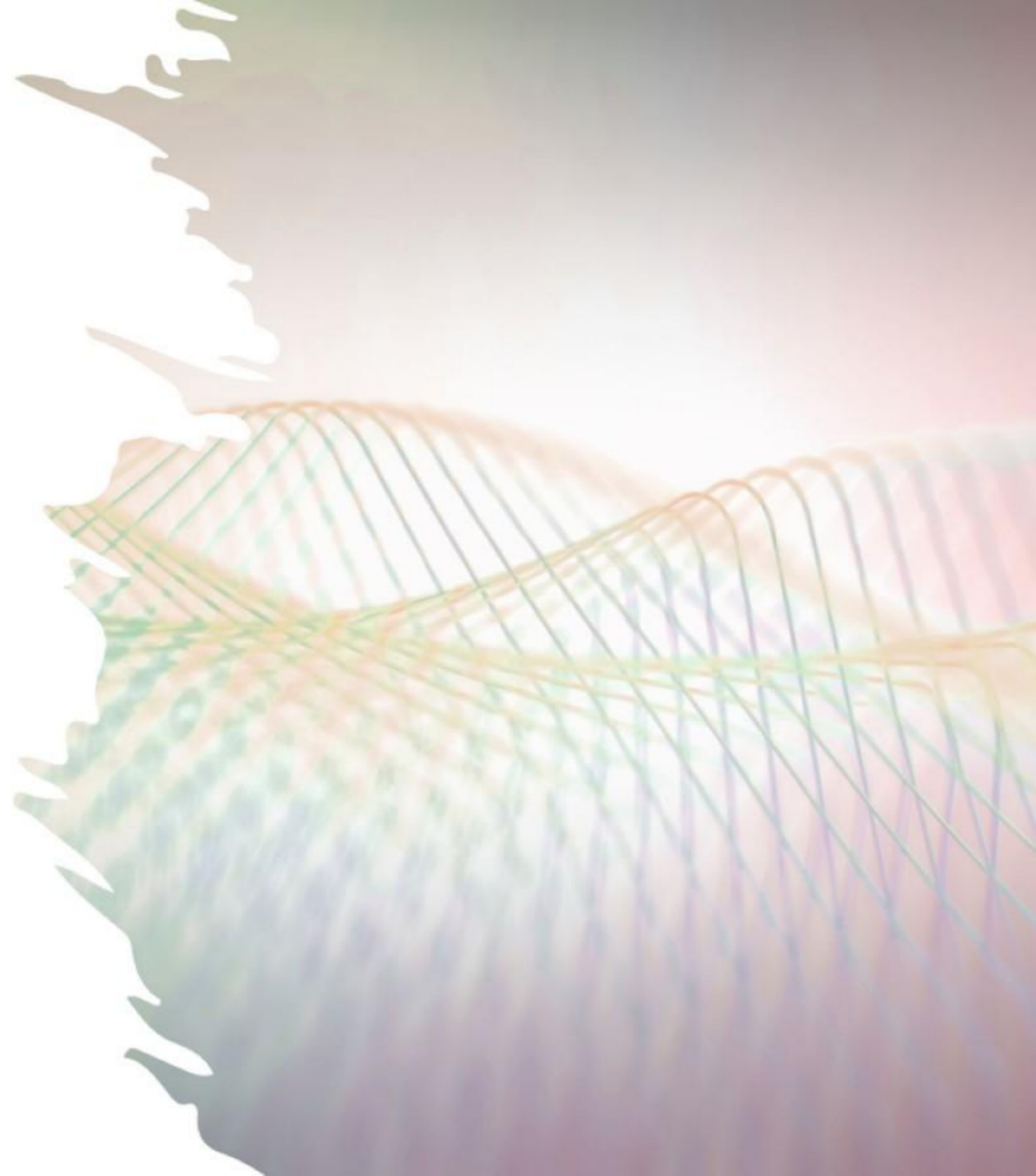


Bioinformatics/ Lec 2

Exploring NCBI: A Guide to Using Biological Databases and Tools

By

Dr Delveen R. Ibrahim



There are several **Nucleotide Sequence Database** that are widely used for bioinformatics analysis. These databases contain information about DNA and RNA sequences, their structures, functions, and related annotations. Here are some prominent nucleic acid databases:

- **GenBank:**

- Run by the National Center for Biotechnology Information (NCBI).
- A big collection of shared DNA and RNA sequences.
- Includes data on genes, genomes, and proteins..

- **European Molecular Biology Library- European Nucleotide Archives (EMBL_ENA):**

- Collaborates with GenBank and the DNA Data Bank of Japan (DDBJ) to share nucleotide sequence data globally.
- Covers a wide range of organisms and sequence types.

❑ **DNA Data Bank of Japan (DDBJ):**

- ❑ Operated by the National Institute of Genetics in Japan.
- ❑ Collaborates with GenBank and EMBL to ensure global data sharing.
- ❑ Like GenBank, it contains nucleotide sequences.

❑ **RefSeq (Reference Sequence):**

- ❑ Maintained by the NCBI.
- ❑ A curated collection of reference sequences for a variety of organisms.
- ❑ Provides annotations and links to related information.

❑ **RNAcentral:**

- ❑ A comprehensive resource for non-coding RNA sequences run by EMBL.
- ❑ Integrates data from various specialized RNA databases.

And many others.....

ncbi.nlm.nih.gov

Google Translate | Gmail - Google Sea... | SABIS® Digital Plat... | Facebook | Syllabus for Microbi... | Email - pppsdam... | My Account | STUDYBLUE | Find a... | Download Movies...

An official website of the United States government. [Learn how you know.](#)

NIH National Library of Medicine
National Center for Biotechnology Information

dsilvern.brahim@u...

All Databases Search

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

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

NCBI Hidden Markov models (HMM) release 11.0 now available

Go to Settings to activate Windows

Release 11.0 of the NCBI protein profile Hidden Markov models (HMM) is available

ddbj.nig.ac.jp/index-e.html

Google Translate | Gmail - Google Sea... | SABIS® Digital Plat... | Facebook | Syllabus for Microbi... | Email - pppsdam... | My Account | STUDYBLUE | Find a... | Download Movies...

