

(3)

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.

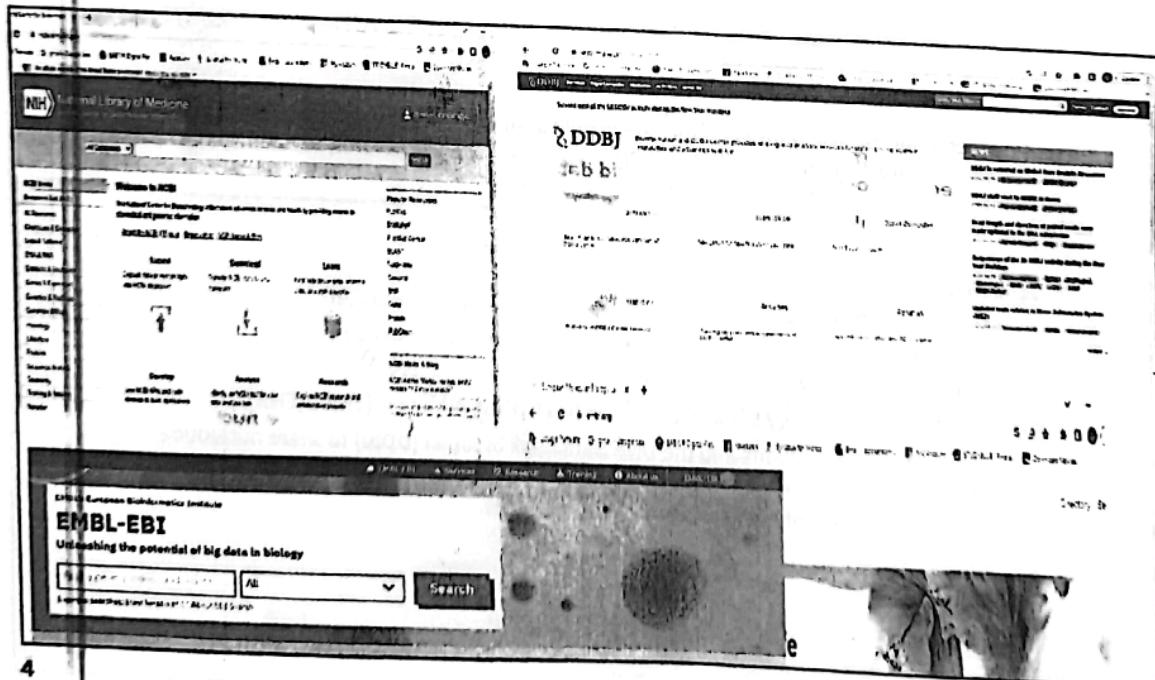
RefSeq (Reference Sequence):

- 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

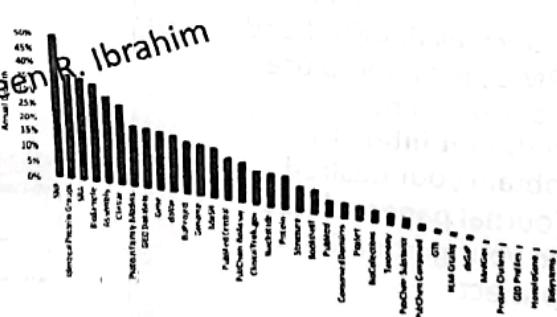
Bioinformatics Tools:

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

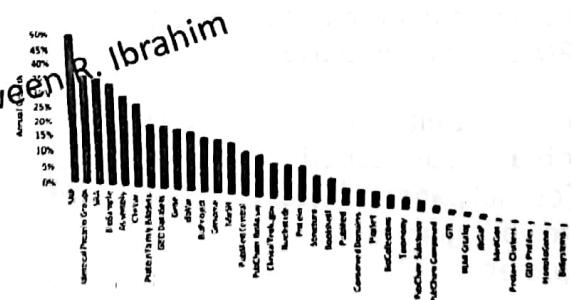
Bioinformatics Tools:

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.

