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? How many databases are hosted by the NCBI?

58 https://www.ncbi.nlm.nih.gov/guide/all/#databases

Which of the following NCBI databases could be considered as primary databases? Protein, Nucleotide, CDD, PubMed, Gene, Genomes, Refseq, BioProjects.

Protein and Nucleotide contain genbank which is primary, however they also contain RefSeq which is secondary, so we have to be carefull when obtaining data from it. Pubmed would be a special case since is not exactly "experimental data", but data is stored "as is" there is no postprocessing to it. BioproJect is difficult to classify It can contain a project that it is secondary database, but itself is apparently primary since data is stored as provided, however it can be modified "a posteriori"

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?

https://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers

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

https://www.ncbi.nlm.nih.gov/search/all/?term=all[Filter]

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.

https://www.ncbi.nlm.nih.gov/search/all/?term=cancer

How many scientific papers contain this word?
https://www.ncbi.nlm.nih.gov/pubmed/?term=cancer 3,980,604

How many nucleotide sequences?
https://www.ncbi.nlm.nih.gov/nuccore/?term=cancer 10985563

How many cancer-related genomics, functional genomics or genetics studies have been stored at NCBI?
https://www.ncbi.nlm.nih.gov/bioproject/?term=cancer 45203

Why does taxonomy database give us one record? (For discussion in class)
https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=6754&lvl=3&lin=f&keep=1&srchmode=1&unlock genus from the family *Cancridae*

Perform a new Global Query but using the word “human”.

How many entries (records) have been obtained for the different databases?
https://www.ncbi.nlm.nih.gov/search/all/?term=human

Would we get the same results if we perform a Global Query using the search expression (homo sapiens),
https://www.ncbi.nlm.nih.gov/search/all/?term=homo%20sapiens

(human[organism])
https://www.ncbi.nlm.nih.gov/search/all/?term=human%5Borganism%5D

or (homo sapiens[organism])?
https://www.ncbi.nlm.nih.gov/search/all/?term=homo+sapiens%5Borganism%5D

Why?
It looks like in some databases human is a clear alias of “homo sapiens”, but not in all of them.(eg: Pubchem databases)

When the [organism] is included, this term is only looked in the field “organism” of the database. And in some databases, there is a “controlled vocabulary” that takes care of the synonyms in this field.

How is the expression [organism] interpreted by each database? Are global search result numbers equal to the search numbers inside each database?

In some databases, it has no meaning and it is ignored, in others specifies where to search for this word or words. In the taxonomy database “human” exactly “homo sapiens” however “homo sapiens” matches three entries, ours and two others which are considered subspecies of “homo sapiens”

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]

Cancer AND human search both terms in any place we can find cases like: https://www.ncbi.nlm.nih.gov/nuccore/DM204094.1 But we have to be careful since cancer AND human[organism] it can also return results from other organisms related to human cancer (viruses or even mouse)

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 GeneBank Nucleotide

NG_011877.2 RefSeq Nucleotide

SRX4644664 SRA

NP_002266.3 Refseq protein

CP027442.1 GeneBank Nucleotide (take care with the genome entry)

PRJNA490405 BioProject

CAB37359.1 GeneBank protein

ADE87724.1 GeneBank 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 (ORGANISM, PUBMED, /db_xref , etc.) From which organism this sequence was obtained? 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? Only Refseq entries have "_" 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? Primates What is the txid for this mammalian order? 9443

How many human protein sequences are there today at the NCBI? 1 915 906

VI. Advance searches

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

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 download automatically 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? 79

From what database the entries belong to?nucleotide

We can download the results in the format "Summary" and use the downloaded file (nuccore_result.txt) to obtain the desired data.

Do all entries represent human sequences? Yes

```
grep "Homo sapiens" nuccore_result.txt |less -NS
```

Or use the left column, species section and add human.

Do all entries represent mRNA sequences? Yes

```
grep " bp " nuccore_result.txt |grep "mRNA"|less
```

Or just look at the left column

Do all sequences belong to the same human subject? No

Do all sequences have the same length? No

```
grep " bp " nuccore_result.txt |sort -n |less
```

Sort by sequence length and look at last page