GenBank (www.ncbi.nlm.nih.gov/genbank/)

It is located in the USA. NCBI since 1992 has provided access to GenBank DNA sequence database through NCBI gateway server freely. The three nucleotide sequence databases GenBank, European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ) coordinate among themselves so that all three of them are updated with the latest findings.

A detailed structure of a nucleotide sequence file format in this database includes the following:

1. **Locus:** This can be defined as a title given by GenBank itself to name the sequence entry. It includes the following:
    a. **Locus Name:** Similar to accession number for the sequence.
    b. **Sequence Length:** Tells the number of bases existing in the sequence.
    c. **Molecule-Type:** Identifies the type of nucleic acid sequence. The various types are mRNA (which is present as cDNA), rRNA, snRNA, and DNA.
    d. **GB Division:** Postulates class of the data according to classification criteria of GenBank.
    e. **Modification Date:** The date on which the record was modified.
2. **Definition:** This denotes the name of the nucleotide sequence.
3. **Accession:** This covers accession number, accession version, and GI number. Accession number can be defined as the unique identifier associated with each nucleotide sequence present in the database. If more than one record is created for a particular sequence then it will have the same accession number but all records will have different versions associated with that accession number.
4. **Keyword:** Defined words that were used to index the entries.
5. **The Source:** This describes organism from which sequences have been obtained. The accepted common name is mentioned first and then the scientific name is mentioned. In the end, the taxonomic lineage according to GenBank is specified.
6. **The Citation:** Includes the journal from which with the sequence was derived as initially the sequences were obtained only from published literature.
7. **Features:** These consist of the information derived from the sequence such as biological source, coding region, exon, intron, promoters, alternate splice patterns, mutations, etc.
8. **Sequence:** Contains the following:
    a. Count of presence of each nucleotide in the sequence,
    b. Whole nucleotide sequence,
    c. Beginning of sequence is determined by keyword “ORIGIN”, and
    d. End is marked as “\\”.

There are many techniques for retrieving and searching data from GenBank.

1 The sequence identifiers can be searched in GenBank along with Entrez Nucleotide.

2 Using BLAST search and then aligning nucleotide sequences to the query sequence.

3 To search the appropriate link and then download nucleotide sequences. (www.ncbi.nlm.nih.gov/genbank/).

Record structure: sequence

Genbank divisions

PRI: primate sequences
ROD: rodent sequences
MAM: other mammalian sequences
VRT: other vertbrate 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
EST: expressed sequence tags
PAT: patent sequences
STS: sequence tag sites
GSS: genome survey sequences
HTC: high throughput cDNA sequences
HTG: high throughput genomic sequences