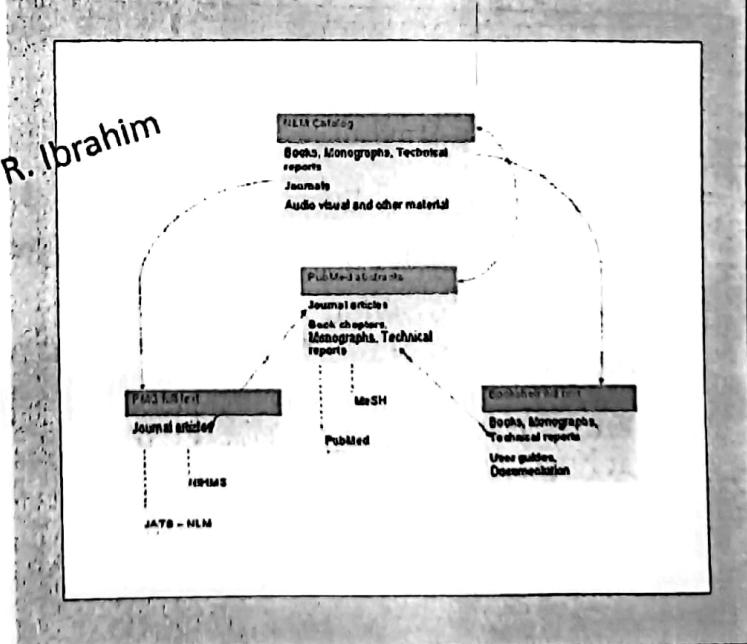
6

- 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.

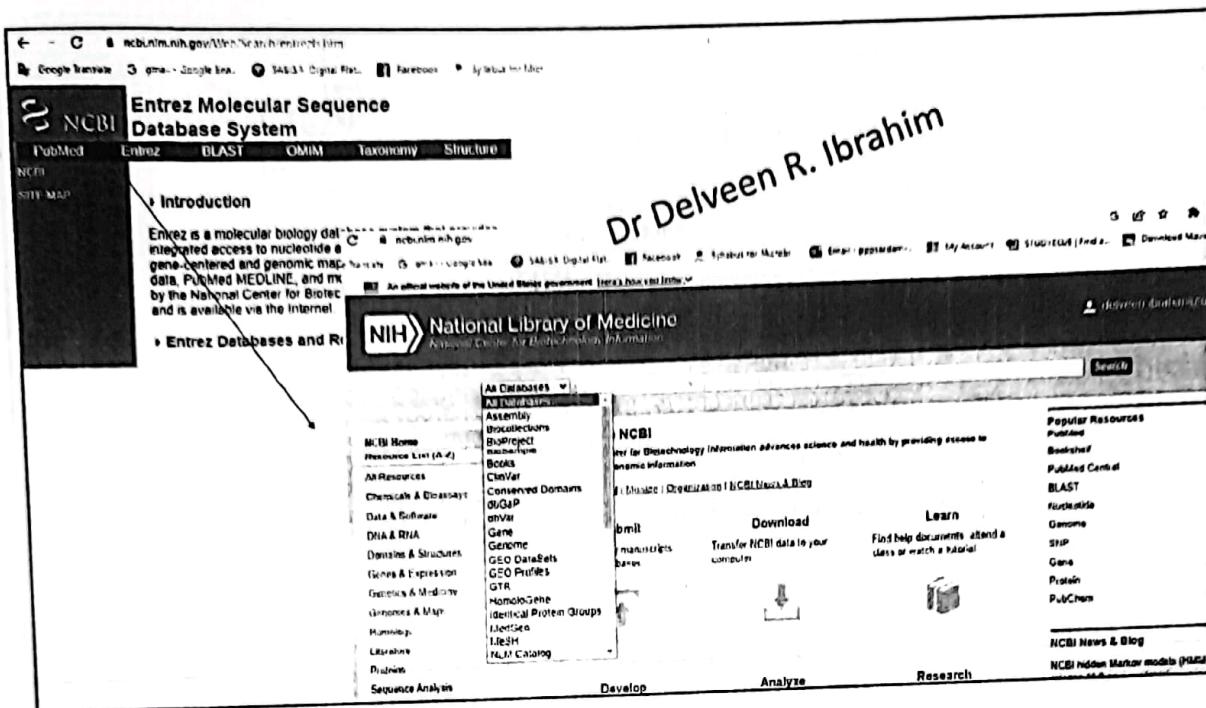
7

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

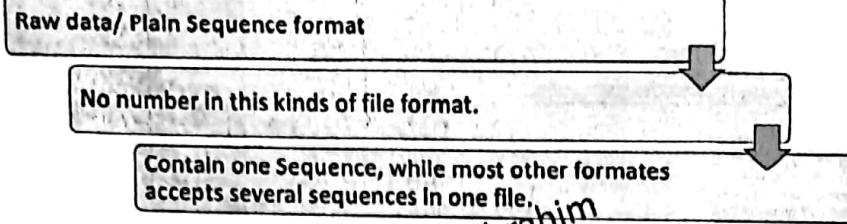
➤ Some of the File format used in Bioinformatics:

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

Why Are There So Many Different Types??

