M. Sc. (Biochemistry)
Semester- III

Course Code	PS03EBIC22	Title of the Course	Bioinformatics
Total Credits of the Course	4	Hours per Week	4

Main Focus of the Course outcomes	Employability	Skill Development	Entrepreneurship
	√	√	

Course Objectives:	1. To get knowledge and awareness of the basic principles and concepts of biology, computer science and mathematics 2. To explore existing software effectively to extract information from large databases and to use this information in computer modelling 3. To get problem-solving skills, including the ability to develop new algorithms and analysis methods. 4. To train student for understanding of the intersection of life and information sciences, the core of shared concepts, language and skills the ability to speak the language of structure-function relationships, information theory, gene expression, and database queries.

Course Content		
Unit	Description	Weightage* (%)
1.	Introduction to Bioinformatics: _ Overview, Internet and bioinformatics, Applications. _ Introduction and Bioinformatics Resources: _ Knowledge of various databases and bioinformatics tools available at theseresources, the major content of the databases, Literature databases: _ Nucleic acid sequence databases: GenBank, EMBL, DDBJ _ Protein sequence databases: SWISS-PROT, TrEMBL, PIR, PDB, SCOP, CATH_ Genome Databases at NCBI, EBI, TIGR, SANGER _ Other Databases of Patterns/Motifs/System Biology (Gene and proteinnetwork database and resources) _ Sequence analysis: _ Various file formats for bio-molecular sequences: genbank, fasta, gcg, msf, nbrfpiretc. _ Basic concepts of sequence similarity, identity and homology, Definitions ofhomologues, orthologues, paralogues, xenologus. _ Scoring matrices: basic concept of a scoring matrix, PAM and BLOSUM series. _ Sequence-based Database Searches: what are sequence-based database searches, BLAST and FASTA algorithms, various versions of basic BLAST and	25%

	FASTA. _ Pairwise and Multiple sequence alignments: basic concepts of sequence alignment, Needleman &Wuncsh, Smith & Waterman algorithms for pairwisealignments, Progressive and hierarchical algorithms for MSA. _ Use of pairwise alignments and Multiple sequence alignment for analysis ofNucleic acid and protein sequences and interpretation of results.	
2.	Gene prediction: _ Gene structure in Prokaryotes and Eukaryotes, Gene prediction methods: NeuralNetworks, Pattern Discrimination methods, Signal sites Predictions, Evaluation ofGene Prediction methods. _ Computational RNA Structure analysis: _ Secondary and tertiary structure of RNA. Various algorithms of RNA folding andtheir analysis. Energy minimization in RNA folding. RNA sequence alignmentbased on secondary structure and its applications in functional genomics andphylogeny. _ Transcriptomics: _ Complete transcript cataloguing and gene discovery sequencing based approach,Microarray based technologies and computation based technologies	25%
3.	**Genomics**:Concepts and tools for genomics and comparative Genomics _ Ancient conserved regions _ Horizontal gene transfer _ Functional classification of genes _ Gene order (synteny) is conserved on chromosomes of related organisms.Prediction of gene function based on a composite analysis. Functional genomics. Putting together all of the information into a genome database. _ **Phylogenetic analysis**: _ Definition and description of phylogenetic trees and various types of trees, Molecular basisof evolution, Method of construction of Phylogenetic trees: Distance based method(UPGMA, NJ), Character Based Method (Maximum Parsimony and Maximum Likelihoodmethod).	25%
4.	Proteomics and Protein Computational Biology: _ Tools for proteomics: Acquisition of protein structure information, databases and applications. _ Structural classification of proteins, Protein structure analysis structure alignment and comparison,	25%

	_ Secondary structure and evaluation: algorithms of Chou Fasman,GOR methods. _ Tertiary Structure: basic principles and protocols, Methods to study 3D structure.prediction of specialized structures. _ Active site prediction, Protein folding, Protein modeling and drug design _ Protein structure comparison and classification: _ Classes, folds, motif, domain; the concepts in 3D structure comparison, purpose ofstructure comparison, algorithms such as FSSP, VAST and DALI. Principles ofprotein folding and methods to study protein folding.	

Teaching-Learning Methodology	Online / Offline / Presentation / Videos	

Evaluation Pattern		
Sr. No.	Details of the Evaluation	Weightage
1.	Internal Written / Practical Examination (As per CBCS R.6.8.3)	15%
2.	Internal Continuous Assessment in the form of Practical, Viva-voce, Quizzes, Seminars, Assignments, Attendance (As per CBCS R.6.8.3)	15%
3.	University Examination	70%

Course Outcomes: Having completed this course, the learner will be able to	
1.	To get introduced to the basic concepts of Bioinformatics and its significance in Biological data analysis.
2.	To get introduced to the basics and advance of sequence alignment and analysis.
3.	To get overview about biological macromolecular structures and structure prediction methods.
4.	To understand the structural organisation, structural properties and various techniques employed in the structure determination of Biological macromolecules – DNA& Protein.
5.	To get exposed to computational methods, tools and algorithms employed for Biological Data Interpretation.
6.	To have hands on training on various computational tools and techniques employed in Biological sequence analysis.

