**Course: Advanced Bioinformatics**

**Module title: Features and Functions**

**Module no. : 210**

**Data Formats and Databases:** The toolbox accesses many of the databases on the Web and other online data sources. It allows you to copy data into the MATLAB Workspace, and read and write to files with standard bioinformatic formats. It also reads many common genome file formats, so that you do not have to write and maintain your own file readers.

**Web-based databases** — You can directly access public databases on the Web and copy sequence and gene expression information into the MATLAB environment.

The sequence databases currently supported are GenBank (getgenbank), GenPept (getgenpept), European Molecular Biology Laboratory (EMBL) (getembl), and Protein Data Bank (PDB) (getpdb). You can also access data from the NCBI Gene Expression Omnibus (GEO) Web site by using a single function (getgeodata).

Get multiply aligned sequences (gethmmalignment), hidden Markov model profiles (gethmmprof), and phylogenetic tree data (gethmmtree) from the PFAM database.

Gene Ontology database — Load the database from the Web into a gene ontology object (geneont). Select sections of the ontology with methods for the geneont object (geneont.getancestors, geneont.getdescendants, geneont.getmatrix, geneont.getrelatives), and manipulate data with utility functions (goannotread, num2goid).

Read data from instruments — Read data generated from gene sequencing instruments (scfread, joinseq, traceplot), mass spectrometers (jcampread), and Agilent® microarray scanners (agferead).

Reading data formats — The toolbox provides a number of functions for reading data from common bioinformatic file formats.

* Sequence data: GenBank (genbankread), GenPept (genpeptread), EMBL (emblread), PDB (pdbread), and FASTA (fastaread)
* Multiply aligned sequences: ClustalW and GCG formats (multialignread)
* Gene expression data from microarrays: Gene Expression Omnibus (GEO) data (geosoftread), GenePix® data in GPR and GAL files (gprread, galread), SPOT data (sptread), Affymetrix® GeneChip® data (affyread), and ImaGene® results files (imageneread)
* Hidden Markov model profiles: PFAM-HMM file (pfamhmmread)

Writing data formats — The functions for getting data from the Web include the option to save the data to a file. However, there is a function to write data to a file using the FASTA format (fastawrite).

BLAST searches — Request Web-based BLAST searches (blastncbi), get the results from a search (getblast) and read results from a previously saved BLAST formatted report file (blastread).

The MATLAB environment has built-in support for other industry-standard file formats including Microsoft Excel and comma-separated-value (CSV) files. Additional functions perform ASCII and low-level binary I/O, allowing you to develop custom functions for working with any data format.