10

10



Examples:

ACAAGATGCCATTGCCCCGGCTCTGCTGCTGCTCCGGGGCACGGCCACCGCTGC
CTGCCCCCTGGAGGGTGGCCCCACGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGA
TAAGGAAAAGCAGCCTCTGACTTCTGCTTGGTTTGAGTGGACCTCCAGGCCAGTGC
CGGGCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGCAGGAAGGGCACCCCC
CAGCAATCCGCGCGGGACAGAATGCCCTGCAAGAACCTCTGGAAAGACCTCTCTCT
GCAAATAAACCTCACCATGAATGCTACGCAAGTTAATTACAGACCTGAA

➤ 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

➤ Genbank format:

- Can contain several sequences.
 - The first section includes the entry's LOCUS, DEFINITION, ACCESSION and VERSION- followed by many others annotations.
 - This ends 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

- 13 -

13

➤ Genbank format:

Examples:

Dr Delveen R. Ibrahim

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 PPT 20-JULY-2008
2 DEFINITION actin, alpha 1, skeletal muscle [Homo sapiens].
3 ACCESSION NP_001091
4 VERSION NP_001091.1 GI:4501881
5 DBSOURCE NEPSEQ; accession NM_001100.3
6 KEYWORDS .
7 SOURCE Homo sapiens (human)
8 ORGANISM Homo sapiens
9
10 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
11 Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini;
Catarrhini; Hominoidea; Homo.
```

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

15

Practical part
Lec 3

16

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

17

- 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

The screenshot shows the NCBI homepage with the 'All Databases' dropdown menu open. The 'Taxonomy' option is highlighted in the list. The main page features sections for 'Develop', 'Analyze', and 'Research', along with links to 'NCBI News & Blog' and 'RefSeq Release 216'.

19

The screenshot shows the NCBI Taxonomy browser for the term 'Aspergillus niger'. The search bar at the top contains 'Aspergillus niger'. Below the search bar, there are tabs for 'Taxonomy' (selected), 'Create strain', 'Units', and 'Advanced'. The main content area displays the taxonomic tree for Aspergillus niger, with a callout pointing to the 'Aspergillus niger species, ascomycetidae fungi' node. On the right side, there are sections for 'Related Information' (Nucleotide, Protein, Assembly) and a 'Send to' button. At the bottom, there is a search interface for 'Display' options and a 'Search' button.

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

NIH National Library of Medicine

Gene Search Advanced

Recent

CFTR (homo sapiens)

CF transmembrane conductance regulator

BC15 ABC07, CFTR.MRP, MRP1.THR.CFTR, dJ76051

Find related data Database [Select]

21

- Click on CFTR gene in the window.

Search results Items: 1 to 20 of 1094

Name/Gene ID	Description	Location	Aliases	MMR
CFTR	CF transmembrane conductance regulator [Homo sapiens (human)]	Chromosome 7, NC_000007.14 (11740025..11766065)	ABC36, ABC07, CFTR.MRP, MRP1, THR-CFTR, dJ76051, CFTR	602421
ATP binding cassette subfamily C member 2	[Homo sapiens (human)]	Chromosome 10, NC_000010.11 (99782G40..99852K94)	ABC30, CMOAT, OJS, MRP2, eMRP	601107
ATP binding cassette subfamily C member 8	[Homo sapiens (human)]	Chromosome 11, NC_000011.10 (173924G0..174768A5, complement)	ABC35, HHF1, M1, MRNBS, MRP8, PIHH, PNMD2, SUR, SUR1, deha2, TNMD2	600509
ATP binding cassette subfamily C member 4	[Homo sapiens (human)]	Chromosome 13, NC_000013.11 (B6019835..B5301461, complement)	MOAT-B, MOAT-E, MRP4	605230
ATP binding cassette subfamily C member 1	[Homo sapiens (human)]	Chromosome 16, NC_000016.10 (16949143..16143053)	ABC20, ABCD, DFNA77, OS-X, MRP, MRP1	158343

CFTR - CF transmembrane conductance regulator [Homo sapiens]

Gene ID 1000 updated on 14-Jan-2023

Summary

Official Symbol: CFTR
Official Full Name: CF transmembrane conductance regulator
Primary source: HGNC/HGNC 1001
See related: Ensemble, ENSG0000001625, MRP1, dJ76051, AllianceGenome, HGNC 1004
Gene type: protein coding
RefSeq status: REVIEWED
Organism: Homo sapiens
Lineage: Eukaryota Metazoa Chordata Craniata, Vertebrata Eutelesiomorpha Mammalia Eutheria Euarchontoglires Primates Haplorhini, Catarrhini Hominoidea Homo
Also known as: CF, MRP1, ABC36, ABC07, CFTR/MRP, THR-CFTR, dJ76051
Summary: This gene encodes a member of the ATP-binding cassette (ABC) transporter superfamily. The encoded protein functions as a chloride

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.

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) AC00001 AC000111 V23510

Conensus CDS CDD352771

UniProtKB/Swiss-Prot P13269 Q2L102

UniProtKB/Trembl AUAQ23P710

Related ENSP00000001014 E121010000001094-1

Conserved Domains (5) SUMMARY

CDD1289 ABCC_CFTR2 ATP-binding cassette domain 2 of CFTR subfamily C

Location: 1208 - 1480

CDD1291 ABCC_CFTR1 ATP-binding cassette domain of the cystic fibrosis transmembrane regulator, subfamily C

Location: 309 - 970

CFTR1271 CFTR_protein cystic fibrosis transmembrane conductor regulator (CFTR)

Location: 1 - 1480

Dggn00064 ABC_membrane ABC transporter transmembrane region

Location: 366 - 1147

23

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

General

Homo sapiens CF transmembrane conductance regulator (CFTR), mRNA

NCBI Reference Sequence NM_000492.4

FASTA

Copy

LOCUS NM_000492 6670 bp mRNA linear PRJ:30-DFC-2023

DEFINITION Homo sapiens CF transmembrane conductance regulator (CFTR), mRNA.

ACCESSION NM_000492

VERSION NM_000492.4

KEYWORDS RefSeq; NM_000492.4

SOURCE Human Genome Project

ORGANISM Homo sapiens

ECOLOGY Eukaryotes; Metazoa; Chordata; Craniota; Vertebrates; Euteleostomi; Amniota; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo.

REFERENCE 1 (Bases 1 to 6670)

COMMENT This gene contains 24 exons and 23 introns and spans approximately 18 megabases on chromosome 7.

Dr. Delveen R. Ibrahim

FASTA =

Homo sapiens CF transmembrane conductance regulator (CFTR), mRNA

NCBI Reference Sequence NM_000492.4

FASTA