7.	To get exposed to various tools and methodologies used in multiple sequence alignment, phylogenetic analysis and genetic diversity analysis observed in biological sequences.
8.	To impart knowledge on chemical databases, various advanced techniques and tools like docking, QSAR studies etc employed in computational drug discovery.
9.	To get knowledge about various approaches in genome sequencing and NGS.

Suggested References:

Sr. No.	References
1.	Bioinformatics: A Beginners Guide, Clavarie and Notredame
2.	Bioinformatics: David Mount
3.	Bioinformatics: Rastogi
4.	Introduction to Bioinformatics: Arthur M. Lesk
5.	Bioinformatics: Principles and applications, Ghosh and Mallick
6.	Bioinformatics: Genes, Proteins and Computer, C A Orengo
7.	Protein Structure Prediction: Methods and Protocols, Webster, David (Southern Cross Molecular Ltd., Bath, UK)

On-line resources to be used if available as reference material

On-line Resources

Nucleotide Sequence Databases (the principal ones)

- NCBI - National Center for Biotechnology Information
- EBI - European Bioinformatics Institute
- DDBJ - DNA Data Bank of Japan

Protein Sequence Databases

- SWISS-PROT &TrEMBL - Protein sequence database and computer annotated supplement
- **UniProt** - UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.
- PIR - Protein Information Resource
- MIPS - Munich Information centre for Protein Sequences

- HUPO - HUman Proteome Organization

Database Searching by Sequence Similarity

- BLAST @ NCBI
- PSI-BLAST @ NCBI
- FASTA @ EBI
- BLAT Jim Kent's Blat is just superb in terms of speed and the integrated view you get for viewing the results

Sequence Alignment

- USC Sequence Alignment Server - align 2 sequences with all possible varieties of dynamic programming
- T-COFFEE - multiple sequence alignment
- ClustalW @ EBI - multiple sequence alignment
- MSA 2.1 - optimal multiple sequence alignment using the Carrillo-Lipman method
- BOXSHADE - pretty printing and shading of multiple alignments
- Splign - Splign is a utility for computing cDNA-to-Genomic, or spliced sequence alignments. At the heart of the program is a global alignment algorithm that specifically accounts for introns and splice signals.
- Spidey - an mRNA-to-genomic alignment program

Protein Domains: Databases and Search Tools

- InterPro - integration of Pfam, PRINTS, PROSITE, SWISS-PROT + TrEMBL
- PROSITE - database of protein families and domains
- Pfam - alignments and hidden Markov models covering many common protein domains
- SMART - analysis of domains in proteins
- ProDom - protein domain database
- PRINTS Database - groups of conserved motifs used to characterise protein families
- Blocks - multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins

Protein 3D Structure

- PDB - protein 3D structure database
- RasMol / Protein Explorer - molecule 3D structure viewers
- SCOP - Structural Classification Of Proteins
- UCL BSM CATH classification
- The DALI Domain Database
- FSSP - fold classification based on structure-structure alignment of proteins
- SWISS-MODEL - homology modeling server
- Structure Prediction Meta-server
- K2 - protein structure alignment
- DALI - 3D structure alignment server
- DSSP - defines secondary structure and solvent exposure from 3D coordinates
- HSSP Database - Homology-derived Secondary Structure of Proteins
- PredictProtein& PHD - predict secondary structure, solvent accessibility, transmembrane helices, and other stuff
- Jpred2 - protein secondary structure prediction
- PSIpred (& MEMSAT &GenTHREADER) - protein secondary structure prediction (& transmembrane helix prediction & tertiary structure prediction by threading)

Phylogeny & Taxonomy

- The Tree of Life
- Species 2000 - index of the world's known species
- TreeBASE - a database of phylogenetic knowledge
- PHYLIP - package of programs for inferring phylogenies
- TreeView - user friendly tree displaying for Macs & Windows

Gene Prediction

- Genscan - eukaryotes
- GeneMark
- Genie - eukaryotes
- GLIMMER - prokaryotes
- tRNAscan - SE 1.1 - search for tRNA genes in genomic sequence
- GFF (General Feature Format) Specification - a standard format for genomic sequence annotation

Metabolic, Gene Regulatory & Signal Transduction Network Databases

- KEGG - Kyoto Encyclopedia of Genes and Genomes
- BioCarta
- **DAVID** - **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery - A useful server to for annotating microarray and other genetic data.
- stke - Signal Transduction Knowledge Environment
- BIND - Biomolecular Interaction Network Database
- EcoCyc
- WIT
- **PathGuide** A very useful collection of resources dealing primarily with pathways
- SPAD - Signaling Pathway Database
- CSNDB - Cell Signalling Networks Database
- PathDB
- Transpath
- DIP - Database of Interacting Proteins
- PFBP - Protein Function and Biochemical Networks

*****