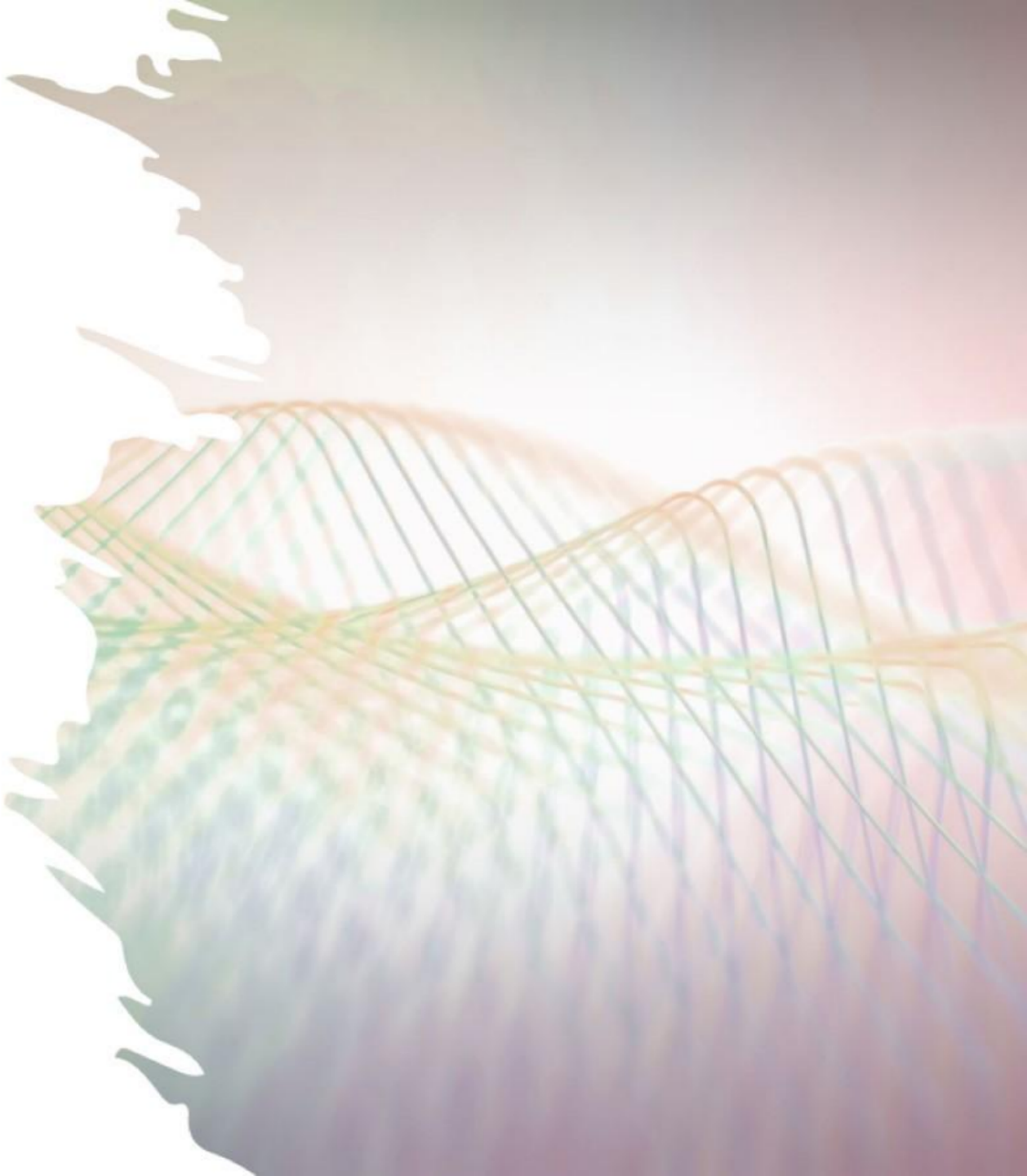
Bioinformatics/ Lec 2

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

An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

All Databases | Search

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
NCBI hidden Markov models (HMM) release 11.0 now available
Release 11.0 of the NCBI protein profile

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

DDBJ Services SuperComputer Statistics Activities About Us | DDBJ Web Sites | Terms Contact

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

DDBJ

Bioinformation 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 Bioinformation and DDBJ Center

NEWS
- DDBJ is selected as Global Core Biodata Resources
- DDBJ staff visit to KOBIC in Korea
- Read length and direction of paired reads were made optional in the DRA submission
- Suspension of the BI-DDBJ activity during the New Year Holidays
- Updated tools related to Mass Submission System (MSS)