```
>NM_000492.4 Homo sapiens CF transmembrane conductance regulator (CFTR), mRNA
GTAGTAACTCTTGCGATTGGAGACCTGAGGCCAGACGCCCTAGCGGGGACCCGGACGCC
ATGAGAGGTCGCTCTTGGAAAAGGGCGACGGGTTGCTCAGAACCTTTTTCAAGCTGGACGAGCGATT
TGAAGAAAAGCATACACAGACCGCTCGGGATTTGTCAGCATATACTACAGATCTCTGCTGTTGATCTGCTG
CAAGCTATCTGAAATTGCGAAAAGGGATGGGGATTAAGAGGCTGGCTTCAAGAGGAAAGATCTCGAAGCTCAT
AAATGCCCTCGGGCGATGTTTTCTGGAGATTATGTTCTGATGAACTTCTTGTATATTGTTGGGGAAATGCA
CGAAAGCGAGTACAGCGCTCTCTCTCTGAGGGATCATAGCTCTCCGATACCCCGGATATAAGGGGGGGAG
CTCTATGCGCGATTAACTTAAAGGCGATAGGGCTTCTTCCTTCCTTATTTGAGGACACTGCTCCCTACGCC
GCCATTTTGGCGCTTCTCATCACATTGGGGATGCGATGAGGAAATAGCTATGTTGAGTTTAAAGGAAAGA
CTTAAAGGCGCTGAAAGCGCTGTTCTGAGGAAATTAAGGATTTAGGATTTGAGCTTCTCCCTTCGCGACAA
CTGAGGAGCAAACTTGGATGAGGGGACTTGCACTTGGGGGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
CTCATGCGGCGATCTGCGGGGGCTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
CCCCCTTTCGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

