Bioinformatics Lec 3

National Centre for Biotechnology Information (NCBI)

By
Dr Delveen R. Ibrahim

1

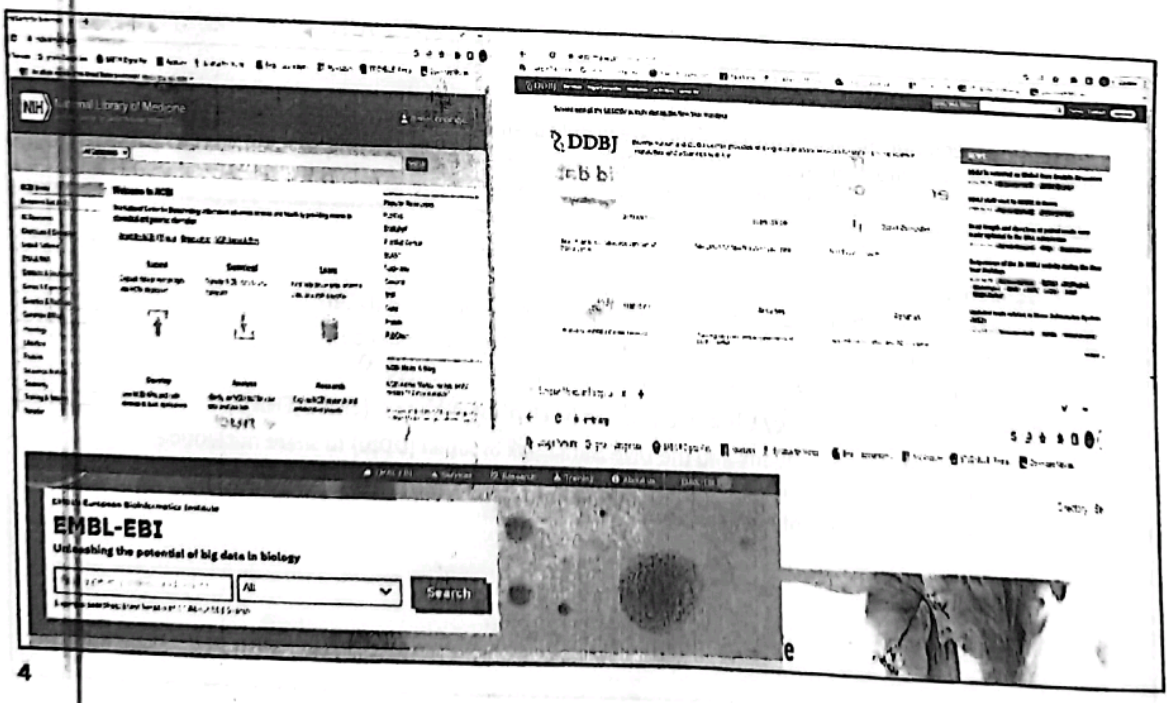
There are several **nucleic acid databases** that are widely used for bioinformatics analysis. These databases contain information about DNA and RNA sequences, their structures, functions, and related annotations. Here are some prominent nucleic acid databases:

- **GenBank:**
  - Managed by the National Center for Biotechnology Information (NCBI).
  - Comprehensive repository of publicly available nucleotide sequences.
  - Includes genomic DNA, RNA, and protein sequences.
- **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.

2

- **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.
  - Similar to 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.
  - Integrates data from various specialized RNA databases.
  - And many others.....

3



4

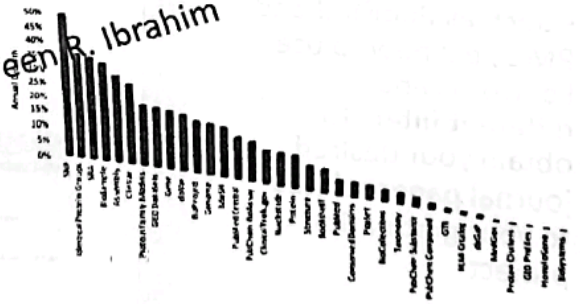
Bioinformatics Tools:

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.
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.

5

NCBI

- NCBI maintains a diverse set of 37 main databases that together contain more than 3.6 billion records , most of which are available 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.

6

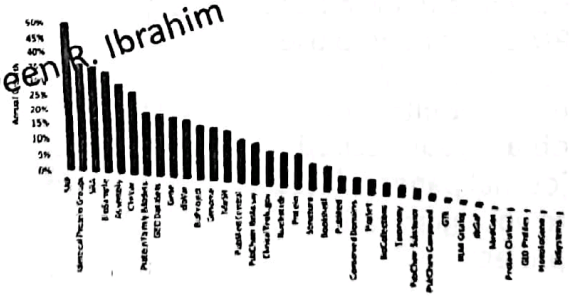
Bioinformatics Tools:

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.
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.