more

European Molecular Biology Lab

embl.org

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 bfl1 | About EBI Search

Directory Se

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

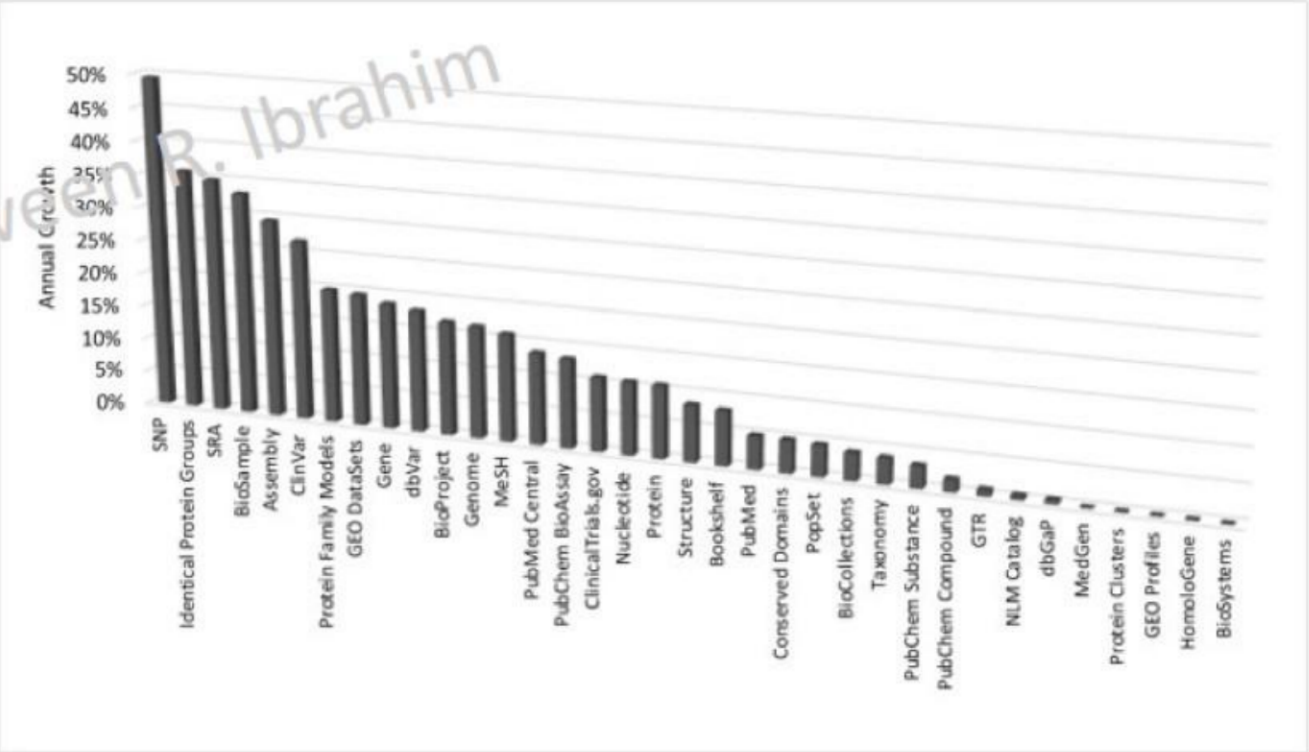
Repository	Full Name	Region	Maintained By
DDBJ	DNA Data Bank of Japan	Asia	National Institute of Genetics (Japan)
GenBank	Genetic Sequence Database	Americas	National Center for Biotechnology Information (NCBI, USA)
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute	Europe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

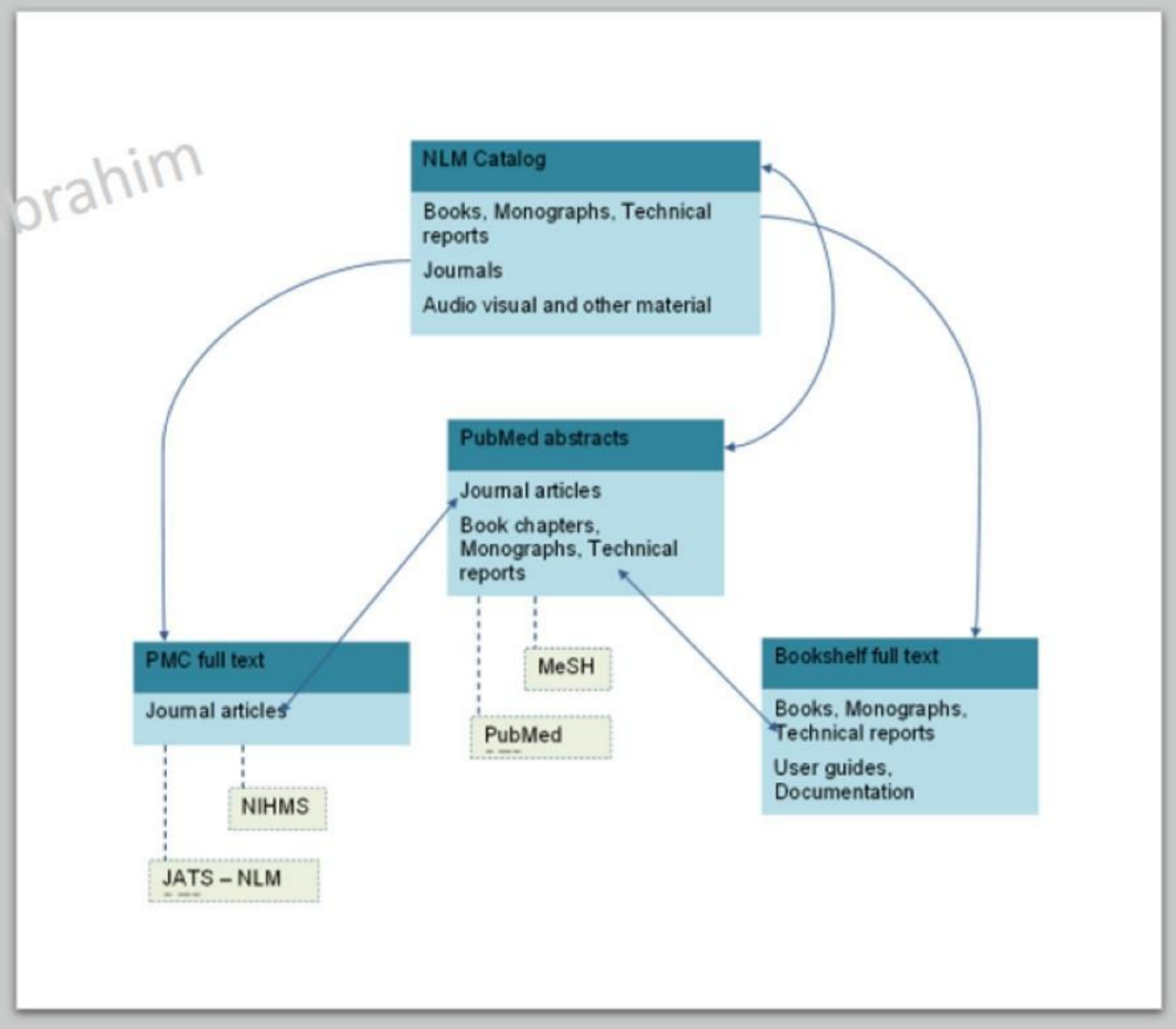
- The NCBI database (http://www.ncbi.nlm.nih.gov/) is a collection of gene, genomic sequence, transcript and proteins
- **NCBI mainly include the following databases:**