DDBJ Services SuperComputer Statistics Activities About Us

DDBJ Web Sites Terms Contact Japanese

Suspension of the BI-DDBJ activity during the New Year Holidays

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

Services
Search, analysis, database services of DDBJ Center

Submission
Navigation for how to submit your data

Super Computer
NIG Supercomputer

Statistics
Statistics of DDBJ Center services

Activities
Training sessions and achievements of DDBJ Center

About us
About Bioinformatics and DDBJ Center

NEWS

DDBJ is selected as Global Core Biodata Resources
2022/12/16 [Announcement](#) [DDBJ Center](#)

DDBJ staff visit to KOBIC in Korea
2022/12/15 [Announcement](#) [DDBJ Center](#)

Read length and direction of paired reads were made optional in the DRA submission
2022/12/12 [Announcement](#) [DRA](#) [DDBJ Center](#)

Suspension of the BI-DDBJ activity during the New Year Holidays
2022/12/12 [Announcement](#) [DDBJ](#) [BioProject](#) [BioSample](#) [DRA](#) [DEA](#) [JGA](#) [AGG](#) [DDBJ Center](#)

Updated tools related to Mass Submission System (MSS)
2022/12/09 [Announcement](#) [DDBJ](#) [DDBJ Center](#)

Activate Windows
Go to Settings to activate Windows.

European Molecular Biology Lab: X

embl.org

Google Translate | Gmail - Google Sea... | SABIS® Digital Plat... | Facebook | Syllabus for Microbi... | Email - pppsdam... | My Account | STUDYBLUE | Find a... | Download Movies...

EMBL-EBI Services Research Training About us

EMBL-EBI

EMBL's European Bioinformatics Institute

EMBL-EBI

Unleashing the potential of big data in biology

Find a gene, protein or chemical All Search

Example searches: blast keratin bf11 | About EBI Search

Directory Se

The three primary databases in the International Nucleotide Sequence Database Collaboration (**INSDC**) are GenBank, DDBJ, and EMBL-EBI

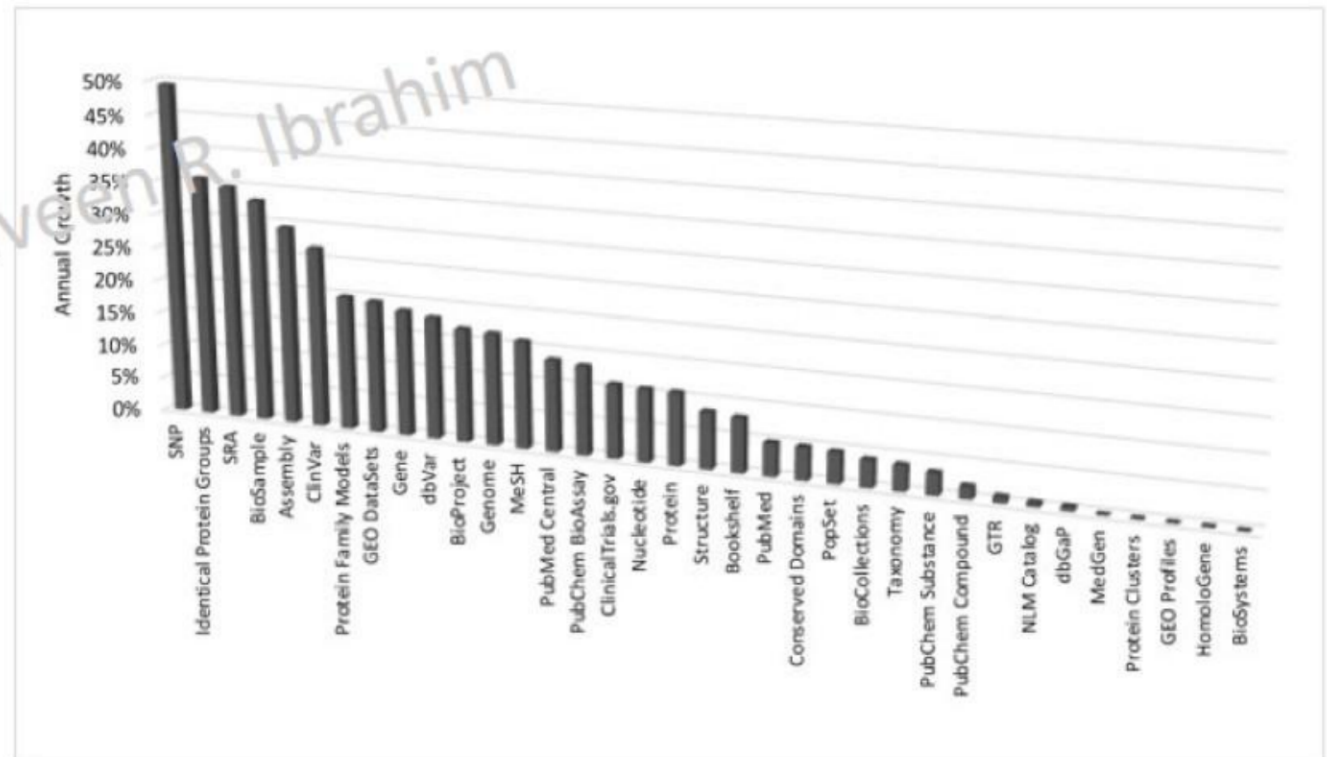
Repository	Full Name	Region	Maintained By
DDBJ	DNA Data Bank of Japan	Asia	National Institute of Genetics (Japan)
GenBank	Genetic Sequence Database	Americas	National Center for Biotechnology Information (NCBI, USA)
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute	Europe	EMBL-EBI, United Kingdom

Bioinformatics Tools:

1. **Sequence Analysis Tools:** Used to analyze DNA, RNA, and protein sequences. Examples include **BLAST** (Basic Local Alignment Search Tool) for sequence similarity searches.
2. **Structural Analysis Tools:** Used to predict and analyze the three-dimensional structures of biological molecules. Examples include tools for protein structure prediction. (AlphaFold2).
3. **Functional Analysis Tools:** Help in understanding the biological function of genes and proteins. **Gene Ontology** (GO) is commonly used for functional annotation.
4. **Pathway Analysis Tools:** Examine the interactions and relationships between biological molecules in pathways. **KEGG** (Kyoto Encyclopedia of Genes and Genomes) is a well-known pathway database.

