Introduction to NCBI: sequence databases, sequence retrieval, sequence file formats

Author : Deepak Khatri

Sequence Databases

The main resources for storing and distributing sequence data are three large databases:

- the NCBI database (www.ncbi.nlm.nih.gov/)
- the European Molecular Biology Laboratory (EMBL) database (www.ebi.ac.uk/embl/)
- the DNA Database of Japan (DDBJ) database (www.ddbj.nig.ac.jp/).

These databases collect all publicly available DNA, RNA and protein sequence data and make it available for free. They exchange data nightly, so contain essentially the same data. Sequences in the NCBI Sequence Database (or EMBL/DDBJ) are identified by an accession number. This is a unique number that is only associated with one sequence, some additional *annotation* data, such as the name of the species it comes from, references to publications describing that sequence, etc. Some of this annotation data was added by the person who sequenced a sequence and submitted it to the NCBI database, while some may have been added later by a human curator working for NCBI. different type of database containes data from:

- nucleotide sequences
- protein sequences
- proteins sequence patterns or motifs
- macromolecular 3D structure
- gene expression data
- metabolic pathways–proteomics data

The NCBI database contains several sub-databases, the most important of which are:

- the NCBI Nucleotide database: contains DNA and RNA sequences
- the NCBI Protein database: contains protein sequences
- EST: contains ESTs (expressed sequence tags), which are short sequences derived from mRNAs
- the NCBI Genome database: contains DNA sequences for whole genomes
- PubMed: contains data on scientific publications

RefSeq

the NCBI database often contains redundant information for a gene, contains sequences of varying quality, and contains both uncurated and curated data.As a result, NCBI has made a special database called RefSeq (reference sequence database), which is a subset of the NCBI database. The data in RefSeq is manually curated, is high quality sequence data, and is non-redundant; this means that each gene (or splice-form of a gene, in the case of eukaryotes), protein, or genome sequence is only represented once.

Retrieve all sequences for an organism or taxon

You can follow this Youtube link other wise Starting with an organism or taxon name...

1. Search the Taxonomy database with the organism name. Accepted common names usually work at all taxonomic levels. Use the scientific name or formal name if no results are obtained with the common name.

2. Click on the desired taxon name in the results. For terminal taxa - generally subspecies, species, or strains - this link leads directly to the summary page. For higher taxa this link will lead to the Taxonomy Browser showing the lower taxa contained within the higher taxon.
3. If necessary, click on the desired taxon link in the Taxonomy Browser to reach the summary page.
4. The number of records in each database are linked in the Entrez records table on the taxon summary page . Click the linked number of records in the table to retrieve all records from the chosen sequence database (Nucleotide, Nucleotide EST, Nucleotide GSS, Protein).

Some Sequence Formats

1. EMBL/Swiss Prot

- The first line of each sequence entry is the ID definition line which contains entry name, dataclass, molecule, division and sequence length.
- XX line contains no data, just a separator
- The AC line lists the accession number.
- DE line gives description about the sequence
- FT precise annotation for the sequence
- Sequence information SQ in the first two spaces.•The sequence information begins on the fifth line of the sequence entry. •The last line of each sequence entry in the file is a terminator line which has the two characters // in the first two spaces.

2. FASTA

- A sequence in Fasta format begins with a single-line description, followed by lines of sequence data.
- The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column.
- It is recommended that all lines of text be shorter than 80 characters in length

3. GenBank/GenPept

- The nucleotide (GenBank) and protein (Gen Pept) database entries are available from Entrez in this format
- Can contain several sequences
- Onesequence starts with: “LOCUS”
- The sequence starts with: "ORIGIN“
- The sequence ends with: "//“

References:

1. https://a-little-book-of-r-for-bioinformatics.readthedocs.io/en/latest/src/chapter3.html
2. https://www.ncbi.nlm.nih.gov/guide/howto/retrieve-seq-org/
3. http://biopython.org/DIST/docs/tutorial/Tutorial.pdf
4. https://sta.uwi.edu/fst/dms/icgeb/documents/18-01-10SequenceformatsanddatabasesinbioinformaticsDGL1.pdf