5

NCBI

- NCBI maintains a diverse set of 37 main databases that together contain more than 3.6 billion records , most of which are available 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.

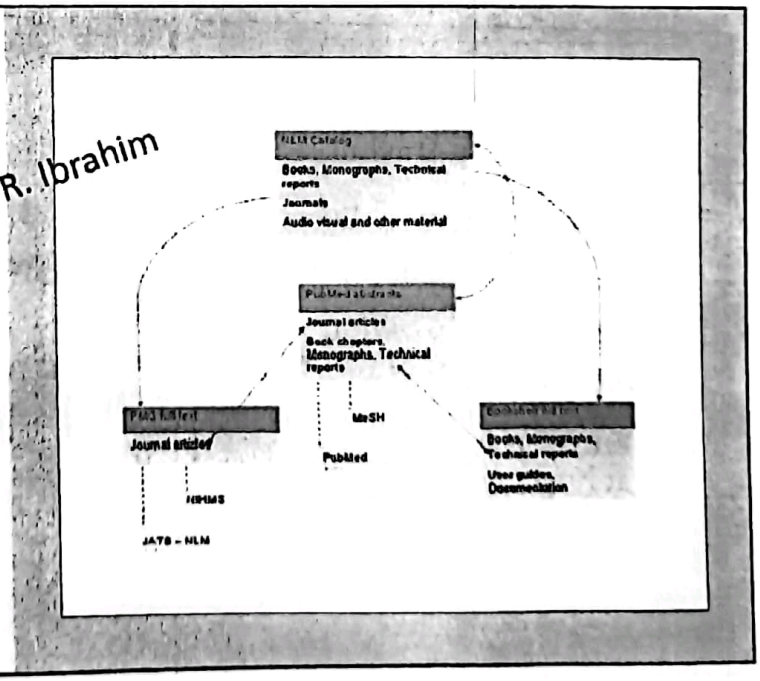
6

- 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.**

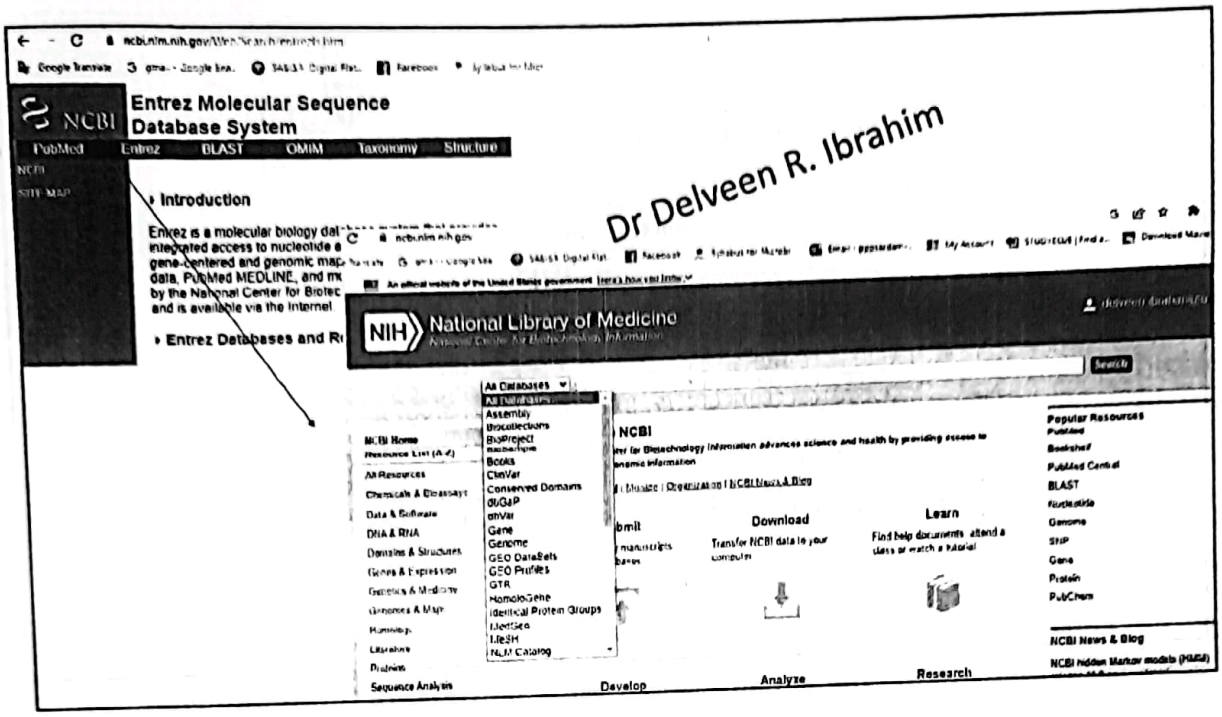
7

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



8



9

- 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 Are There So Many Different Types??