NCBI (National Center for Biotechnology Information)

- As of October 2024, the (NCBI) maintains a diverse set of 31 repositories and knowledgebases containing approximately 4.6 billion records. Most of these records are accessible through the Entrez retrieval system at <https://www.ncbi.nlm.nih.gov/search/>

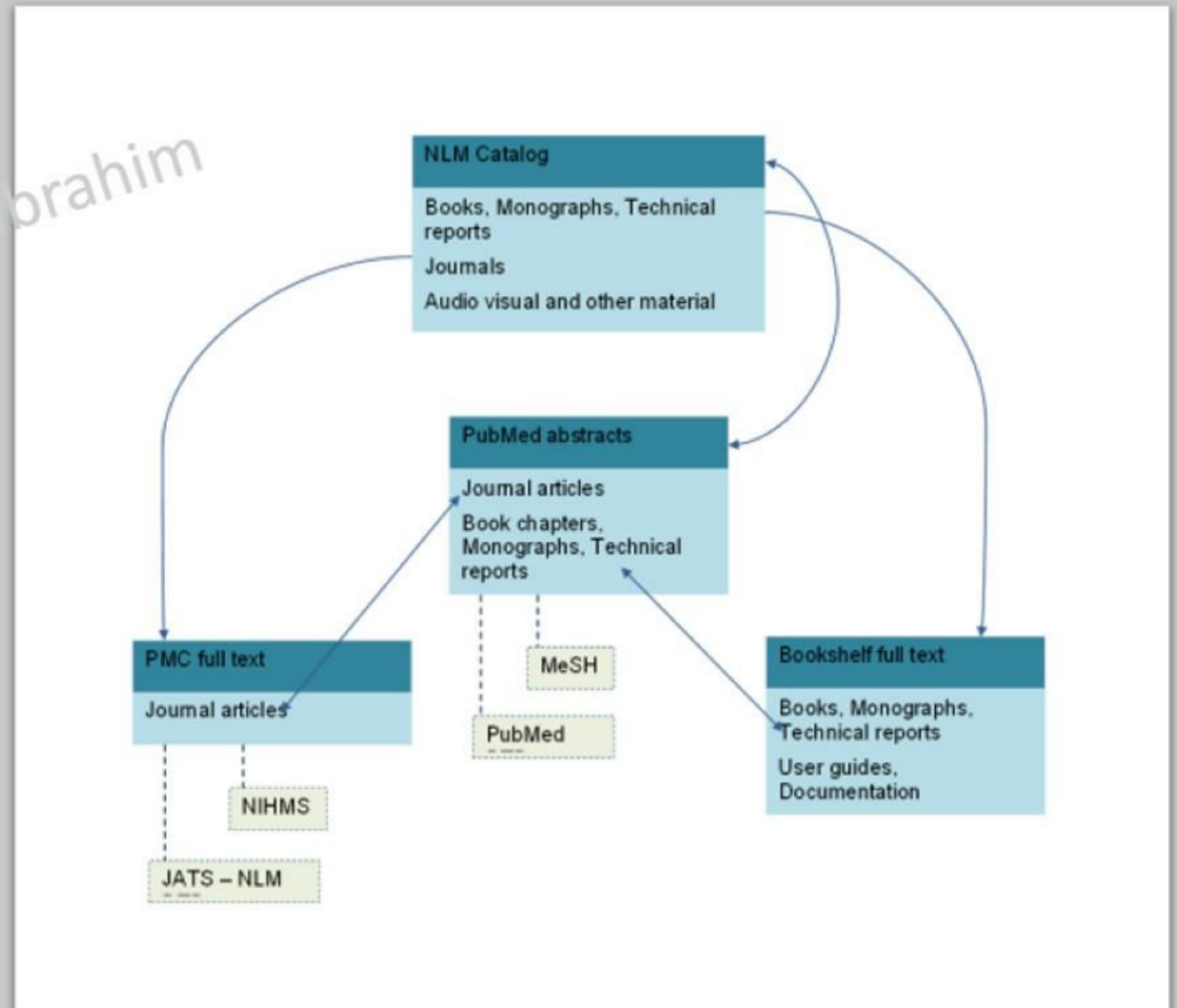


Annual growth rates of the number of records in each NCBI database as of 4 September 2021.

- The NCBI database (<http://www.ncbi.nlm.nih.gov/>) is a collection of gene, genomic sequence, transcript and proteins
- **NCBI mainly include the following databases:**
 - **Literature :** ex, PubMed & PMC
 - **Genome: ex,** Nucleotide & Taxonomy
 - **Proteins: ex,** Protein and Protein Clusters
 - **Genes: ex,** HomoloGene and Gene
 - **Chemicals: ex,** PubChem Substance
- Besides NCBI includes many analysis tools such as **BLAST , Prime-BLAST and ORF-finder.**

Overview of NCBI Literature Resources:

- Such as PubMed and PMC , ex: how to use keywords and different filters to obtain your desired journal paper related to your graduation project



NCBI Entrez Molecular Sequence Database System

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI
SITE MAP

Introduction

Entrez is a molecular biology database system that provides integrated access to nucleotide and protein sequence data, gene-centered and genomic map data, PubMed MEDLINE, and more. It is maintained by the National Center for Biotechnology Information and is available via the Internet.

Entrez Databases and Resources

ncbi.nlm.nih.gov

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine

National Center for Biotechnology Information

delveen.ibrahim@u...

All Databases ▾

- All Databases
- Assembly
- Biocollections
- BioProject
- BioSample
- Books
- ClinVar
- Conserved Domains
- dbGaP
- dbVar
- Gene
- Genome
- GEO DataSets
- GEO Profiles
- GTR
- HomoloGene
- Identical Protein Groups
- MedGen
- MeSH
- NLM Catalog

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

NCBI

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

[Home](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Download

Learn

Develop

Analyze

Research

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

NCBI hidden Markov models (HMM) release 11.0 now available

➤ **Some of the File format used in Bioinformatics:**

- Raw data/ plain text
- GenBank format
- FASTA format
- FASTQ
- EMBL
- SAM/BAM
- VCF
- GFF
- ASN.1

Dr Delveen R. Ibrahim

Why Is There Such A Variety of Types?

Raw data/ Plain Sequence format

No number in this kinds of file format.

Contain one Sequence, while most other formates accepts several sequences in one file.

Examples:

ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGGCCACGGCCACCGCTGC
CCTGCCCCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAA
TAAGGAAAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGC
CGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCC
CAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAAGTTCTTCTGGAAGACCTTCTCCTCCT
GCAAATAAAACCTCACCCATGAATGCTCACGCAAGTTTAATTACAGACCTGAA

➤ FASTA format:

- Can contain several sequences.
- Each sequence in FASTA format begins with a single-line description, followed by lines of sequence data.
- The description line must begin with a greater-than (">") symbol in the first column .

