<u>**Short course on Bioinformatics**</u>

Major Bioinformatics Resources: NCBI, EBI, ExPASy, RCSB:

- The knowledge of various databases and bioinformatics tools available at these resources, organization of databases: data contents and formats, purpose and utility in Life Sciences

Open access bibliographic resources and literature databases:

- Open access bibliographic resources related to Life Sciences viz., PubMed, BioMed Central, Public Library of Sciences (PloS)

Sequence databases:

- Formats, querying and retrieval
- Nucleic acid sequence databases: GenBank, EMBL, DDBJ;
- Protein sequence databases: Uniprot-KB: SWISS-PROT, TrEMBL, PIR-PSD
- Repositories for high throughput genomic sequences: EST, STS GSS, etc.;
- Genome Databases at NCBI, EBI, TIGR, SANGER
- Viral Genomes
- Archeal and Bacterial Genomes;
- Eukaryotic genomes with special reference to model organisms (Yeast, Drosophila, C. elegans, Rat, Mouse, Human, plants such as Arabidopsis thaliana, Rice, etc.)

3D Structure Database: PDB, NDB

- Chemical Structure database: Pubchem
- Gene Expression database: GEO, SAGE

Derived Databases:

- Knowledge of the following databases with respect to: basic concept of derived databases, sources of primary data and basic principles of the method for deriving the secondary data, organization of data, contents and formats of database entries, identification of patterns in given sequences and interpretation of the same
- Sequence: InterPro, Prosite, Pfam, ProDom, Gene Ontology
- Structure classification database: CATH,SCOP, FSSP
- Protein-Protein interaction database: STRING

Compilation of resources:

- NAR Database and Web server Issues and other resources published in Bioinformatics related journals

Sequence Analysis:

File formats:

- Various file formats for bio-molecular sequences: GenBank, FASTA, GCG, MSF etc

Basic concepts:

- Sequence similarity, identity and homology, definitions of homologues, orthologues, paralogues

Scoring matrices:

- basic concept of a scoring matrix, Matrices for nucleic acid and proteins sequences, PAM and BLOSUM series, principles based on which these matrices are derived

Pairwise sequence alignments:

- Basic concepts of sequence alignment: local and global alignments, Needleman and Wunsch, Smith and Waterman algorithms for pairwise alignments, gap penalties, use of pairwise alignments for analysis of Nucleic acid and protein sequences and interpretation of results.

Multiple sequence alignments (MSA):

- The need for MSA, basic concepts of various approaches for MSA (e.g. progressive, hierarchical etc.). Algorithm of CLUSTALW and PileUp and their application for sequence analysis (including interpretation of results), concept of dandrogram and its interpretation

Database Searches:

- Keyword-based searches using tools like ENTREZ and SRS
- Sequence-based searches: BLAST and FASTA

Sequence patterns and profiles:

- Basic concept and definition of sequence patterns, motifs and profiles, various types of pattern representations viz. consensus, regular expression (Prosite-type) and sequence profiles; profile-based database searches using PSI-BLAST, analysis and interpretation of profile-based searches

Taxonomy and phylogeny:

- Basic concepts in systematics, taxonomy and phylogeny; molecular evolution; nature of data used in Taxonomy and Phylogeny, Definition and description of phylogenetic trees and various types of trees

Protein and nucleic acid properties:

- Computation of various parameters using proteomics tools at the ExPASy server and EMBOSS

Comparative genomics:

- Basic concepts and applications, whole genome alignments: understanding significance. Artemis as an example

Sequence analysis: Scoring matrices:

- Detailed method of derivation of the PAM and BLOSUM matrices

Pairwise sequence alignments:

- Needleman and Wuncsh, Smith and Waterman algorithms and their implementation

Multiple sequence alignments (MSA):

Use of HMM-based Algorithm for MSA (e.g. SAM method)
Sequence patterns and profiles:

- **Repeats:** Tandem and Interspersed repeats, repeat finding, Motifs, consensus, position weight matrices
- Algorithms for derivation of and searching sequence patterns: MEME, PHI-BLAST, SCanProsite and PRATT
- Algorithms for generation of sequence profiles: Profile Analysis method of Gribskov, HMMer, PSI-BLAST

Protein and nucleic acid properties: e.g. Proteomics tools at the ExPASy server and EMBOSSTaxonomy and phylogeny:

- Phylogenetic analysis algorithms such as maximum Parsimony, UPGMA, Transformed Distance, Neighbors-Relation, Neighbor-Joining, Probabilistic models and associated algorithms such as Probabilistic models of evolution and maximum likelihood algorithm, Bootstrapping methods, use of tools such as Phylip, Mega, PAUP
- Analysis of regulatory RNAs: Databases and tools

  Fee: Rs 40,000