Nucleic acid Sequence databases

Dr. Maya C. Nair
Govt. Sanskrit College, Pattambi, Kerala, India
Nucleic acid sequence databases

- Databases which combine information on nucleotide sequences
- There are different nucleic acid databases
  - EMBL
  - GenBank
  - DDBJ
  - GSDB
  - TRANSFAC
  - GeneCards
EMBL

- Europe’s primary nucleotide sequence resource
- From European Bioinformatics Institute (EBI)
- Main source for DNA and RNA sequences
- Sequences are either direct submissions from individuals, genome sequencing groups, or from scientific literature and patent applications
- Database is produced in collaboration with GenBank and DDBJ
- Each of the three groups collects a portion of the total sequence data reported worldwide
- All new and updated database entries are exchanged between the groups on a daily basis
EMBL (contd..)

- Information can be retrieved from EMBL using the SRS Sequence Retrieval System
- This links the principal DNA and protein sequence databases with motif, structure, mapping and other specialist databases
- EMBL may be searched with query sequences via the EBI’s Web interfaces to the BLAST and FastA programs
GenBank

- Is the DNA database from the National Center for Biotechnology Information (NCBI)
- It is an annotated collection of all publicly available DNA sequences
- Incorporates sequences from publicly available sources, primarily from direct author submissions and large scale sequencing projects
- Exchanges data with EMBL and DDBJ
- Diversity and enormous size of GenBank have made it to split into smaller discrete divisions
GenBank (contd..)

- Artificial separation into discrete divisions facilitates, fast, specific searches by restricting queries to particular database subsets
- Information can be retrieved from GenBank using Entrez integrated retrieval system
- This combines data from the principal DNA and protein sequence databases with information from genome maps and protein structures
- GenBank may be searched with user query sequences by means of BLAST programme
Structure of GenBank entries

• A GenBank release includes
  • the sequence files
  • indices created on various database fields (e.g., author index, reference index etc.)
  • information derived from the database

• Most commonly used is the sequence entry file
  – it contains the sequence itself and descriptive information relating to it
Structure of GenBank entry

• Each entry consists of a number of key words, relevant associated sub-key words and an optional Feature Table

• its end is indicated by a // terminator

❖ Positioning

❖ Keywords begin in column 1
❖ sub-key words begin in column 3
❖ a code defining part of the Feature Table begins in column 5
❖ any line beginning with a blank character is considered a continuation from the key word or sub key word
Key words

- **LOCUS** – introduces a short label for the entry that may suggest the function of the sequence
  - This line also summarizes,
    - the number of bases
    - source of sequence data
    - section of data base
    - date of submission

- **DEFINITION** - this line contains concise description of the sequence

- **ACCESSION** – gives the accession number, a unique constant code assigned to each entry

- **NID** - supplies a nucleotide identifier
• **KEYWORDS** – introduces a list of short phrases, describing gene products and other relevant information about the entry

• **SOURCE** - provides information on the tissue from which the data have been derived. The sub key word ORGANISM illustrates the biological classification

• **REFERENCE** – records indicate the portion of sequence data to which the cited literature refers
  
  – sub Key words AUTHORS, TITLE and JOURNAL provide a structure for the citation
  
  – MEDLINE sub-key word is a pointer to an online medical literature information resource, which allows the abstract of a given article to be viewed
• **FEATURES** – introduces a section that has its own structure and its purpose is to describe properties of the sequence in detail (this is the feature table)
  
  – within the table, links to other databases are made through the ‘/db_xref’ qualifier
  
  – co-ordinates are given for 5’ – untranslated region, for the coding sequence, for the 3’ – untranslated region, the polyadenylation sequence etc.
  
  – protein translation and location of the signal and mature peptides are also given
• **BASE COUNT** – details the frequency of occurrence of the different base types in the sequence

• **ORIGIN** – indicates the location of the first base of the sequence within the genome

  The nucleotide sequence follows and the entry is terminated by the // marker
Other Nucleic acid databases

- **GSDB – Genome Sequence DataBase** One of the four public nucleotide databases worldwide. This is made available through the website of the National Centre for Genome Resources.

- **TRANSFAC** – is a transcription factor database at the Gesellschaft für Biotechnologische Forschung mbH, Germany. It compiles data about gene regulatory DNA sequences and protein factors binding to them.

- **GENECARDS** – is a database of human genes, their products and their involvement in diseases.

- **DDBJ** – is the DNA database of Japan, began in 1986 in collaboration with EMBL and GenBank. Database is produced, maintained and distributed at the National Institute of Genetics, Japan.
Genomic Resources
(Boutique databases)
GENOME DATABASES

• The purpose of these databases is to bring a focus
  – to species-specific genomics
  – to particular sequencing techniques
    • Eg. *Saccharomyces cerevisiae* (SGD)
      – *Caenorhabditis elegans* (ACeDB)
      – *Drosophila melanogaster* (FlyBase)
      – *Rat* (RGD)
      – Human (GDB)
UniGene

- provide a transcript map by sets of non-redundant gene-oriented clusters derived from GenBank sequences
- Collection represents genes from many organisms
- HGI (section 5.9.3) represents only human genes
- Each cluster relate to a unique gene and include related information (such as tissue type in which the gene is expressed, map location etc.)
- UniGene has been used by experimentalists to select reagents for gene mapping projects
- Also used for large-scale gene expression analysis
- This resource can be accessed via NCBI’s home page