  - **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.nlm.nih.gov/Web/Search/entrezfs.html

Google Translate | gmail - Google Sea... | SABIS® Digital Plat... | Facebook | Syllabus for Micr

NCBI

Entrez Molecular Sequence Database System

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

NCBI

SITE MAP

Introduction

Entrez is a molecular biology dat
integrated access to nucleotide a
gene-centered and genomic map
data, PubMed MEDLINE, and mo
by the National Center for Biotec
and is available via the Internet.

Entrez Databases and Re

ncbi.nlm.nih.gov

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

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 | Search

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

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

l | Mission | Organization | NCBI News & Blog

bmit
r manuscripts
bases

Download
Transfer NCBI data to your computer

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

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

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:

ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGC
CCTGCCCCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAA
TAAGGAAAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGC
CGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCC
CAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCT
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 .

symbol of >

Examples:

Sequence

Description

```
>NZ_AP022255.1:23069-23944 Aeromonas caviae strain WP8-S18-ESBL-04 plasmid pWP8-S18-ESBL-
04_1, complete sequence
ATGGTTAAAAAATCACTGCGTCAGTTCACGCTGATGGCGACGGCAACCGTCACGCTGTTGTTAGGAAGTG
TGCCGCTGTATGCGCAAACGGCGGACGTACAGCAAAAACTTGCCGAATTAGAGCGGCAGTCGGGAGGAAG
ACTGGGGGTGGCATTGATTAACACAGCGGATAATTCGCAAATACTTTATCGTGCTGATGAGCGCTTCGCG
ATGTGCAGCACCAGTAAAGTGATGGCCGCGGCCGCGGTGCTGAAGAAAAGTGAAAGCGAACCGAATCTGT
TAAATCAGCGAGTTGAGATCAAAAAATCTGACCTGGTTAACTATAATCCGATTGCGGAAAAGCACGTCAA
TGGGACGATGTCACTGGCTGAGCTTAGCGCGGCCGCGCTACAGTACAGCGATAACGTGGCGATGAATAAG
CTGATTGCTCACGTTGGCGGCCCGGCTAGCGTCACCGCGTTCGCCCGACAGCTGGGAGACGAAACGTTCC
GTCTCGACCGTACCGAGCCGACGTTAAACACCGCCATTCCGGGCGATCCGCGTGATACCACTTCACCTCG
```