10

10

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

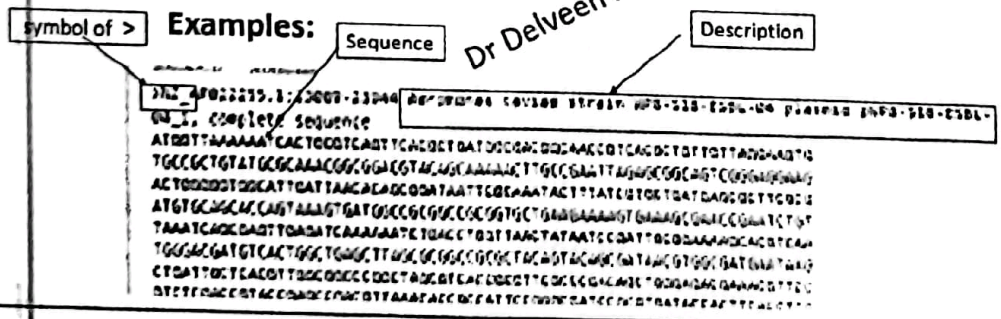
➢ **FASTA format:**

-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 .

Examples:



➢ **Genbank format:**

-Can contain several sequences.

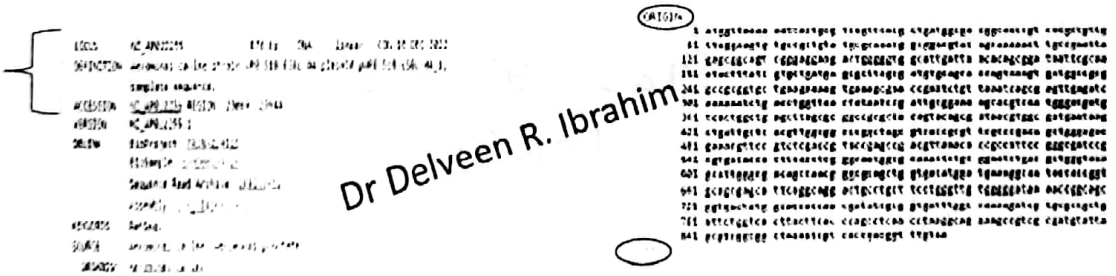
- The first section Includes the entry's LOCUS, DEFINITION, ACCESSION and VERSION- followed by many others annotations.
- This end by " ORIGION" which is the start of the sequence, and the end of the sequence is marked by two slashes ("//").
- The Genbank format allows for the storage of information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

13

13

➢ **Genbank format:**

Examples:



14

14

GenPept is a database of GenBank gene products, namely the translation of all CDS (coding sequence) features with a translation qualifier. GenPept is not an official release from the NCBI but is thoroughly maintained and each new release of GenBank.

GenPept Format

GenPept format is text-based and derived from the parent GenBank format. Following is an example of a GenPept record, note that only the beginning of the record is shown:

```
1 LOCUS       NP_001091               377 aa          linear   PRI 20-MAY-2008
2 DEFINITION  actin, alpha 1, skeletal muscle [Homo sapiens].
3 ACCESSION   NP_001091
4 VERSION     NP_001091.1  GI:4501881
5 DBSOURCE    REFSEQ: accession NM_001100.3
6 KEYWORDS    .
7 SOURCE      Homo sapiens (human)
8   ORGANISM  Homo sapiens
9             Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
10            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
11            Catarrhini; Hominidae; Homo.
```

