Topic 1& 2

Created	@November 9, 2023 6:35 PM
Reviewed	☐

1. **Introduction: Organizing biological knowledge in databases.**

Tables:

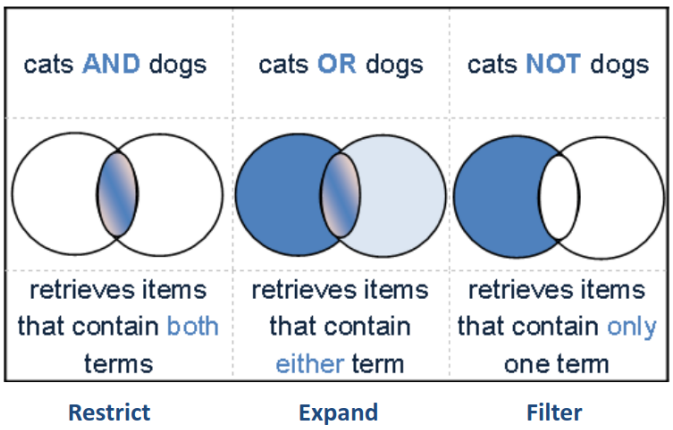
- Columns (fields):
  - Attributes of tables (ex. citation table, title, journal, etc)
- rows (records):
  - Actual data
  - Whereas fields describe what data is stored, the rows of a table are where the actual data is stored.

Documents:

- Key →value

Identifier: essentially a name of a database, table or table column. Unique identifier.

The primary key (accession number or accession code)



- To find all records for this species: txid10090[porgn]

    - [porgn] →Important to select sequences really belonging to this organism.
- To see how many eukaryotic species there are:
    - go to taxonomy home page and click on Statistics
    - Eukaryota[SubTree] AND species[Rank] AND specified[prop]
        - specified[prop] →Helps to restrict the output of this list to species with formal Linnean binomial names
- [Primary organism] = [porgn]
- When we have an asterisk means that it replaces 0 to n number of characters anywhere in the accession number the asterisk is it.
- If we have a file with the accession numbers:
    - If we want to know how many records are there listed:
        - less -NS 1.txt

2. **Biological raw data in public databanks and repositories for primary DNA sequences.**

Taxonomical divisions —> gbdiv_pri[PROP]

- PRI - primate sequences
- ROD - rodent sequences
- MAM - other mammalian sequences
- VRT - other vertebrate sequences
- INV - invertebrate sequences
- PLN - plant, fungal, and algal sequences
- BCT - bacterial sequences
- VRL - viral sequences
- PHG - bacteriophage sequences
- SYN - synthetic sequences

- UNA - unannotated sequences
- ENV - environmental sampling sequences

Others

- PAT - patent sequences

High-throughput or functional divisions and other sub-databases

(gbdiv_htg[PROP])

- EST – Expressed Sequence Tag (1st pass single read cDNA)
- GSS – Genome Survey Sequences (1st pass single read gDNA)
- HTG – High Throughput Genomic (incomplete sequences of genomic clones)
- STS – Sequence Tagged Site (PCR-based mapping)
- HTC - High-throughput cDNA

(”wgs”[Properties])

- TSA - Transcriptome shotgun data (transcriptome shotgun assembly projects)
- WGS - Whole genome shotgun data (whole genome shotgun projects)

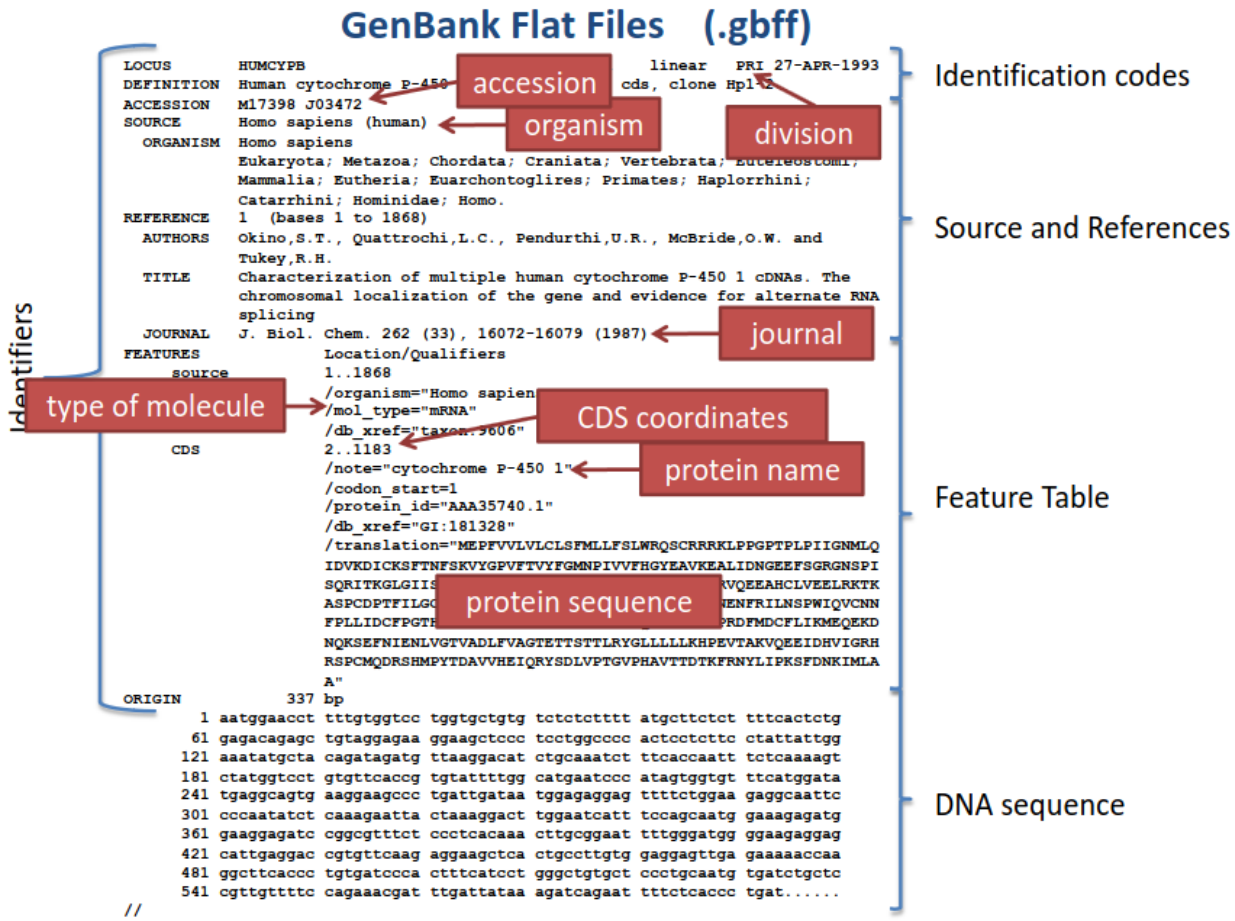
Others:

- CON- Contains no sequence data, but rather instructions for the assembly of reads.

ADVANCED SEARCHES —> Search field descriptions and tags at NCBI

[Accession]	[ACCN]	The accession number assigned by NCBI.
[All Fields]	[ALL]	All terms from all search fields in the database.
[Author]	[AU]	All authors from all references in the records.
[EC/RN Number]	[ECNO]	Enzyme Commission (EC) number for an enzyme activity.
[Feature Key]	[FKEY]	Biological features listed in the Feature Table of the sequence records.

[Gene Name]	[GENE]	Gene names annotated on database records.
[Issue]	[ISS]	The issue number of the journals cited on sequence records
[Journal]	[JOUR]	The name of the journals cited on sequence records.
[Keyword]	[KYWD]	Keywords applied by submitter
[Modification Date]	[MDAT]	The date of most recent modification of a sequence record.
[Organism]	[ORGN]	The scientific and common names for the complete taxonomy of organisms
[Properties]	[PROP]	Molecular type, source database, and other properties of the sequence
[Protein Name]	[PROT]	The names of protein products as annotated on sequence records.
[Publication Date]	[PDAT]	The date that records were made public in Entrez.
[Sequence Length]	[SLEN]	The total length of the sequence
[Text Word]	[WORD]	Text on a sequence record that is not indexed in other fields.
[Title]	[TI]	Words and phrases found in the title of the sequence record.



Examples:

- find in PubMed all the scientific papers containing the word “human” in the title and published in the last five years but are not review articles:
  - human[title] AND ("2018/09/21"[DP] : "2023/09/19"[DP]) NOT Review[PT]
- How many eukaryotic species (w/ formal names) are there currently in the taxonomy database?:
  - Eukaryota[SubTree] AND species[Rank] AND specified[prop]
- What is the total number of records for this taxon at the nucleotide database to this day?
  - txid10090[porgn] (10090 —> it is the identifier the taxid)
- find all the sequence from the house mouse (*Mus musculus*) stored in the nucleotide database at the NCBI :

    - txid10090[Primary organism]
    - mus musculus[Primary Organism]
    - mus musculus[porgn]
    - house mouse[porgn]
- What does the asterisk mean ? NC_0000*[Accession] AND Human[Organism]
    - The asterisk replaces 0 to n number of characters at the end of the accession number.
- If we have a list of accession numbers and we have to know some tips about it, we can use the commands:
    - `less -NS Acc_List.txt`
    - `grep -h \> *fasta|cut -d \[ -f 2|sort|uniq`
    - `grep -h \> *fasta|less -NS`
- Find sequences that are codifying at nucleotide database (CDS):
    - CDS[fkey]
    - nucleotide_protein[filter]
- Localize all of the records in the NCBI nucleotide database containing rabbit genomic sequences that are codifying.
    - rabbit[porgn] AND biomol_genomic[PROP] AND CDS[fkey]
- How many of these sequences come from the mitochondrial compartment:
    - rabbit[porgn] AND biomol_genomic[PROP] AND CDS[fkey] AND mitochondrion[filter]
- How many complete mammalian mitochondrial chromosomes can be found at NCBI nucleotide database:
    - mammalia[porgn] AND mitochondrion[filter] AND "complete genome"[TI]
- One of the proteins encoded in the mitochondrial chromosomes is cytochrome b. The best strategy to search for sequences encoding for cytochrome b proteins in NCBI nucleotide database is to include in a query the expression:
    - AND (cytochrome b[protein name])

- find out the NCBI genomic reference sequence encoding for human (us) cytochrome b protein.
    - "human"[Primary Organism] AND "cytochrome b"[protein name] AND refseq[filter]
- It is possible to find complete genome sequences in the NCBI nucleotide database for any organism. Create a search strategy trying to find at the nucleotide database all the complete or nearly complete genomic sequences for Human immunodeficiency virus type 1.
    - txid11676[Primary organism] AND "complete genome"[title]
    - Human immunodeficiency virus type 1 —> NC_001802.1 —> txid11676
- There are several genomes at the NCBI without annotations. That means none of the genetic elements like CDSs, genes or any other features annotated as “misc_feature”, have been indicated in the sequence. Search the records in the NCBI nucleotide database for HIV-1 (Human immunodeficiency virus type 1) complete genomes without any annotation.
    - txid11676[porgn] AND "complete genome"[title] NOT CDS[fkey] NOT Gene[fkey] NOT misc_feature[fkey]