BIF501 – BIOINFORMATICS-2
Assignment No. 1
Fall 2021
Total Marks: 5
Due date: 16-12-2021

Instructions:

- **Make sure that you upload the solution file before due date. No assignment will be accepted through e-mail after the due date.**
- The assignment should consist of at least 800 words.
- Use the font style "Times New Roman" and font size "12".
- Compose your document in MS-Word, any file created in any formatwill not be accepted and marked zero.
- Use black and blue font colors only.
- Solution guidelines
- To solve this assignment, you should have good command over lectureson nucleotide and proteins databases.
- This is not a group assignment, it is an individual assignment so be careful and avoid copying others' work
- Give the answer according to question only and avoid irrelevant details.

Please note that your assignment will not be graded if:

- It is submitted after due date
- The file you uploaded does not open
- The file you uploaded is copied from someone else
- It is in some format other than .doc
- Cheating or copying/plagiarism of assignment is strictly prohibited. The cheated or plagiarized assignment will be marked 'Zero'.

Question
Give a comprehensive account of nucleotide and protein databases available on internet.

Primary databases of Protein
The PRIMARY databases hold the experimentally determined protein sequences inferred from the conceptual translation of the nucleotide sequences. This, of course, is not experimentally derived information, but has arisen as a result of interpretation of the nucleotide sequence information and consequently must be treated as potentially containing misinterpreted information. There is a number of primary protein sequence databases and each requires some specific consideration.
a. Protein Information Resource (PIR) – Protein Sequence Database (PIR-PSD):

- The PIR-PSD is a collaborative endeavor between the PIR, the MIPS (Munich Information Centre for Protein Sequences, Germany) and the JIPID (Japan International Protein Information Database, Japan).
- The PIR-PSD is now a comprehensive, non-redundant, expertly annotated, object-relational DBMS.
- A unique characteristic of the PIR-PSD is its classification of protein sequences based on the superfamily concept.
- The sequence in PIR-PSD is also classified based on homology domain and sequence motifs.
- Homology domains may correspond to evolutionary building blocks, while sequence motifs represent functional sites or conserved regions.
- The classification approach allows a more complete understanding of sequence function-structure relationship.

b. SWISS-PROT

- The other well-known and extensively used protein database is SWISS-PROT. Like the PIR-PSD, this curated proteins sequence database also provides a high level of annotation.
- The data in each entry can be considered separately as core data and annotation.
- The core data consists of the sequences entered in common single letter amino acid code, and the related references and bibliography. The taxonomy of the organism from which the sequence was obtained also forms part of this core information.
- The annotation contains information on the function or functions of the protein, post-translational modification such as phosphorylation, acetylation, etc., functional and structural domains and sites, such as calcium binding regions, ATP-binding sites, zinc fingers, etc., known secondary structural features as for examples alpha helix, beta sheet, etc., the quaternary structure of the protein, similarities to other protein if any, and diseases that may arise due to different authors publishing different sequences for the same protein, or due to mutations in different strains of an described as part of the annotation.

TrEMBL (for Translated EMBL) is a computer-annotated protein sequence database that is released as a supplement to SWISS-PROT. It contains the translation of all coding sequences present in the EMBL Nucleotide database, which have not been fully annotated. Thus it may contain the sequence of proteins that are never expressed and never actually identified in the organisms.

c. Protein Databank (PDB):

- PDB is a primary protein structure database. It is a crystallographic database for the three-dimensional structure of large biological molecules, such as proteins.
- In spite of the name, PDB archive the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as antibiotic gramicidin and complexes of protein and nucleic acids.

- The database holds data derived from mainly three sources: Structure determined by X-ray crystallography, NMR experiments, and molecular modeling.