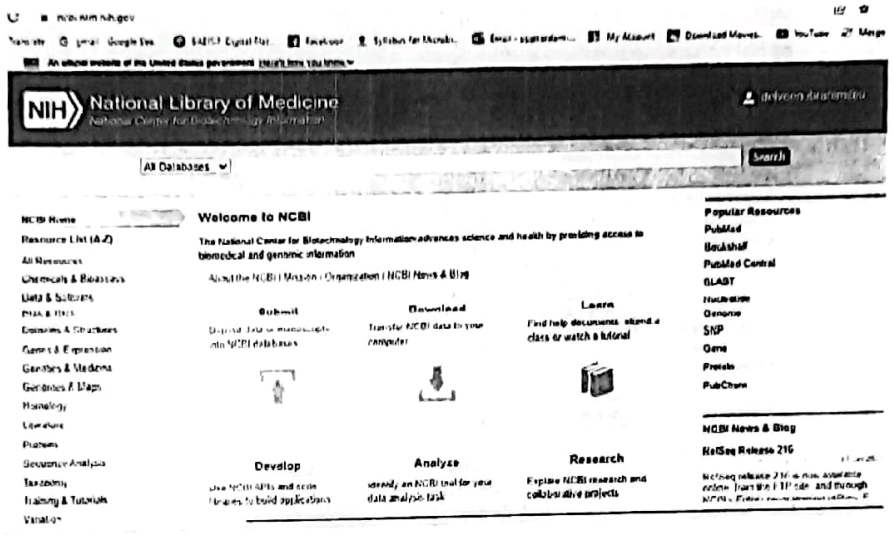
It is a multi-sequence format and each sequence is terminated by a double-slash sign ('//').

15

Practical part
Lec 3

16

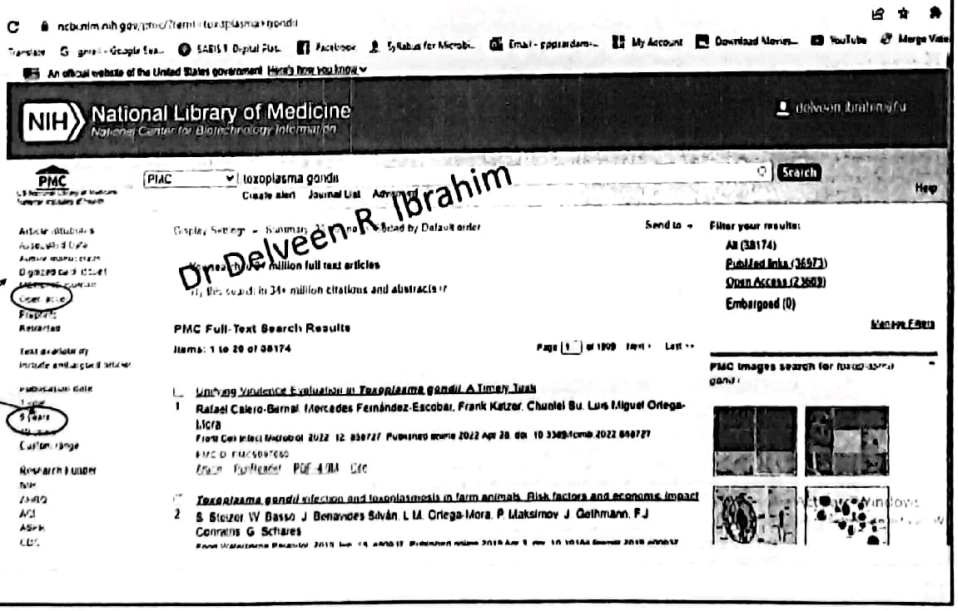
1. First open https://www.ncbi.nlm.nih.gov , you will see the following webpage



17

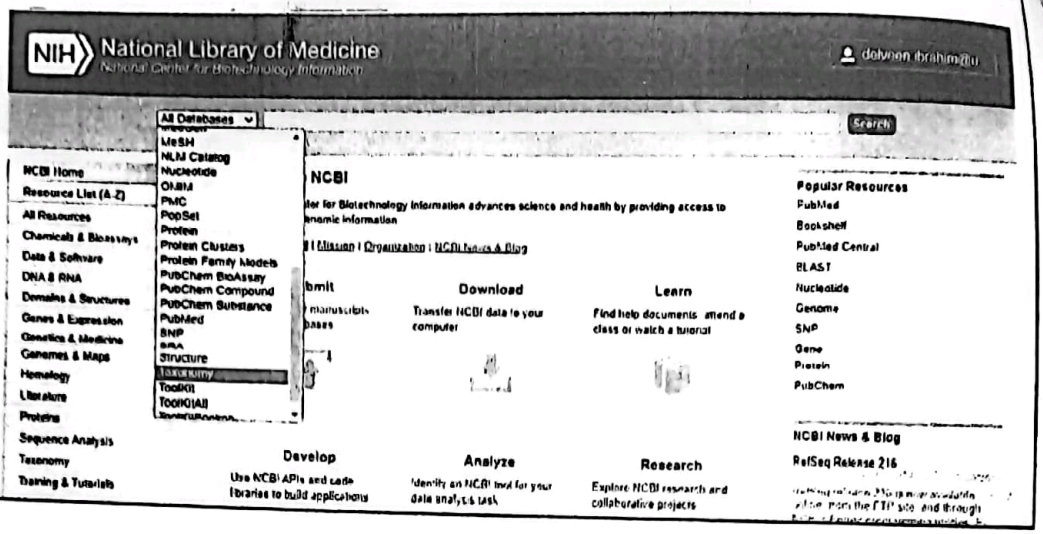
2. For literature you can use PMC and PubMed as I explained for you before