➢ 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 **" ORIGION" 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: SAMD00194626
            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 cgggaggaag actgggggtg gcattgatta acacagcgga taattcgcaa
      181 atactttatc gtgctgatga gcgcttcgcg atgtgcagca ccagtaaagt gatggccgcg
      241 gccgcggtgc tgaagaaaag tgaaagcgaa ccgaatctgt taaatcagcg agttgagatc
      301 aaaaaatctg acctggttaa ctataatccg attgcggaaa agcacgtcaa tgggacgatg
      361 tcactggctg agcttagcgc ggccgcgcta cagtacagcg ataacgtggc gatgaataag
      421 ctgattgctc acgttggcgg cccggctagc gtcaccgcgt tcgcccgaca gctgggagac
      481 gaaacgttcc gtctcgaccg taccgagccg acgttaaaca ccgccattcc gggcgatccg
      541 cgtgatacca cttcacctcg ggcaatggcg caaactctgc ggaatctgac gctgggtaaa
      601 gcattgggcg acagccaacg ggcgcagctg 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
//
```

National Library of Medicine
National Center for Biotechnology Information

delveen.ibrahim@u...

Protein | Protein | Search
Advanced | Help

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 lpevrdglkp vhrrvlyamf
     61 dsgfrpdrsh aksarsvaet mgnyhphgdv siydtlvrma qpwslryplv dgqgnfgspg
    121 ndppaamryt earltplame mlreideetv dfipnydgrv qeptvlpsrf pnllangsgg
    181 iavgmatnip phnlrelada vfwalenhda deeetlaavm grvkgpdfpt aglivgsqgt
    241 adayktgrgs irmrgvveve edsrgrtslv itelpyqvnh dnfitsiaeq vrdgklagis
    301 niedqssdrv glrivieikr davakvvinn lykhtqlqts fganmlaivd gvprtlrldq
    361 liryyvdhql dvivrrttyr lrkanerahi lrglvkalda ldevialira setvdiarag
    421 lielldidei qaqaildmql rrlaalerqr iiddlakiea eiadledila kperqrgivr
    481 delaeivdrh gddrrtriia adgdvsdedl iaredvvvti tetgyakrtk tdlyrsqkrg
    541 gkgvqgaglk qddivahffv csthdlilff ttqgrvyrak aydlpeasrt argqhvanll
    601 afqpeeriaq viqirgytda pylvlatrng lvkkskltdf dsnrsggiva vnlrdndelv
    661 gavlcsaddd lllvsangqs irfsatdeal rpmgratsgv qgmrfniddr llslnvvreg
    721 tyllvatsgg yakrtaieey pvqgrggkgv ltvmydrrrg rlvgalivdd dselyavtsg
    781 ggvirtaarq vrkagrqtkg vrlmnlgegd tllaiarnae esgddnavda ngadqtgn
//
```

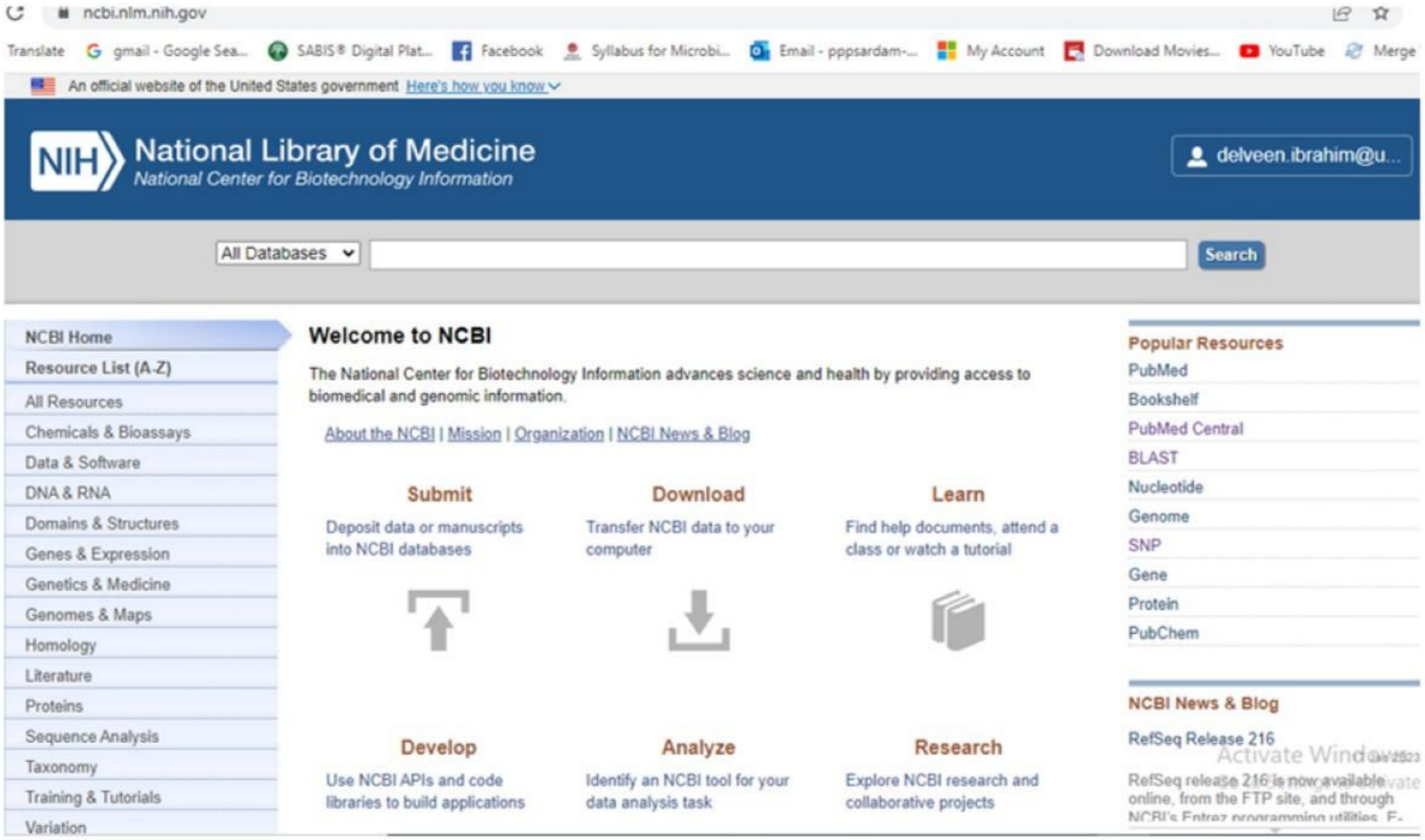
GenPept is a database format that provides protein sequence records derived from nucleotide coding sequences (CDS) in the GenBank database