Examples:

symbol of >

Sequence

Description

```
>NZ_AP022255.1:23069-23944 Aeromonas caviae strain WP8-S18-ESBL-04 plasmid pWP8-S18-ESBL-04_1, complete sequence
ATGGTTAAAAAATCACTGCGTCAGTTACGCTGATGGCGACGGCAACCGTCACGCTGTTGTTAGGAAGTG
TGCCGCTGTATGCGCAAACGGCGGACGTACAGCAAAAACCTTGCCGAATTAGAGCGGCAGTCGGGAGGAAG
ACTGGGGGTGGCATTGATTAACACAGCGGATAATTCGCAAATACTTTATCGTGCTGATGAGCGCTTCGCG
ATGTGCAGCACCAGTAAAGTGATGGCCGCGGCCGCGGTGCTGAAGAAAAGTGAAAGCGAACCGAATCTGT
TAAATCAGCGAGTTGAGATCAAAAAATCTGACCTGGTTAACTATAATCCGATTGCGGAAAAGCACGTCAA
TGGGACGATGTCACTGGCTGAGCTTAGCGCGGCCGCGCTACAGTACAGCGATAACGTGGCGATGAATAAG
CTGATTGCTCACGTTGGCGGCCCGGCTAGCGTCACCGCGTTTCGCCGACAGCTGGGAGACGAAACGTTCC
GTCTCGACCGTACCGAGCCGACGTTAAACACCGCCATTTCGGGCGATCCGCGTGATACCACTTCACTCG
```

➤ Genbank format:

- Can contain several sequences.
- The first section includes the entry's LOCUS, DEFINITION, ACCESSION and VERSION- followed by many others annotations.
- This end by " **ORIGIN**" which is the start of the sequence, and the end of the sequence is marked by two slashes ("//").
- The Genbank format allows for the storage of information in addition to a DNA/protein sequence. It holds much more information than the FASTA format

➤ Genbank format:

Examples:

LOCUS NZ_AP022255 876 bp DNA linear CON 16-DEC-2022
DEFINITION Aeromonas caviae strain WP8-S18-ESBL-04 plasmid pWP8-S18-ESBL-04_1,
complete sequence.
ACCESSION [NZ_AP022255](#) REGION: 23069..23944
VERSION NZ_AP022255.1
DBLINK BioProject: [PRJNA224116](#)
BioSample: [SAM00194626](#)
Sequence Read Archive: [DRR199462](#)
Assembly: [GCF_014169735.1](#)
KEYWORDS RefSeq.
SOURCE Aeromonas caviae (Aeromonas punctata)
ORGANISM [Aeromonas caviae](#)

ORIGIN

```
1 atggttaaaa aatcactgcg tcagttcacg ctgatggcga cggcaaccgt cacgctgttg
61 ttaggaagtg tgccgctgta tgcgcaaacg gcggacgtac agcaaaaact tgccgaatta
121 gagcggcagt cgggagggaag actgggggtg gcattgatta acacagcggg taattcgcaa
181 atactttatc gtgctgatga gcgcttcgcg atgtgcagca ccagtaaagt gatggccgcg
241 gccgcgggtg tgaagaaaag tgaagcggaa ccgaatctgt taaatcagcg agttgagatc
301 aaaaaatctg acctgggtta ctataatccg attgcggaaa agcacgtcaa tgggacgatg
361 tcactggctg agcttagcgc ggccgcgcta cagtacagcg ataacgtggc gatgaataag
421 ctgattgctc acgttggcgg cccggctagc gtcaccgctg tcgcccgaca gctgggagac
481 gaaacgttcc gtctcgaccg taccgagcgg acgttaaaca ccgccattcc gggcgatccg
541 cgtgatacca cttcacctcg ggcaatggcg caaactctgc ggaatctgac gctgggtaaa
601 gcattgggcg acagccaacg ggcgagctg gtgacatgga tgaaaggcaa taccaccggt
661 gcagcgagca ttcaggcagg actgcctgct tcctgggttg tgggggataa aaccggcagc
721 ggtgactatg gcaccaccaa cgatatcgcg gtgatttggc caaaagatcg tgcgccgctg
781 attctggtca cttacttcac ccagcctcaa cctaaggcag aaagccgtcg cgatgtatta
841 gcgtcggcgg ctaaaatcgt caccgacggt ttgtaa
```

//

GenPept

Send to:

Change region shown

Customize view

GyrA [Mycobacterium tuberculosis]

GenBank: AVV29511.1

[Identical Proteins](#)
[FASTA](#)
[Graphics](#)

Go to:

LOCUS AVV29511 838 aa linear BCT 03-APR-2018

DEFINITION GyrA [Mycobacterium tuberculosis].

ACCESSION AVV29511

VERSION AVV29511.1

DBSOURCE accession [MG995190.1](#)

KEYWORDS

SOURCE Mycobacterium tuberculosis (Mycobacterium tuberculosis variant tuberculosis)

ORGANISM [Mycobacterium tuberculosis](#)

Bacteria; Bacillati; Actinomycetota; Actinomycetes; Mycobacteriales; Mycobacteriaceae; Mycobacterium; Mycobacterium tuberculosis complex.

REFERENCE 1 (residues 1 to 838)

AUTHORS Daum,L.T. and Rodriguez,J.D.