You can use different filters

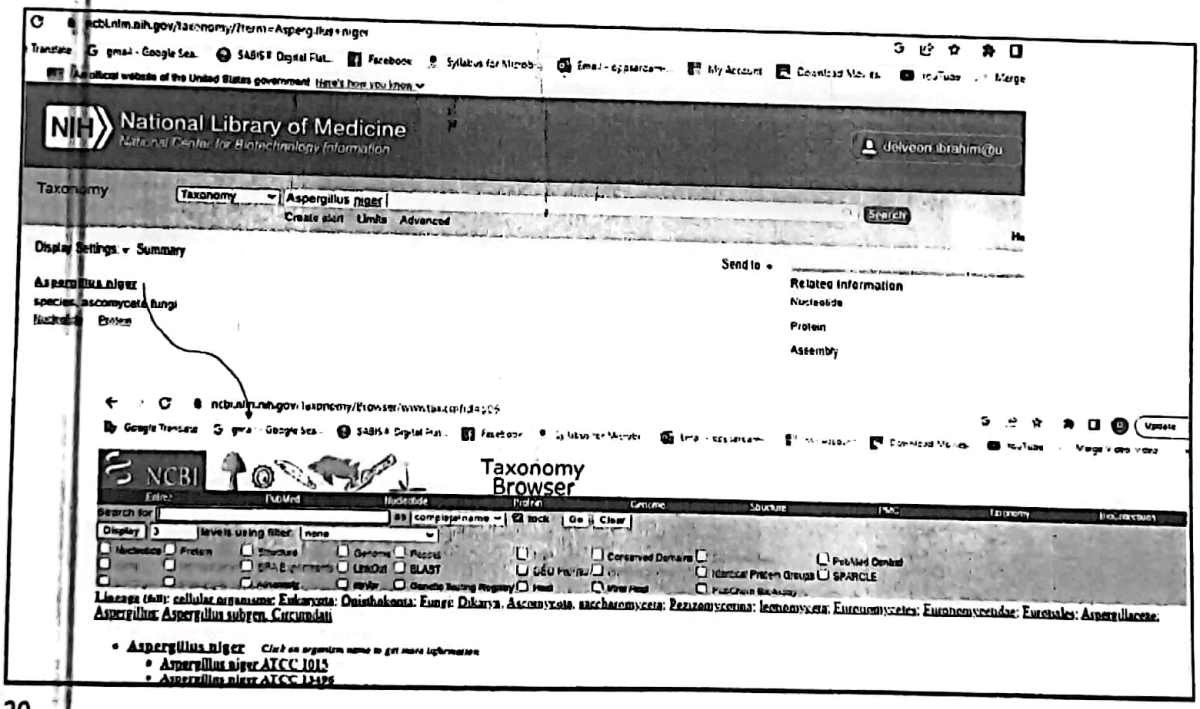


18

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



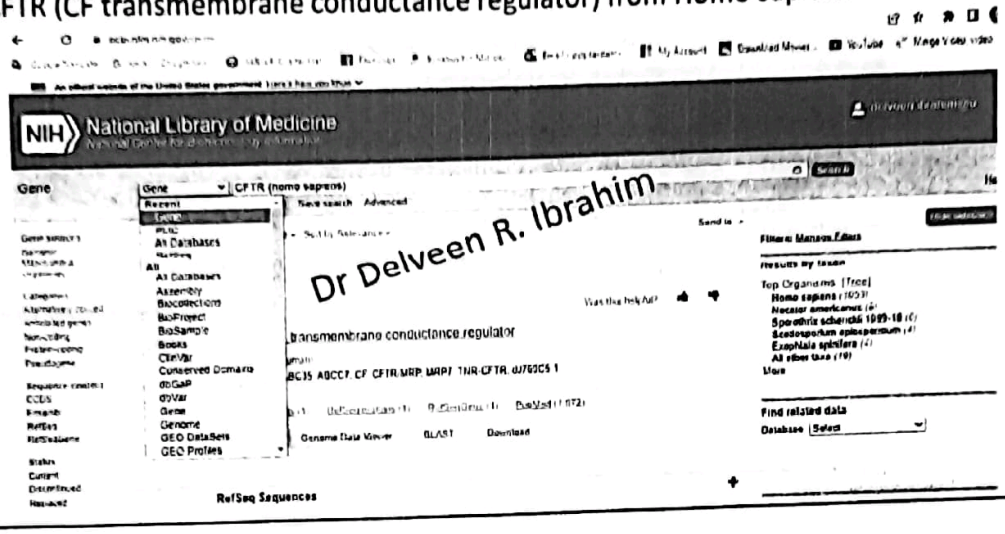
19



20

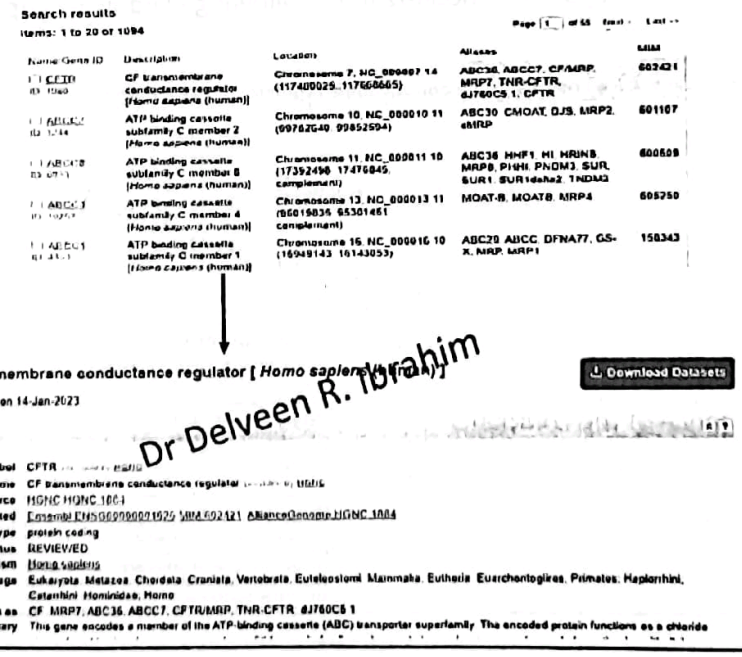
