Practical Session #1: Introduction to NCBI and Entrez databases.

I. Explore the National Center for Biotechnology information (NCBI) website and get familiar with its design and environment and its major databases.

https://www.ncbi.nlm.nih.gov/

What NCBI is? **National Center for Biotechnology Information**

How many databases are hosted by the NCBI? **37**

Which of the following NCBI databases could be considered as primary databases?
Protein(kind of both, includes from GenPept(primary), and RefSeq(Secondary)), **Nucleotide**(same case before), CDD, PubMed, Gene**(comes from Nucleotides)**, Genomes**(comes from Genes>Nucleotides)**, Refseq**(secondary comes from Nucleotides, Proteins)**, BioProjects.

II. The Entrez system

Perform a Global Query at the NCBI through the Entrez using the expression (all[Filter]). Which database contains the largest number of records? **Protein**

Now we are interested to find all the information at the NCBI related to the group of diseases in humans known as cancer. Type the word "cancer" in the search box on the NCBI homepage and run the search (Global Query). Note that the query is interpreted differently in different databases. How many scientific papers contain this word? How many nucleotide sequences?

Scientific papers: 4940163 (PubMed)

Nucleotide: 11008415

How many cancer-related functional genomics studies have been stored at NCBI? **45741 (BioProject)** Why does taxonomy database give us one record? (For discussion in class)

Because it refers to the name of a crab, not the illness

Perform a new Global Query but using the word "human". How many entries (records) have been obtained for the different databases?**There are lots of database results** Would we get the same results if we perform a Global Query using the search expression (homo sapiens), (human[organism]) or (homo sapiens[organism])? Why? How is the expression [organism] interpreted by each database?

Yes, because the [organism] tag specifies the search location, limiting results

If you are interested in studying human cancer, which of the following strategies would produce a more useful set of results in a Global Query at the NCBI?

cancer AND human

cancer[organism] AND human

cancer AND human[organism]

cancer OR human[organism]

III. At NCBI each record is assigned a UID "unique integer identifier" for internal tracking. In sequence databases this unique identifier is also known as the Accession number.

What NCBI database the following UIDs belong to?

CM000253.1 → **Nucleotide**

NG_011877.2 → **Nucleotide**

SRX4644664 → **Genome**

NP_002266.3 → **Protein**

CP027442.1 → **Nucleotide**

PRJNA490405 → **BioProject**

CAB37359.1 → **Protein**

ADE87724.1 → **Protein**

IV. Open the NCBI entry with accession number NG_011877 and get familiar with the format and the different fields used to store sequence information. This will open in the GenBank Flat File Format.

What does this entry represent? Do you think this entry provides cross-references (links) to other databases? **Yes** From which organism this sequence was obtained? **A homo sapiens** What is the UID or identifier for this organism in the Taxanomy database? **9606** What does the underscore "_" in the accession number stand for? Display the entry in FASTA format. What happened?

V. In the Taxonomy database explore all the information related to the organism *Homo sapiens*.

Look at the lineage for this taxon. What order do humans belong to? What is the txid for this mammalian order?

How many human protein sequences are there today at the NCBI?

VI. Advance searches

With which of these strategies will you find all the human sequences stored in the nucleotide database at the NCBI?

A. txid9606[Primary organism]

B. homo sapiens[Primary Organism]

C. homo sapiens[porgn]

D. human[porgn]

VI. Batch Entrez

Use Batch Entrez to upload a file of GIs or accession numbers from the Nucleotide or Protein databases, or upload a list of record identifiers from other Entrez databases. Batch Entrez will automatically download the corresponding records.

In this exercise we will retrieve from the NCBI database all sequences related to Homo sapiens tumor protein 53 (TP53) published on a paper with PubMed accession number PMC3675194. Find the flat text file in the course page with the list of numbers referenced in this paper.

1. Save the text file locally in your computer.
2. Open Batch Entrez.

https://www.ncbi.nlm.nih.gov/sites/batchentrez

3. Select the database from which the list of accessions will be queried.
4. Use the "Browse" button to select the filename containing the list of identifiers from your system directory.
5. After pressing the "Retrieve" button you will see a list of record summaries. Retrieve them!
6. Optionally, select a format in which to display the data for viewing, and/or saving. Select "Send to file" to save the file.

Try to solve this exercise using the ncbi web page and standard unix commands (grep, sort, less, wc, etc)

How many records are on the list?

From what database the entries belong to?

Do all entries represent human sequences?

Do all entries represent mRNA sequences?

Do all sequences belong to the same human subject?

Do all sequences have the same length?

If it does not work, you can obtain the same results with this link:
https://www.ncbi.nlm.nih.gov/nuccore/?term=KC820708:KC820786[pacc]

You can download all this sequences as fasta files using e-utilities, under linux / OSX / Cygwin, etc:

1) Download the PMC3675194-List_IDs.txt file
2) Go to the downloaded directory in a command line shell (eg: bash)
3) Execute the following 2 commands:
    a. dos2unix PMC3675194-List_IDs.txt
    b. cat PMC3675194-List_IDs.txt |xargs -tI% wget -O %.fasta "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=%&rettype=fasta&retmode=text"

Note that the command “b” starts with “cat” and ends with “text”. The command “a” is important, you should know why.