Secondary Databases of Protein

The secondary databases are so termed because they contain the results of analysis of the sequences held in primary databases. Many secondary protein databases are the result of looking for features that relate different proteins. Some commonly used secondary databases of sequence and structure are as follows:

a. PROSITE:

- A set of databases collects together patterns found in protein sequences rather than the complete sequences. PROSITE is one such pattern database.
- The protein motif and pattern are encoded as "regular expressions".
- The information corresponding to each entry in PROSITE is of the two forms – the patterns and the related descriptive text.

b. PRINTS:

- In the PRINTS database, the protein sequence patterns are stored as 'fingerprints'. A fingerprint is a set of motifs or patterns rather than a single one.
- The information contained in the PRINT entry may be divided into three sections. In addition to entry name, accession number and number of motifs, the first section contains cross-links to other databases that have more information about the characterized family.
- The second section provides a table showing how many of the motifs that make up the fingerprint occurs in the how many of the sequences in that family.
- The last section of the entry contains the actual fingerprints that are stored as multiple aligned sets of sequences, the alignment is made without gaps. There is, therefore, one set of aligned sequences for each motif.

c. MHCPep:

- MHCPep is a database comprising over 13000 peptide sequences known to bind the Major Histocompatibility Complex of the immune system.
- Each entry in the database contains not only the peptide sequence, which may be 8 to 10 amino acid long but in addition has information on the specific MHC molecules to which it binds, the experimental method used to assay the peptide, the degree of activity and the binding affinity observed , the source protein that, when broken down gave rise to this peptide along with other, the positions along the peptide where it anchors on the MHC molecules and references and cross-links to other information.

d. Pfam

- Pfam contains the profiles used using Hidden Markov models.

- HMMs build the model of the pattern as a series of the match, substitute, insert or delete states, with scores assigned for alignment to go from one state to another.
- Each family or pattern defined in the Pfam consists of the four elements. The first is the annotation, which has the information on the source to make the entry, the method used and some numbers that serve as figures of merit.
- The second is the seed alignment that is used to bootstrap the rest of the sequences into the multiple alignments and then the family.
- The third is the HMM profile.
- The fourth element is the complete alignment of all the sequences identified in that family.

Primary databases of nucleotide sequences

- There are three chief databases that store and make available raw nucleic acid sequences to the public and researchers alike: **GenBank, EMBL, DDBJ**.
- They are referred to as the primary nucleotide sequence databases since they are the repository of all nucleic acid sequences.
- GenBank is physically located in the USA and is accessible through the NCBI portal over the intern.
- EMBL (European Molecular Biology Laboratory) is in UK and DDJB (DNA databank of Japan) is in Japan.
- All three accept nucleotide sequence submissions and then exchange new and updated data on a daily basis to achieve optimal synchronization between them.
- These three databases are primary databases, as they house original sequence data.
- They collaborate with Sequence Read Archive (SRA), which archives raw reads from high-throughput sequencing instruments.

a. GenBank

The GenBank sequence database is open access, annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced and maintained by the National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration (INSDC), receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank has become an important database for research in biological fields and has grown in recent years at an exponential rate by doubling roughly every 18 months.

b. EMBL (European Molecular Biology Laboratory)

The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database is a comprehensive collection of primary nucleotide sequences maintained at the European Bioinformatics Institute (EBI). Data are received from genome sequencing centers, individual scientists and patent offices.

c. DDBJ (DNA databank of Japan)

It is located at the National Institute of Genetics (NIG) in the Shizuoka prefecture of Japan. It is the only nucleotide sequence data bank in Asia. Although DDBJ mainly receives its data from Japanese researchers, it can accept data from contributors from any other country.

2. Secondary databases of nucleotide sequences

- Many of the secondary databases are simply sub-collection of sequences culled from one or the other of the primary databases such as GenBank or EMBL.
- There is also usually a great deal of value addition in terms of annotation, software, presentation of the information and the cross-references.
- There are other secondary databases that do not present sequences at all, but only information gathered from sequences databases.

a. Omniome Database:
Omniome Database is a comprehensive microbial resource maintained by TIGR (The Institute for Genomic Research). It has not only the sequence and annotation of each of the completed genomes, but also has associated information about the organisms (such as taxon and gram stain pattern), the structure and composition of their DNA molecules, and many other attributes of the protein sequences predicted from the DNA sequences.
It facilitates the meaningful multi-genome searches and analysis, for instance, alignment of entire genomes, and comparison of the physical proper of proteins and genes from different genomes etc.

b. FlyBase Database:
A consortium sequenced the entire genome of the fruit fly *D. Melanogaster* to a high degree of completeness and quality.

c. ACeDB:
It is a repository of not only the sequence but also the genetic map as well as phenotypic information about the *C. Elegans* nematode worm.