4. Gene (transcript) sequence retrieval from NCBI database:

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



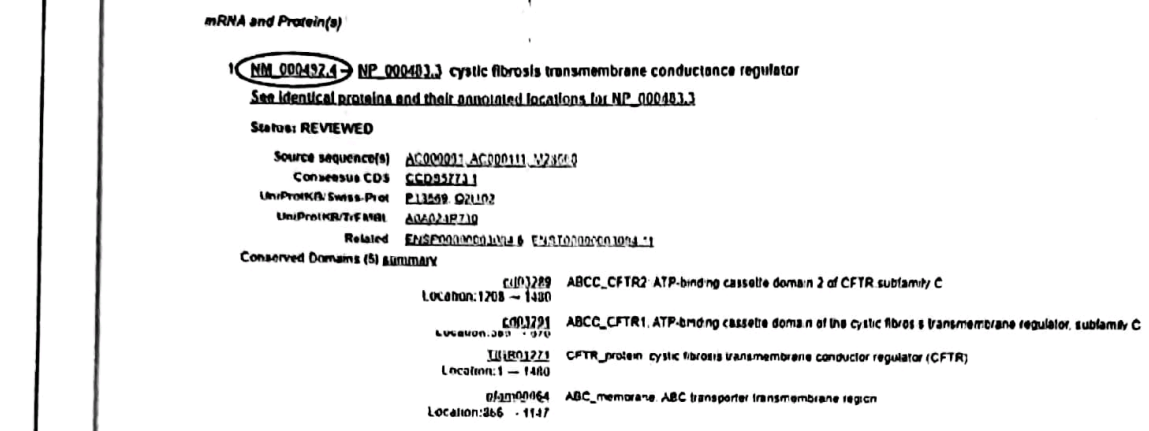
21

- Click on CFTR gene in the window.