TITLE Next-Generation Sequencing for Characterizing Drug Resistance-Conferring Mycobacterium tuberculosis Genes from Clinical Isolates in the Ukraine

JOURNAL J. Clin. Microbiol. (2018) In press

Avast Secure Browser

ORIGIN

```

1 mtdttlppdd sldriepvdi qqemqrsyid yamsvivgra lpevrldgkpv hrrvlyamf
61 dsgrfprdrsh aksarsvaet mgnypghgdv siydtlvrra qpwslyrplv dgagnfgspg
121 ndppaamryt earltplame mlreideetv dfipnydgrv qeptvlpsrf pnllangsgg
181 iavgmatnip phnlrelada vfwalenhda deeetlaavm grvkgpdfpt aglivgsqgt
241 adayktgrgs imrgvveve edsngtrslv itelpyqvnv dnfitsiaeq vrdgklagis
301 niedqssdrv glrivieikr davakvinn lykhtqlqts fganmlaivd gvprtlrdq
361 liryvvdhql dvivrtrtyr lrkanerahi lrglvkalda ldevialira setvdiaarg
421 lielldidei qaqailldmql rrlaalerqr iiddlakiea eiadledila kperqgivr
481 delaeivdrh gddrrtriia adgdvsdedl iaredvvvti tetgyakrtk tdlrsqkrq
541 gkgvqgaglk qddivahffv csthdlilff ttqgrvyrak aydlpeasrt argqhvanll
601 afqpeeriaq viqirgytda pylvlatrng lvkkskltdf dsnrsggiva vnlrdndelv
661 gavlcsaddl lllvsangqs irfsatdeal rpmgratsgv qgmrfniddr llslnvvreg
721 tyllvatsgg yakrtaieey pvqgrgkgv ltvmydrnrg rlvgalivdd dselyavtsg
781 ggvirtaarq vrkagrtkqg vrlmnlgedd tllaiarnae esgddnavda ngadtqtn

```

//

GenPept is a database format that provides protein sequence records derived from nucleotide coding sequences (CDS) in the GenBank database

Exploring NCBI:
Practical section
Lec 3

1. First open <https://www.ncbi.nlm.nih.gov> ,
you will see the following webpage

ncbi.nlm.nih.gov

Translate gmail - Google Sea... SABIS® Digital Plat... Facebook Syllabus for Microbi... Email - pppsardam-... My Account Download Movies... YouTube Merge

An official website of the United States government [Here's how you know](#)

 **National Library of Medicine**
National Center for Biotechnology Information

delveen.ibrahim@u...

All Databases

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

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



Popular Resources

[PubMed](#)

[Bookshelf](#)

[PubMed Central](#)

[BLAST](#)

[Nucleotide](#)

[Genome](#)

[SNP](#)

[Gene](#)

[Protein](#)

[PubChem](#)

NCBI News & Blog

RefSeq Release 216

RefSeq release 216 is now available online, from the FTP site, and through NCBI's Entrez programming utilities. F...

2. For literature you can use PMC and PubMed as I explained for you before

You can use different filters

ncbi.nlm.nih.gov/pmc/?term=toxoplasma+gondii

Translate Gmail - Google Sea... SABIS® Digital Plat... Facebook Syllabus for Microbi... Email - pppsardam-... My Account Download Movies... YouTube Merge Video

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information

delveen.ibrahim@u...

PMC US National Library of Medicine National Institutes of Health

Search **toxoplasma gondii**

Create alert Journal List Advanced Help

Display Settings: Summary, 20 per page, Sorted by Default order

Send to: Filter your results:

All (38174)

[PubMed links \(36973\)](#)

[Open Access \(23689\)](#)

Embargoed (0)

[Manage Filters](#)

Article attributes

Associated Data

Author manuscripts

Digitized back issues

MEDLINE journals

Open access

Preprints

Retracted

Text availability

Include embargoed articles

Publication date

1 year

5 years

10 years

Custom range...

Research Funder

NIH

AHRQ

ACL

ASPR

CDC

You searched for **2 million full text articles**

Try this search in **34+ million citations and abstracts**

PMC Full-Text Search Results

Items: 1 to 20 of 38174

<< First < Prev Page 1 of 1909 Next > Last >>

☐ [Unifying Virulence Evaluation in *Toxoplasma gondii*: A Timely Task](#)

1. Rafael Calero-Bernal, Mercedes Fernández-Escobar, Frank Katzer, Chunlei Su, Luis Miguel Ortega-Mora
Front Cell Infect Microbiol. 2022; 12: 868727. Published online 2022 Apr 28. doi: 10.3389/fcimb.2022.868727
PMCID: PMC9097680
[Article](#) [PubReader](#) [PDF-4.0M](#) [Cite](#)

☐ [Toxoplasma gondii infection and toxoplasmosis in farm animals: Risk factors and economic impact](#)

2. S. Stelzer, W. Basso, J. Benavides Silván, L.M. Ortega-Mora, P. Maksimov, J. Gethmann, F.J. Conraths, G. Schares
Food Waterborne Parasitol. 2019 Jun; 15: e00037. Published online 2019 Apr 3. doi: 10.1016/j.fawpar.2019.e00037

PMC Images search for *toxoplasma gondii*

Activate Windows
Go to Settings to activate Windows

3. Taxonomy, for defining the exact taxonomy for an organism you can use the taxonomy section from NCBI , you can find it on left upper side, under all Databases

The screenshot shows the NCBI website interface. At the top, the NIH logo and 'National Library of Medicine' text are visible, along with a user login 'delveen.ibrahim@u...'. Below the header, a search bar and a dropdown menu labeled 'All Databases' are present. The dropdown menu is open, showing a list of databases including MeSH, NLM Catalog, Nucleotide, OMIM, PMC, PopSet, Protein, Protein Clusters, Protein Family Models, PubChem BioAssay, PubChem Compound, PubChem Substance, PubMed, SNP, SRA, Structure, **Taxonomy** (highlighted), ToolKit, ToolKitAll, and ToolKitBooksh. On the left side, a vertical navigation menu lists various resources like 'NCBI Home', 'Resource List (A-Z)', 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The main content area features several sections: 'NCBI' with a description, 'Submit' for manuscripts and bases, 'Download' for transferring data, 'Learn' for help documents and tutorials, 'Develop' for using NCBI APIs, 'Analyze' for identifying tools, and 'Research' for exploring research projects. On the right side, there are sections for 'Popular Resources' (PubMed, Bookshelf, PubMed Central, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem) and 'NCBI News & Blog' (RefSeq Release 216, Activate Windows).

Taxonomy Taxonomy Aspergillus niger Search

Create alert Limits Advanced

Display Settings: Summary

Aspergillus niger
species, ascomycete fungi
[Nucleotide](#) [Protein](#)

Send to:

Related information

Nucleotide

Protein

Assembly

ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=5061



Taxonomy
Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy BioCollections

Search for as complete name lock Go Clear

Display 3 levels using filter: none

☐ Nucleotide ☐ Protein ☐ Structure ☐ Genome ☐ Popset ☐ SNP ☐ Conserved Domains ☐ GEO Datasets ☐ PubMed Central
☐ Gene ☐ HomoloGene ☐ SRA Experiments ☐ LinkOut ☐ BLAST ☐ GEO Profiles ☐ Protein Clusters ☐ Identical Protein Groups ☐ SPARCLE
☐ BioProject ☐ BioSample ☐ Assembly ☐ dbVar ☐ Genetic Testing Registry ☐ Host ☐ Viral Host ☐ PubChem BioAssay

Lineage (full): [cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Fungi](#); [Dikarya](#); [Ascomycota](#); [saccharomyceta](#); [Pezizomycotina](#); [leotiomyceta](#); [Eurotiomycetes](#); [Eurotiomycetidae](#); [Eurotiales](#); [Aspergillaceae](#); [Aspergillus](#); [Aspergillus subgen. Circumdati](#)

o **Aspergillus niger** Click on organism name to get more information.

