURNAL     Unpublished
REFERENCE   2 (bases 1 to 1653)
AUTHORS     Huang,H.J.
TITLE       Direct Submission
JOURNAL     Submitted (26-NOV-1999) Huang H.J., Department of Biology,
            Cheng-Kung University, 1, University Rd., Tainan 701, TAIWAN

FEATURES             Location/Qualifiers
     source           1..1653
                    /organism="Oryza sativa"
                    /mol_type="mRNA"
                    /db_xref="taxon:4530"
     gene             1..1653
                    /gene="mapk4"
     CDS              225..1334
                    /gene="mapk4"
                    /function="protein kinase"
                    /codon_start=1
                    /product="MAPK4 protein"
                    /protein_id="CAB61889.1"
                    /db_xref="GOA:Q5Z859"
                    /db_xref="InterPro:IPR000719"
                    /db_xref="InterPro:IPR002290"
                    /db_xref="InterPro:IPR003527"
                    /db_xref="InterPro:IPR008271"
                    /db_xref="InterPro:IPR011009"
                    /db_xref="InterPro:IPR017441"
                    /db_xref="InterPro:IPR017442"
                    /db_xref="UniProtKB/Swiss-Prot:Q5Z859"
                    /translation="MVMVDPPIKSGKSHYTHWQTLFEIDTKVVPKPIGRGAYG
                    IVCSSINRATNEKVAIKKINNVFDRVDAIIRLRLKLLRHLRHEHVIKLDIIMPVH
                    RRSFKDYYLVYELHDTDLHQIISKSQPLSNDHCQYFLFQLLRGLKYLHSAGILHRDLK
                    PGNLLVNAACDLKICDFGLARTNNTKGFMT EYVTRHYRAPELLCCDNVGTSIDVW
                    SVGCIFAELLGRKPIFPPTGTECLNQLKLIIVNLGTMSEADIEFIDNPKARKYIKTLPTV
                    PGILPITSMYPQAHPLAIDLKQMLKLVFDPKSRISVTEALEHPYMSPLYDPSANPPVQPV
                    IDLDIDENLGVDMIREMIMQEMLHYHPEVVAQVNM"
```

Header

Title, Taxonomy,
Citation

Features

No. of bases, amin
acids, ORF start
position, end
position and
encoded protein
sequences

Sequence

```
ORIGIN
1  gcccgctcag gtcagctcag ctcttcaaaa taggtgggag gtgctctcgc ctccccctcc
61  cggcgcttc cgctccaga tcgcgcgcgc ccgcccgcg cg-gctcgcct cgagcggag
121  ctatagctgc cgcttcgat tcgacgctgc atgtactggc gagaggcgta cgctggcctg
181  gctctcccgc acgcgccct cgctctctcc cgttcatta caagatggtg atgatggtag
241  accctcctaa tggcatggga aaccaaggga agcattacta cacaatgtgg caaacgctat
301  ttgagattga caccaaagt gtgccaatca agcccatcgg aagaggggct tatggaatag
361  tttgctcctc tataaacctg gcaaccaacg aaaaagttag aataaagaag atcaacaact
421  tctttgacaa ccgtgtggat gcactgagga ccctaagggg gctgaaacta ctgcgacact
481  tgcgccatga aaatgttatt gccttgaaag atataatgat gccagtacac agaaggagtt
541  tcaaagatgt atacttgggt tatgaactca tggacacaga tctacaccag ataattaat
601  catctcaacc tctttctaag gaccactgtc aatatttctt ttttcagcta ctccgaggct
```