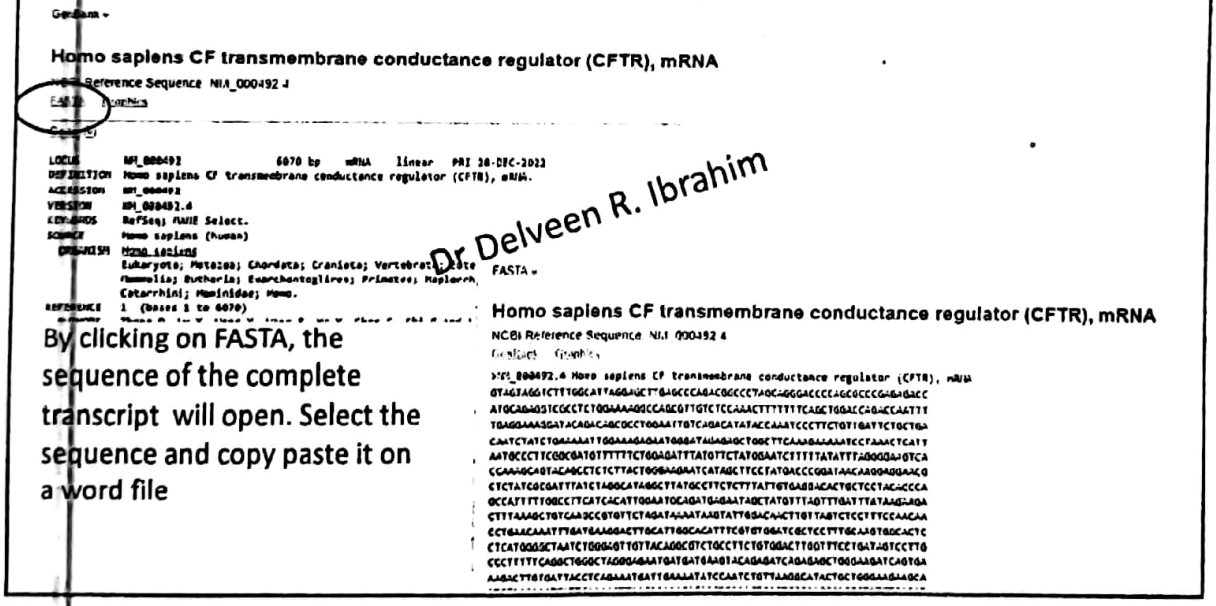
22

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



23

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

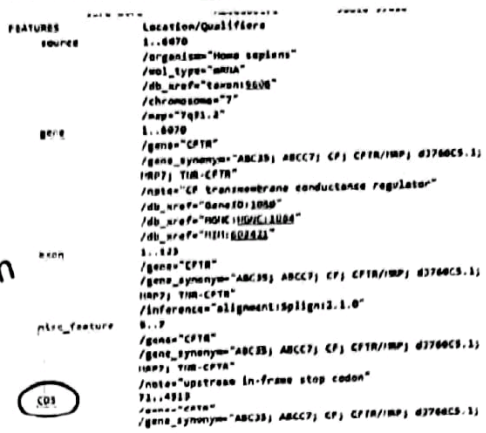


By clicking on FASTA, the sequence of the complete transcript will open. Select the sequence and copy paste it on a word file

Dr Delveen R. Ibrahim

24

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

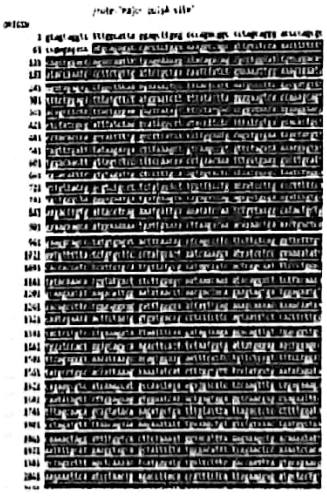


25

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.

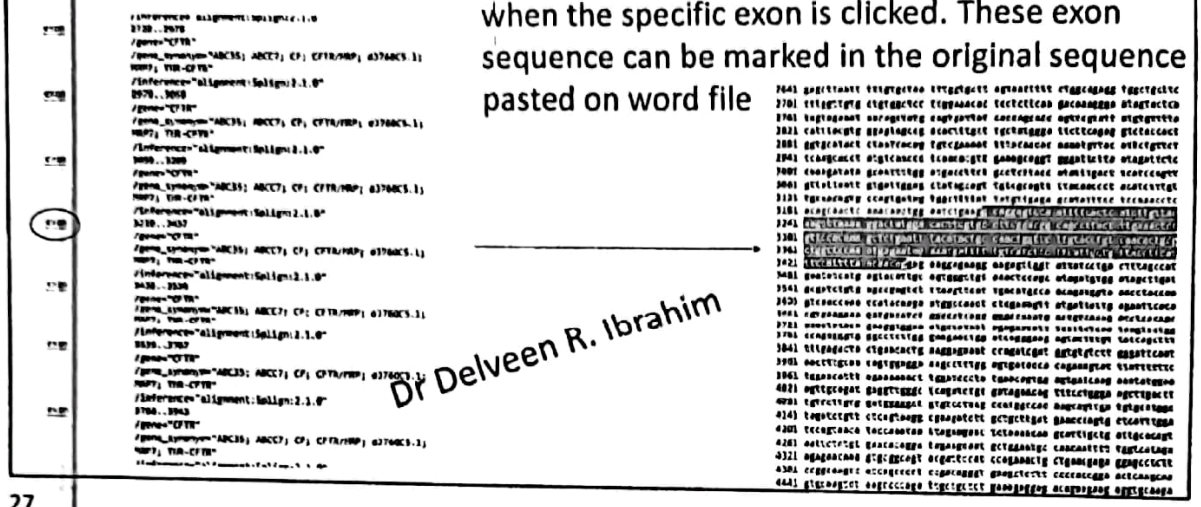
Like this CDS, 5' UTR and 3' UTR sequence of any gene or transcript can be retrieved for further analysis, cloning and other purposes