- Aspergillus niger ATCC 1015
- Aspergillus niger ATCC 13496

4. Gene (transcript) sequence retrieval from NCBI database:

- On left upper side select gene. Type name of gene or gene symbol and organism name. For example, CFTR (CF transmembrane conductance regulator) from Homo sapiens

The screenshot displays the NCBI Gene database interface for the CFTR gene. The 'Gene' dropdown menu is open, showing a list of categories including 'Recent', 'Gene', 'PMC', 'All Databases', 'Assembly', 'Biocollections', 'BioProject', 'BioSample', 'Books', 'ClinVar', 'Conserved Domains', 'dbGaP', 'dbVar', 'Gene', 'Genome', 'GEO DataSets', and 'GEO Profiles'. The main content area shows the gene name 'CFTR (homo sapiens)', a description 'transmembrane conductance regulator', and various links for data retrieval like 'RefSeq proteins', 'RefSeqGene', and 'PubMed'. The right sidebar shows 'Results by taxon' and 'Find related data' sections.

- Click on CFTR gene in the window.

Search results

Items: 1 to 20 of 1094

<< First < Prev Page 1 of 55 Next > Last >>

Name/Gene ID	Description	Location	Aliases	MIM
☐ CFTR ID: 1080	CF transmembrane conductance regulator [<i>Homo sapiens</i> (human)]	Chromosome 7, NC_000007.14 (117480025..117668665)	ABC35, ABCC7, CF/MRP, MRP7, TNR-CFTR, dJ760C5.1, CFTR	602421
☐ ABCC2 ID: 1244	ATP binding cassette subfamily C member 2 [<i>Homo sapiens</i> (human)]	Chromosome 10, NC_000010.11 (99782640..99852594)	ABC30, CMOAT, DJS, MRP2, cMRP	601107
☐ ABCC8 ID: 6833	ATP binding cassette subfamily C member 8 [<i>Homo sapiens</i> (human)]	Chromosome 11, NC_000011.10 (17392498..17476845, complement)	ABC36, HHF1, HI, HRINS, MRP8, PHHI, PNDM3, SUR, SUR1, SUR1delta2, TNDM2	600509
☐ ABCC4 ID: 10257	ATP binding cassette subfamily C member 4 [<i>Homo sapiens</i> (human)]	Chromosome 13, NC_000013.11 (95019835..95301451, complement)	MOAT-B, MOATB, MRP4	605250
☐ ABCC1 ID: 4363	ATP binding cassette subfamily C member 1 [<i>Homo sapiens</i> (human)]	Chromosome 16, NC_000016.10 (15949143..16143053)	ABC29, ABCC, DFNA77, GS-X, MRP, MRP1	158343



CFTR CF transmembrane conductance regulator [*Homo sapiens* (human)]

[Download Datasets](#)

Gene ID: 1080, updated on 14-Jan-2023

Summary

Official Symbol	CFTR provided by HGNC
Official Full Name	CF transmembrane conductance regulator provided by HGNC
Primary source	HGNC:HGNC:1884
See related	Ensembl:ENSG00000001626 MIM:602421 ; AllianceGenome:HGNC:1884
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	CF; MRP7; ABC35; ABCC7; CFTR/MRP; TNR-CFTR; dJ760C5.1
Summary	This gene encodes a member of the ATP-binding cassette (ABC) transporter superfamily. The encoded protein functions as a chloride

- In the window go down in the mRNA and proteins section. Now click on NM_000492.4 to get nucleotide sequence for the mRNA and NP_000483.3 for protein sequence.

mRNA and Protein(s)

1 [NM_000492.4](#) – [NP_000483.3](#) cystic fibrosis transmembrane conductance regulator

[See identical proteins and their annotated locations for NP_000483.3](#)

Status: REVIEWED

Source sequence(s)	AC000061 , AC000111 , M28668
Consensus CDS	CCDS5773.1
UniProtKB/Swiss-Prot	P13569 , Q2I102
UniProtKB/TrEMBL	A0A024R730
Related	ENSP00000003084.6 , ENST00000003084.11

Conserved Domains (5) [summary](#)

cd03289 Location:1208 → 1480	ABCC_CFTR2; ATP-binding cassette domain 2 of CFTR, subfamily C
cd03291 Location:389 → 670	ABCC_CFTR1; ATP-binding cassette domain of the cystic fibrosis transmembrane regulator, subfamily C
TIGR01271 Location:1 → 1480	CFTR_protein; cystic fibrosis transmembrane conductor regulator (CFTR)
pfam00664 Location:866 → 1147	ABC_membrane; ABC transporter transmembrane region

- After clicking on NM_000492.4 the following window will open.

GenBank ▾

Homo sapiens CF transmembrane conductance regulator (CFTR), mRNA

NCBI Reference Sequence: NM_000492.4

[FASTA](#) [Graphics](#)

[Go to](#) 

LOCUS NM_000492 6070 bp mRNA linear PRI 28-DEC-2022
 DEFINITION Homo sapiens CF transmembrane conductance regulator (CFTR), mRNA.
 ACCESSION NM_000492
 VERSION NM_000492.4
 KEYWORDS RefSeq; MANE Select.
 SOURCE Homo sapiens (human)
 ORGANISM [Homo sapiens](#)
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Eute
 Mammalia; Eutheria; Euarchontoglires; Primates; Haplorh
 Catarrhini; Hominidae; Homo.
 REFERENCE 1 (bases 1 to 6070)
 AUTHORS Zhang G, Li Y, Wang Y, Wang B, He Y, Chen C, Shi G and