Exploring NCBI:

Practical section

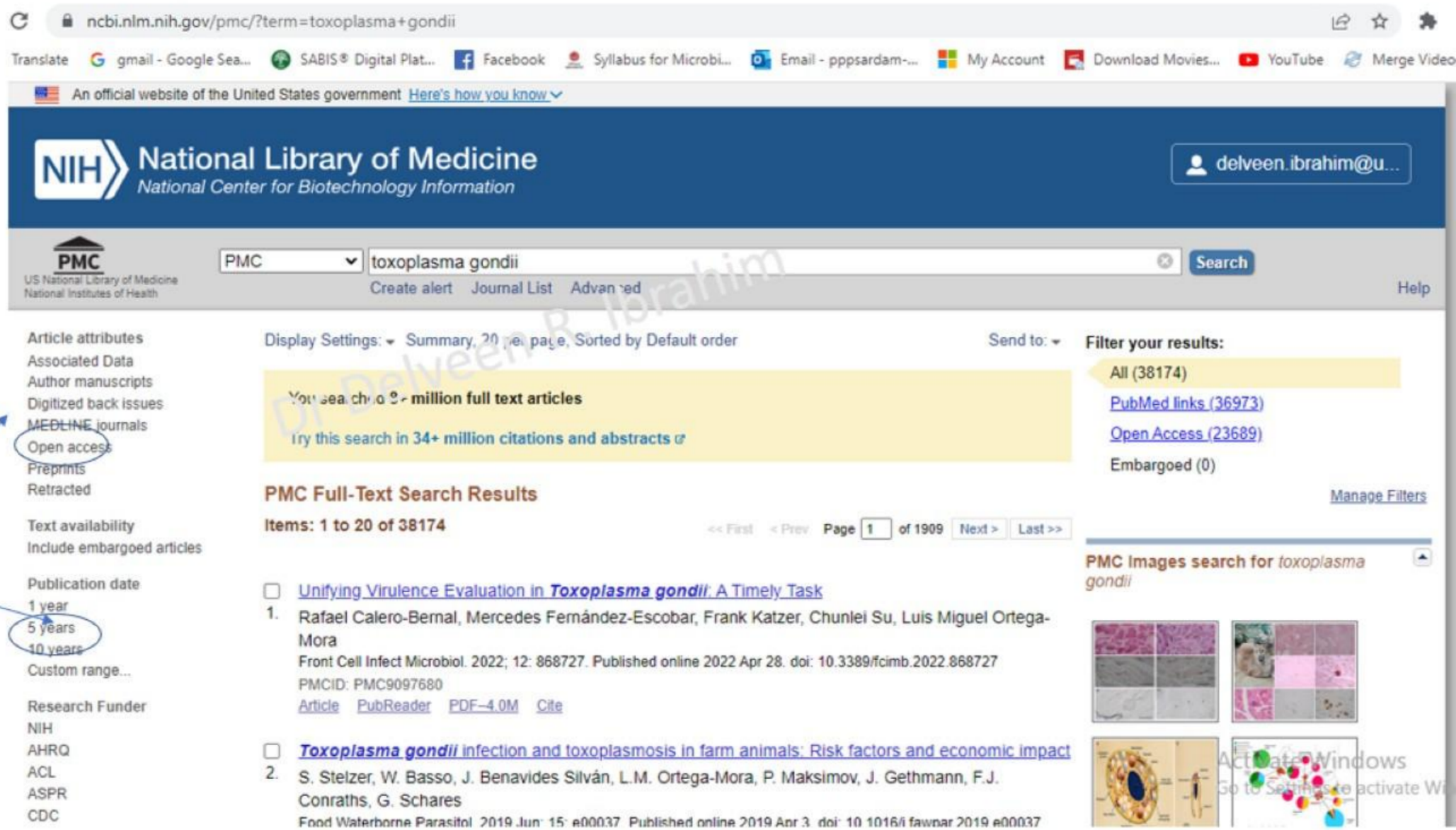
Lec 3

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

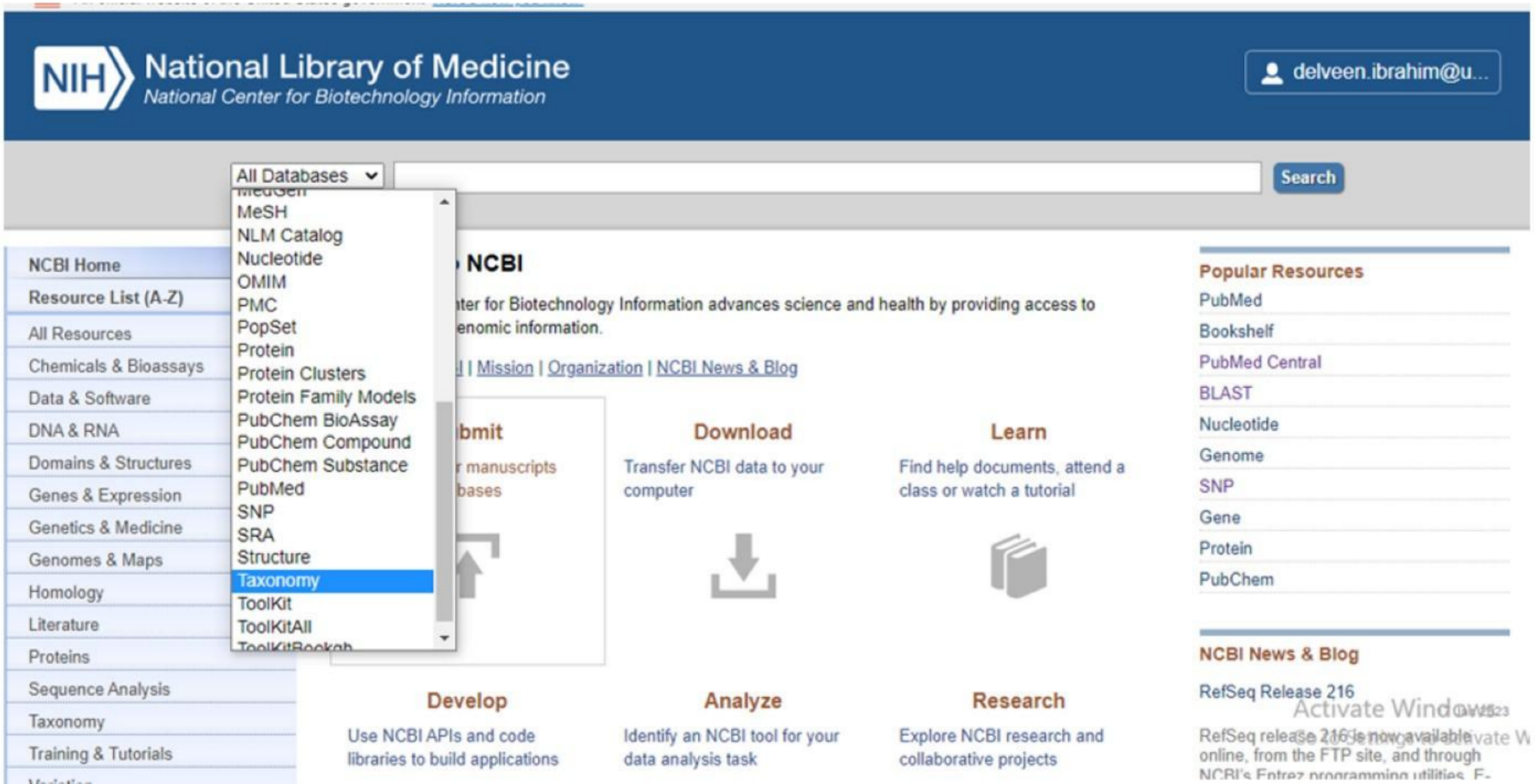


2. For literature you can use PMC and PubMed as I explained for you before

You can use different filters



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



ncbi.nlm.nih.gov/taxonomy/?term=Aspergillus+niger

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

NIH National Library of Medicine
National Center for Biotechnology Information

delveen.ibrahim@u...

Taxonomy | Taxonomy | Aspergillus niger | Search

Create alert Limits Advanced

Display Settings: Summary

Send to:

Aspergillus niger
species, ascomycete fungi
Nucleotide Protein

Related information
Nucleotide
Protein
Assembly

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

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

NCBI 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

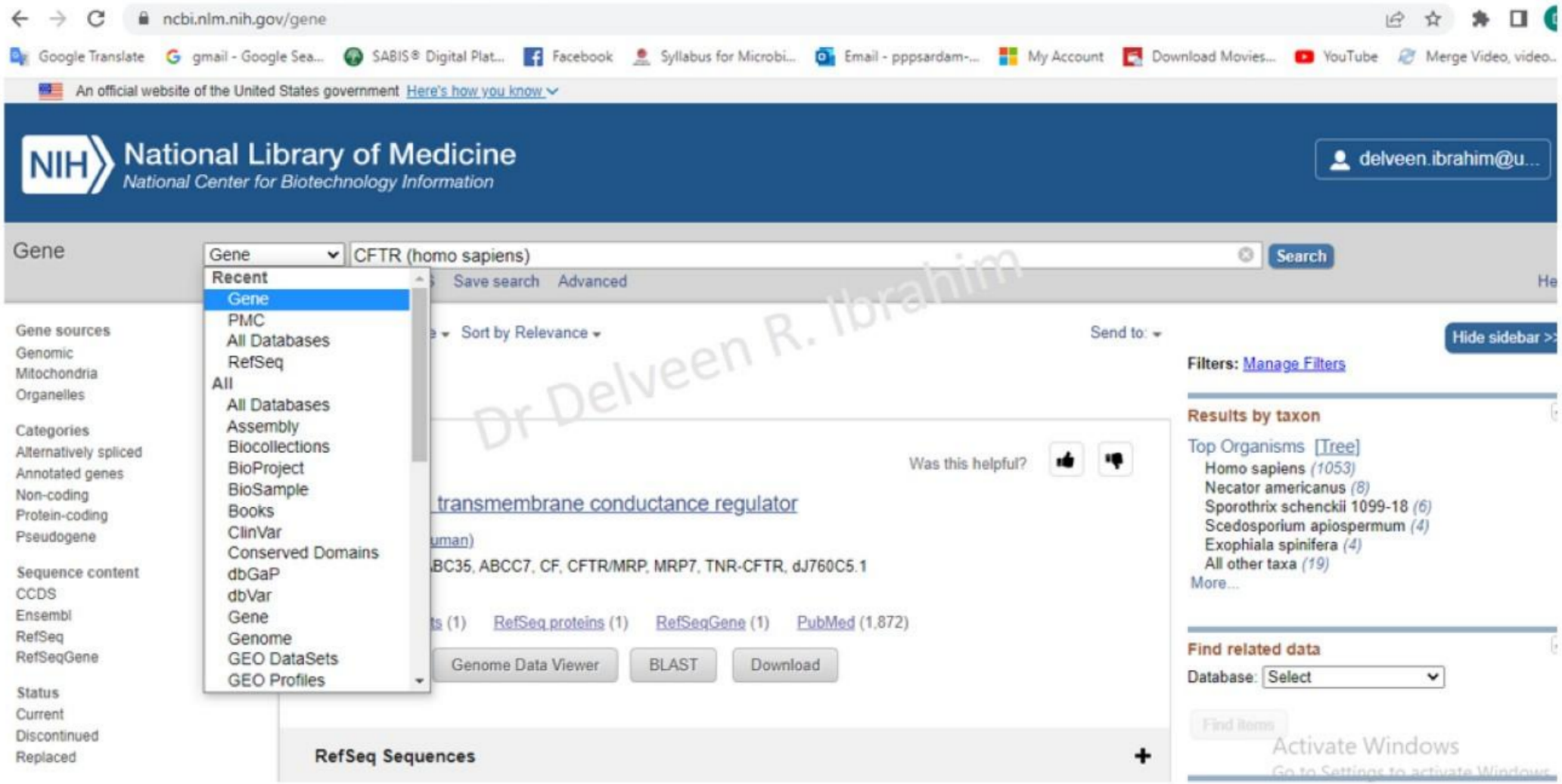
Nucleotide, Protein, Structure, Genome, Popset, SNP, Conserved Domains, GEO Datasets, PubMed Central
Gene, HomoloGene, SRA Experiments, LinkOut, BLAST, GEO Profiles, Protein Clusters, Identical Protein Groups, SPARCLE
BioProject, BioSample, Assembly, dbVar, Genetic Testing Registry, Host, Viral Host,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

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

- **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



- 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 [*Homo sapiens* (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 [*Homo sapiens* (human)]	Chromosome 10, NC_000010.11 (99782640..99852594)	ABC30, CMOAT, DJS, MRP2, cMRP	601107
ABCC8 ID: 6833	ATP binding cassette subfamily C member 8 [*Homo sapiens* (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 [*Homo sapiens* (human)]	Chromosome 13, NC_000013.11 (95019835..95301451, complement)	MOAT-B, MOATB, MRP4	605250
ABCC1 ID: 4363	ATP binding cassette subfamily C member 1 [*Homo sapiens* (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; Ute
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrh
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 6070)
```


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

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
GTAGTAGGTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACC
ATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGTTGTCTCCAAACTTTTTTTCAGCTGGACCAGACCAATTT
TGAGGAAAGGATACAGACAGCGCCTGGAATTGTCAGACATATACCAAATCCCTTCTGTTGATTCTGCTGA
CAATCTATCTGAAAAATTGGAAAGAGAATGGGATAGAGAGCTGGCTTCAAAGAAAAATCCTAAACTCATT
AATGCCCTTCGGCGATGTTTTTTCTGGAGATTTATGTTCTATGGAATCTTTTTATATTTAGGGGAAGTCA
CCAAAGCAGTACAGCCTCTCTTACTGGGAAGAATCATAGCTTCCTATGACCCGGATAACAAGGAGGAACG
CTCTATCGCGATTTATCTAGGCATAGGCTTATGCCTTCTCTTTATTGTGAGGACACTGCTCCTACACCCA
GCCATTTTTGGCCTTCATCACATTGGAATGCAGATGAGAATAGCTATGTTTAGTTTGATTTATAAGAAGA
CTTTAAAGCTGTCAAGCCGTGTTCTAGATAAAATAAGTATTGGACAACTTGTTAGTCTCCTTTCCAACAA
CCTGAACAAATTTGATGAAGGACTTGCATTGGCACATTTCGTGTGGATCGCTCCTTTGCAAGTGGCACTC
CTCATGGGGCTAATCTGGGAGTTGTTACAGGCGTCTGCCTTCTGTGGACTTGGTTTCCTGATAGTCCTTG
CCCTTTTTCAGGCTGGGCTAGGGAGAATGATGATGAAGTACAGAGATCAGAGAGCTGGGAAGATCAGTGA
AAGACTTGTGATTACCTCAGAAATGATTGAAAATATCCAATCTGTTAAGGCATACTGCTGGGAAGAAGCA
```