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

24

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

Dr Delveen R. Ibrahim

```

FEATURES          ID: ENSP00000000000000000000000000000000
source           Location/Qualifiers
                1..6070
                /organism="Homo sapiens"
                /mol_type="mRNA"
                /db_xref="GeneID:1050"
                /chromosome="7"
                /map="7q31.2"
                1..6070
                /gene="CFTR"
                /gene_symbol="ABC38; ABCC7; CF; CFTR/IMP; d3780C5.1"
                IMP; TIM-CFTR
                /notes="CF transmembrane conductance regulator"
                /db_xref="GeneID:1050"
                /db_xref="MIM:115011;LocusID:1050"
                /db_xref="HGNC:HGNC:1050"
                1..623
                /gene="CFTR"
                /gene_symbol="ABC38; ABCC7; CF; CFTR/IMP; d3780C5.1"
                IMP; TIM-CFTR
                /inference="alignment:SpliceN2.1.0"
                6..7
                /gene="CFTR"
                /gene_symbol="ABC38; ABCC7; CF; CFTR/IMP; d3780C5.1"
                IMP; TIM-CFTR
                /notes="Upstream In-frame stop codon"
                73..491
                /gene="CFTR"
                /gene_symbol="ABC38; ABCC7; CF; CFTR/IMP; d3780C5.1"

```

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.

```
278 /Inference="allignment:alignm:2.1.0"
279 .._TIR
280 /gene="CTTR"
281 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
282 HBP; TIR-CTTR"
283 /Inference="allignment:alignm:2.1.0"
284 .._TIR
285 /gene="CTTR"
286 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
287 HBP; TIR-CTTR"
288 /Inference="allignment:alignm:2.1.0"
289 .._TIR
290 /gene="CTTR"
291 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
292 HBP; TIR-CTTR"
293 /Inference="allignment:alignm:2.1.0"
294 .._TIR
295 /gene="CTTR"
296 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
297 HBP; TIR-CTTR"
298 /Inference="allignment:alignm:2.1.0"
299 .._TIR
300 /gene="CTTR"
301 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
302 HBP; TIR-CTTR"
303 /Inference="allignment:alignm:2.1.0"
304 .._TIR
305 /gene="CTTR"
306 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
307 HBP; TIR-CTTR"
308 /Inference="allignment:alignm:2.1.0"
309 .._TIR
310 /gene="CTTR"
311 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
312 HBP; TIR-CTTR"
313 /Inference="allignment:alignm:2.1.0"
314 .._TIR
315 /gene="CTTR"
316 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
317 HBP; TIR-CTTR"
318 /Inference="allignment:alignm:2.1.0"
319 .._TIR
320 /gene="CTTR"
321 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
322 HBP; TIR-CTTR"
323 /Inference="allignment:alignm:2.1.0"
324 .._TIR
325 /gene="CTTR"
326 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
327 HBP; TIR-CTTR"
328 /Inference="allignment:alignm:2.1.0"
329 .._TIR
330 /gene="CTTR"
331 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
332 HBP; TIR-CTTR"
333 /Inference="allignment:alignm:2.1.0"
334 .._TIR
335 /gene="CTTR"
336 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
337 HBP; TIR-CTTR"
338 /Inference="allignment:alignm:2.1.0"
339 .._TIR
340 /gene="CTTR"
341 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
342 HBP; TIR-CTTR"
343 /Inference="allignment:alignm:2.1.0"
344 .._TIR
345 /gene="CTTR"
346 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
347 HBP; TIR-CTTR"
348 /Inference="allignment:alignm:2.1.0"
349 .._TIR
350 /gene="CTTR"
351 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
352 HBP; TIR-CTTR"
353 /Inference="allignment:alignm:2.1.0"
354 .._TIR
355 /gene="CTTR"
356 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
357 HBP; TIR-CTTR"
358 /Inference="allignment:alignm:2.1.0"
359 .._TIR
360 /gene="CTTR"
361 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
362 HBP; TIR-CTTR"
363 /Inference="allignment:alignm:2.1.0"
364 .._TIR
365 /gene="CTTR"
366 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
367 HBP; TIR-CTTR"
368 /Inference="allignment:alignm:2.1.0"
369 .._TIR
370 /gene="CTTR"
371 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
372 HBP; TIR-CTTR"
373 /Inference="allignment:alignm:2.1.0"
374 .._TIR
375 /gene="CTTR"
376 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
377 HBP; TIR-CTTR"
378 /Inference="allignment:alignm:2.1.0"
379 .._TIR
380 /gene="CTTR"
381 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
382 HBP; TIR-CTTR"
383 /Inference="allignment:alignm:2.1.0"
384 .._TIR
385 /gene="CTTR"
386 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
387 HBP; TIR-CTTR"
388 /Inference="allignment:alignm:2.1.0"
389 .._TIR
390 /gene="CTTR"
391 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
392 HBP; TIR-CTTR"
393 /Inference="allignment:alignm:2.1.0"
394 .._TIR
395 /gene="CTTR"
396 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
397 HBP; TIR-CTTR"
398 /Inference="allignment:alignm:2.1.0"
399 .._TIR
400 /gene="CTTR"
401 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
402 HBP; TIR-CTTR"
403 /Inference="allignment:alignm:2.1.0"
404 .._TIR
405 /gene="CTTR"
406 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
407 HBP; TIR-CTTR"
408 /Inference="allignment:alignm:2.1.0"
409 .._TIR
410 /gene="CTTR"
411 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
412 HBP; TIR-CTTR"
413 /Inference="allignment:alignm:2.1.0"
414 .._TIR
415 /gene="CTTR"
416 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
417 HBP; TIR-CTTR"
418 /Inference="allignment:alignm:2.1.0"
419 .._TIR
420 /gene="CTTR"
421 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
422 HBP; TIR-CTTR"
423 /Inference="allignment:alignm:2.1.0"
424 .._TIR
425 /gene="CTTR"
426 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
427 HBP; TIR-CTTR"
428 /Inference="allignment:alignm:2.1.0"
429 .._TIR
430 /gene="CTTR"
431 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
432 HBP; TIR-CTTR"
433 /Inference="allignment:alignm:2.1.0"
434 .._TIR
435 /gene="CTTR"
436 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
437 HBP; TIR-CTTR"
438 /Inference="allignment:alignm:2.1.0"
439 .._TIR
440 /gene="CTTR"
441 /gene_synonym="ARCC5; ARCC7; CP; CTTR/HBP; d3760C5.1;
442 HBP; TIR-CTTR"
443 /Inference="allignment:alignm:2.1.0"
444 .._TIR
```