FASTA ▾

Homo sapiens CF transmembrane conductance regulator (CFTR), mRNA

NCBI Reference Sequence: NM_000492.4

[GenBank](#) [Graphics](#)

```
>NM_000492.4 Homo sapiens CF transmembrane conductance regulator (CFTR), mRNA
GTAGTAGGTCCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCAGCGCCCGAGAGACC
ATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGTTGTCTCCAACTTTTTTCAGCTGGACCAGACCAATTT
TGAGGAAAGGATACAGACAGCGCCTGGAATTGTCTAGACATATACCAAATCCCTTCTGTTGATTCTGCTGA
CAATCTATCTGAAAAATTGGAAAGAGAATGGGATAGAGAGCTGGCTTCAAAGAAAAATCCTAACTCATT
AATGCCCTTCGGCGATTGTTTTTCTGGAGATTATGTTCTATGGAATCTTTTATATTTAGGGGAAGTCA
CCAAAGCAGTACAGCCTCTCTTACTGGGAAGAATCATAGCTTCCTATGACCCGGATAACAAGGAGGAACG
CTCTATCGCGATTTATCTAGGCATAGGCTTATGCCTTCTCTTTATTGTGAGGACACTGCTCTACACCCA
GCCATTTTTGGCCTTCATCACATTGGAATGCAGATGAGAATAGCTATGTTTAGTTTGATTATAAGAAGA
CTTTAAAGCTGTCAAGCCGTGTTCTAGATAAAATAAGTATTGGACAACCTGTTAGTCTCCTTTCCAACAA
CCTGAACAAATTTGATGAAGGACTTGCAATTGGCACATTTCTGTGGATCGCTCCTTTGCAAGTGGCACTC
CTCATGGGGCTAATCTGGGAGTTGTTACAGGCGTCTGCCTTCTGTGGACTTGGTTTCTGATAGTCCTTG
CCCTTTTTCAGGCTGGGCTAGGGAGAATGATGATGAAGTACAGAGATCAGAGAGCTGGGAAGATCAGTGA
AAGACTTGTGATTACCTCAGAAATGATTGAAAATATCCAATCTGTTAAGGCATACTGCTGGGAAGAAGCA
```

By clicking on FASTA, the sequence of the complete transcript will open. Select the sequence and copy paste it on a word file

- **Identify CDS (Coding sequence),**. Go back in the previous window. Scroll down and click on CDS

FEATURES	Location/Qualifiers
source	1..6070 /organism="Homo sapiens" /mol_type="mRNA" /db_xref="taxon:9606" /chromosome="7" /map="7q31.2"
gene	1..6070 /gene="CFTR" /gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1; MRP7; TNR-CFTR" /note="CF transmembrane conductance regulator" /db_xref="GeneID:1080" /db_xref="HGNC:HGNC:1884" /db_xref="MIM:602421"
exon	1..123 /gene="CFTR" /gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1; MRP7; TNR-CFTR" /inference="alignment:Splign:2.1.0"
misc feature	5..7 /gene="CFTR" /gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1; MRP7; TNR-CFTR" /note="upstream in-frame stop codon"
CDS	71..4513 /gene="CFTR" /gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1;

After clicking on CDS, the CDS region in the complete transcript will be highlighted. The CDS starts with start codon (ATG) and ends with stop codon. The sequence before the CDS is called as 5' UTR and after CDS region is called as 3' UTR (untranslated region).

Now, you can select it and copy then paste on a word document as before.

Like this CDS, 5' UTR and 3' UTR sequence of any gene or transcript can be retrieved for further analysis, cloning and other purposes

/note="major polyA site"

ORIGIN

```
1 gtaglaggtc ttggcatta ggagcttag ccagacggc cctagcagg acccagcgc
61 ccgagagacc atgcagaggt cgcctctga aaagccagc gtgtctcca aactttttt
121 cagctgacc agaccattt tggagaaag atacagacg cgcctgaat tgcagacat
181 ataccaatc ccttctgtt attctctga caatctatc gaaaatttg aaagagaa
241 gtagagag ctgcttcaa agaaaatcc taactcatt aatgccctt gccatgttt
301 ttctgaga ttatgttct atgaatctt ttatatta ggggaagtc caaagcagt
361 acagccttc ttactgaa gaatcatag ttctatgac ccgataaca agaggaacg
421 ttctatcag atttatctg gcatagctt atgctcttc ttatgttga gacactgct
481 tctacacca gccattttt gcccttaca cattagaatg cagatgaga tagctatgt
541 tagttgatt tataaaga ctttaagct gtcaagcgt gtctagata aataagtat
601 tggacaact gtagcttc ttccaaca cctgaaca ttgatgaag gacttcatt
661 gacatttc gtatgacg ctccttga agtgcactc ctatgggct taacttggg
721 ttgttacag gcctgactc tctgtgact tgtttctg atagtcctg ccttttca
781 tctgacta gggagatga tgatgaatg cagatcag agagctgga agatcagta
841 agactttg attacctag aatgattg aatatcaa tctgttaag catactctg
901 gaaagaga atgaaaaa tgaatgaaa cttaagcaa acagaactg aactgactg
961 aaagcagcc tatgtgat actcaatag ctacgcttc ttcttctag gttctttgt
1021 ttgttttt tctgtcttc cctatgact aatcaagg atcatctcc gaaaatttt
1081 taccaccatc tcatctga ttgtctgc catggcgtc actggaat ttccctggc
1141 ttacaaaca tgtatgact ccttggagc aataaaca atacagatt tcttcaaaa
1201 tcaagatat aagcattg aatataact aagactaca gaagtatga tggagaatt
1261 aacagcttc tggagagg gttttggga attatttga aaagcaaac aaacaataa
1321 aatagaaa acttcaatg gtatgacg cctctctc agtaattct cactcttgc
1381 actcctgtc ctgaagata ttaattcaa gatagaaga gacatttgt tgcatttgc
1441 tgaatcact ggaagagca agacttact tctaatgtg attatggag aactggagc
1501 ttcaagagt aaatttggc acagtgaag aatttcatt ttcttctag ttcttggat
1561 tatgctgac accattaa aaatatcat cttagtgtt tctatgatg aatatagat
1621 tgaagctc atcaagcat gccaatga agagcactc tcaagttg cagagaaga
1681 caatatagt cttgagag gtgaatcac actgagtg gttcaacga caagatttc
1741 tttagcaag cagatata aagatctga ttgtatga ttgactctc cttttgata
1801 tctagtatt ttacagaa aagaatttt tgaactgt gtctgaaac tgaatctat
1861 caaaactagg attttgtc ctctaaaat ggaacttta aagaagctg acaaaattt
1921 aattttgat gaagtagca gctatttta tggacattt tcgaactcc aaatttaca
1981 tccagacttt agctcaaac tcatgggtg ttattcttc gaccaattt gtcagaaag
2041 aagaattca atcctaact agacttaca cgtttctc ttagaaggag atgtctctg
2104
```