26

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

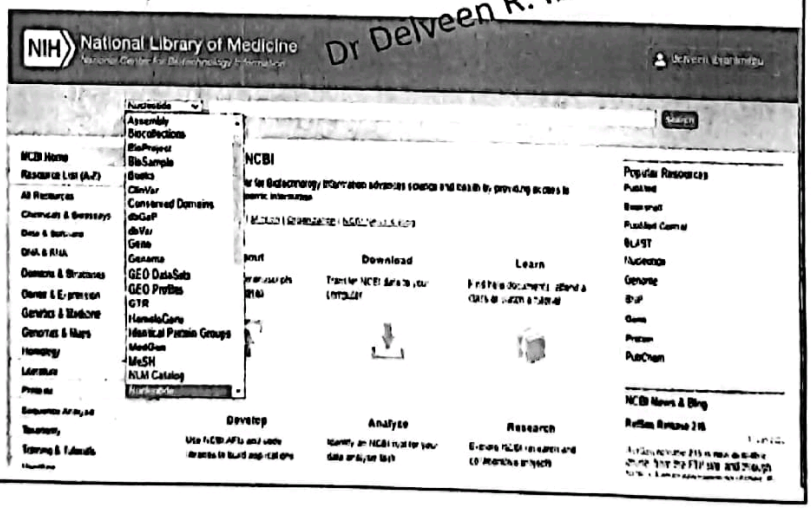


Dr Delveen R. Ibrahim

27

Finding Nucleotide sequence by using **accession number**, or **key words** such as the gene name and organism by searching the Nucleotide section in NCBI

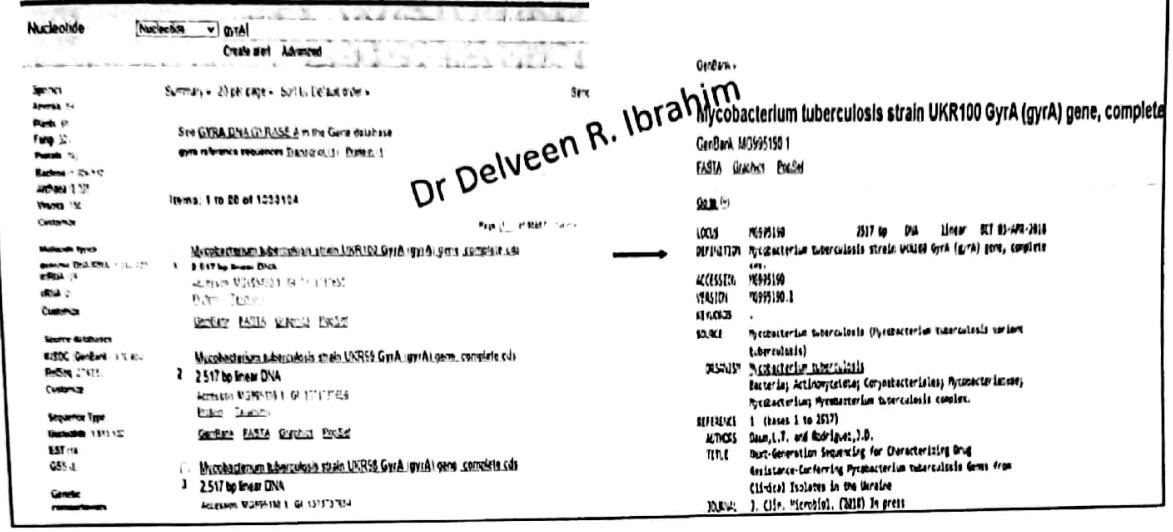
On left upper side select **Nucleotide.** Type accession number , name of gene or gene symbol and organism name



Dr Delveen R. Ibrahim

28

Type *gyrA* (gyrase subunit A) for example, you will get the window below. Click on the first option , you will get the GenBank and you can do copy the complete gene or the CDS just like before



29

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.

30