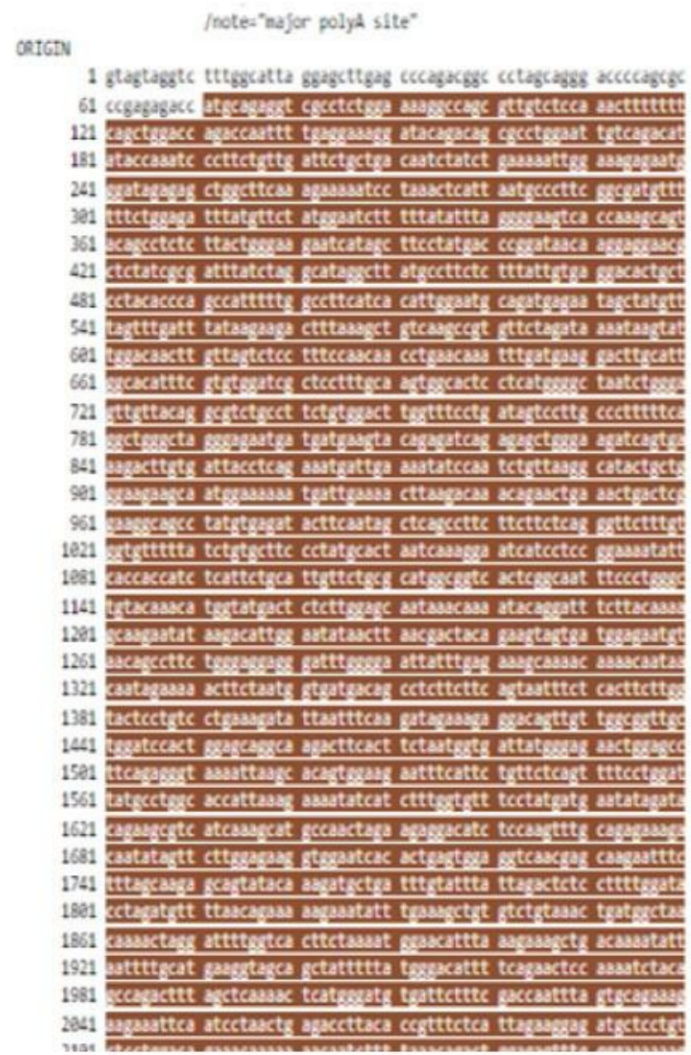
- **Identify CDS (Coding sequence),.** Go back in the previous window. Scroll down and click on CDS

```
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



Number of exons and their sequence can also be studied.

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

```
/inference="alignment:Splign:2.1.0"
exon    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"
```

```
2641 gagcttaatt tttgtgctaa tttggtgctt agtaattttt ctggcagagg tggctgcttc
2701 tttggttgtg ctgtggctcc ttggaaacac tcctcttcaa gacaaaggga atagtactca
2761 tagtagaaat aacagctatg cagtgattat 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 tgtcgcagtt ttacaaccct acatctttgt
3121 tgcaacagtg ccagtgatag tggcttttat tatgttgaga gcatatttcc tccaaacctc
3181 acagcaactc aaacaactgg aatctgaagg caggagtcca attttcactc atcttgttac
3241 aagcttaaaa ggactatgga cacttcgtgc cttcggacgg cagccttact ttgaaactct
3301 gttccacaaa gctctgaatt tacatactgc caactggttc ttgtacctgt caacactgcg
3361 ctggttccaa atgagaatag aaatgatttt tgtcatcttc ttcattgctg ttaccttcat
3421 ttccatttta acaacagagag aaggagaagg aagagttggt attatcctga ctttagccat
3481 gaatatcatg agtacattgc agtgggctgt aaactccagc atagatgtgg atagcttgat
3541 gcgatctgtg agccgagtct ttaagttcat tgacatgcca acagaaggta aacctaccaa
3601 gtcaaccaaa ccatacaaga atggccaact ctcgaaagtt atgattattg agaattcaca
3661 cgtgaagaaa gatgacatct ggccctcagg gggccaaatg actgtcaaag atctcacagc
3721 aaaatacaca gaaggtggaa atgccatatt agagaacatt tccttctcaa taagtcctgg
3781 ccagagggtg ggcctcttgg gaagaactgg atcagggaag agtactttgt tatcagcttt
3841 tttgagacta ctgaacactg aaggagaaat ccagatcgat ggtgtgtctt gggattcaat
3901 aactttgcaa cagtggagga aagcctttgg agtgatacca cagaaagtat ttatttttc
3961 tggaacattt agaaaaaact tggatcccta tgaacagtgg agtgatcaag aaatatggaa
4021 agttgcagat gaggttgggc tcagatctgt gatagaacag tttcctggga agcttgactt
4081 tgtccttgtg gatgggggct gtgtcctaag ccatggccac aagcagttga tgtgcttggc
4141 tagatctgtt ctcagtaagg cgaagatctt gctgcttgat gaacccagtg ctcatttgga
4201 tccagtaaca taccaaataa ttagaagaac tctaaaacaa gcatttgctg attgcacagt
4261 aattctctgt gaacacagga tagaagcaat gctggaatgc caacaatttt tggtcataga
4321 agagaacaaa gtgcggcagt acgattccat ccagaaactg ctgaacgaga ggagcctctt
4381 ccggcaagcc atcagcccct ccgacagggt gaagctcttt ccccaccgga actcaagcaa
4441 gtgcaagtct aagccccaga ttgctgctct 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

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