Number of exons and their sequence can also be studied.

exon
/inference="alignment:Splign:2.1.0"
2728..2978
/gene="CFTR"
/gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1;
MRP7; TNR-CFTR"
/inference="alignment:Splign:2.1.0"
exon
2979..3058
/gene="CFTR"
/gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1;
MRP7; TNR-CFTR"
/inference="alignment:Splign:2.1.0"
exon
3059..3209
/gene="CFTR"
/gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1;
MRP7; TNR-CFTR"
/inference="alignment:Splign:2.1.0"
exon
3210..3437
/gene="CFTR"
/gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1;
MRP7; TNR-CFTR"
/inference="alignment:Splign:2.1.0"
exon
3438..3538
/gene="CFTR"
/gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1;
MRP7; TNR-CFTR"
/inference="alignment:Splign:2.1.0"
exon
3539..3787
/gene="CFTR"
/gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1;
MRP7; TNR-CFTR"
/inference="alignment:Splign:2.1.0"
exon
3788..3943
/gene="CFTR"
/gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1;
MRP7; TNR-CFTR"
/inference="alignment:Splign:2.1.0"

The corresponding sequence will be highlighted when the specific exon is clicked. These exon sequence can be marked in the original sequence pasted on word file

2641 gagcttaatt tttgtgctaa tttgggtgctt agtaattttt ctggcagagg tggctgcttc
2701 tttggttggt ctgtggctcc ttggaacac tcctcttcaa gacaaggga atagtactca
2761 tagtagaaat aacagctatg cagtgtattat caccagcacc agttcgtatt atgtgtttta
2821 catttacgtg ggagtagccg acactttgct tgctatggga ttcttcagag gtctaccact
2881 ggtgcatact ctaatcacag tgtcgaaaat tttacaccac aaaatgttac attctgttct
2941 tcaagcacct atgtcaaccc tcaacacgtt gaaagcaggt gggattctta atagattctc
3001 caaagatata gcaattttgg atgaccttct gcctcttacc atatttgact tcatccagtt
3061 gttattaatt gtgattggag ctatagcagt tgtcgagtt ttacaaccct acatctttgt
3121 tgcaacagtg ccagttagat tggcttttat tatgttgaga gcataattcc tccaaacctc
3181 acagcaactc aaacaactgg aatctgaagg caggagtcca attttctctc atcttgttag
3241 aagcttaaaa ggactatgga cacttcgtgc cttcggacgg cagccttact ttgaaactct
3301 gttccacaaa gctctgaatt tacatactgc caactgggtc ttgtacctgt caaactgtcg
3361 ctggttccaa atgagaatag aaatgatttt tgtcatcttc ttcattgctg ttaccttcat
3421 ttccatttta acaacaggag aaggagaaag aagagttggt attatcctga ctttagccat
3481 gaatatcatg agtacattgc agtgggctgt aaactccagc atagatgtgg atagcttgat
3541 gcgatctgtg agccgagtct ttaagttcat tgacatgcca acagaaggta aacctacca
3601 gtcaacaaa ccatacaaga atggccaact ctgaaagtt atgattattg agaattcaca
3661 cgtgaagaaa gatgacatct ggccctcagg ggcccaaatg actgtcaaat atctcacagc
3721 aaaatacaca gaagggtgaa atgccaatatt agagaacatt tccttctcaa taagtcttgg
3781 ccagagggtg ggccctcttg gaagaactgg atcagggaag agtactttgt tatcagcttt
3841 tttgagacta ctgaacactg aaggagaaat ccagatcgat ggtgtgtctt gggattcaat
3901 aactttgcaa cagtggagga aagcctttgg agtgatacca cagaaagtat ttattttttc
3961 tggaaacatt agaaaaaact tggatcccta tgaacagtgg agtgatcaag aaatatggaa
4021 agttgcagat gaggttgggc tcagatctgt gatagaacag tttcctggga agcttgactt
4081 tgtccttggt gatgggggct gtgtcctaag ccattggccac aagcagttga tgtgcttggc
4141 tagatctgtt ctcatgaagg cgaagatctt gctgcttgat gaacccagtg ctcatttgga
4201 tccagtaaca taccaaataa ttagaagaac tctaaaaaa gcatttggctg attgcacagt
4261 aattctctgt gaacacagga tagaagcaat ctgggaatgc caacaatttt tggatcagga
4321 agagaacaaa gtgcggcagt acgattccat ccagaaactg ctgaacgaga ggagcctctt
4381 ccggcaagcc atcagccctt ccgacagggt gaagctcttt cccaccgga actcaagcaa
4441 gtcaagctc aagccccaga ttgtctctct gaaagaggag acagaagaag aggtgcaaga

How to Use the Gene Database in NCBI

1. Go to the NCBI website: <https://www.ncbi.nlm.nih.gov>



2. Select 'Gene' database from the dropdown or search directly



3. Enter the gene name, symbol, or ID (e.g., BRCA1, TP53)



4. Explore results and click on the desired gene



5. Review the gene page: Summary, Genomic context, RefSeq, Variants



6. Download data in desired format (e.g., FASTA, GenBank)



7. Use tools like BLAST for sequence analysis

Dr Delveen R. Ibrahim

Some of the questions and their answers which is related to biological database and file formats:

What are the common file formats in bioinformatics?

The FASTA file format is one of the most widely used bioinformatics file types. FASTQ is also used broadly due to the widespread adoption of next-generation sequencing. Other common file types include SAM, **BAM**, CRAM, BED, **VCF**, GFF, and GTF, mentioned in this lecture.

What are data types in bioinformatics?

Data types in bioinformatics can be DNA sequences, RNA sequences, amino acid sequences, methylation sequences, three-dimensional protein structures, and more, listed in lecture 2.

Why do we have different sequence file formats?

Scientists are using bioinformatic data for many different purposes, and other file types include different kinds of information concerning a DNA, RNA, or protein sequence. These various file types may be used for compatibility with additional bioinformatics software or storage efficiency.

<https://youtu.be/QLcmEqBayr0>

A Quick Overview of the NCBI website