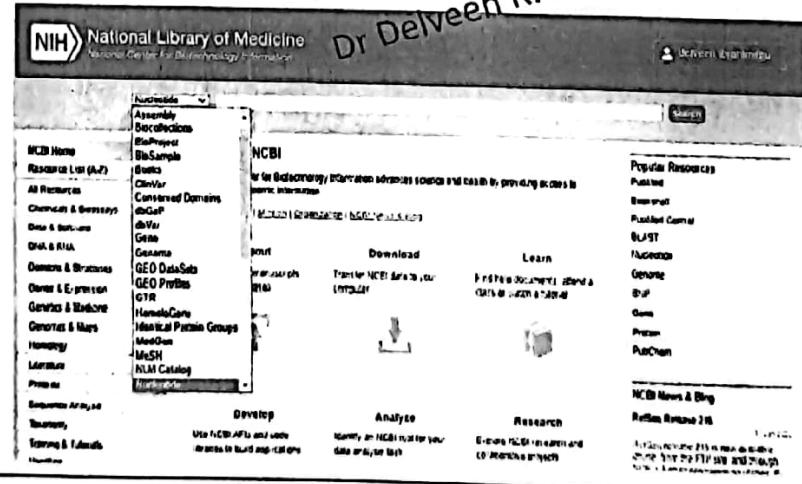
The corresponding sequence will be highlighted when the specific exon is clicked. These exon sequence can be marked in the original sequence pasted on word file

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



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

The screenshot shows a search interface for nucleotide sequences. The search term 'gyrA' has been entered. The results page displays several entries, with the top result being the 'Mycobacterium tuberculosis strain UKR100 GyrA (gyrA) gene, complete cds'. This entry includes details such as the accession number M399510, length 2517 bp, and a reference to a paper by Dr. Delveen R. Ibrahim et al. in the journal 'J. Clin. Microbiol.'.

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