Indian Institute of Information Technology, Allahabad
Department of Applied Sciences
Course Syllabus Format
MDM: Biological Data Analytics

1. **Name of the Course:** Proteomics and Genomics

2. **LTPS structure of the course:** 3-0-0-0

3. **Objective of the course**

- To understand the structure, organization, and functions of nucleic acids and proteins in biological systems.
- To explain the molecular mechanisms of DNA replication, transcription, translation, and regulation of gene expression relevant to genomics studies.
- To develop fundamental knowledge of protein structure, protein synthesis, and biomolecular interactions associated with proteomics and modern biological research.

Course plan:

Component	Unit	Topics for Coverage	Chapter No
Component 1 (Mid-sem)	1	**Structure and Functions of Nucleic Acids:** Nucleic acid as the genetic material, Chemical structure of DNA and RNA, Watson-Crick model, melting temperature (Tm), major groove and minor groove in DNA, Conformation of nucleic acids: A-, B-, Z-, DNA, types of coding and non-coding RNA, DNA sequencing methods: Sanger sequencing and Next Generation Sequencing (NGS).	
	2	**DNA replication and gene expression**: DNA Replication, Transcription and Translation, post-transcriptional modifications of RNA, gene expression analysis tools.	
	3	**Protein synthesis and processing**: Ribosome, Messenger RNA, tRNA structure and function, Charging of tRNA, Codon and anticodon, Central dogma of molecular biology, Mechanism of protein synthesis, post-translational modifications of proteins, Principles of gene regulation.	
Component 2 (End-sem)	4	**Structure and Functions of proteins:** Structure and types of amino acid, peptide bond, disulfide bond, amino acid metabolism, Primary, secondary, tertiary and quaternary structure of proteins, Functions of protein, Protein denaturation, Electrophoresis: SDS-PAGE, Native PAGE and 2D gel electrophoresis, Protein quantification methods, Proteomics data analysis tools.	

Text Book and References

1. Freifelder D: Molecular Biology, 5th edition. Narosa Publishing House, India
2. Berg JM, Tymoczko JL, Gatto GJ and Stryer L (2015) Biochemistry, 8th Edition, WH Freeman & Co., New York.
4. Watson JD, Baker TA, Bell SP, Gann A, Levine M and Losick R: Molecular Biology of the Gene, 6th edition, Cold Spring Harbour Laboratory Press, Pearson Publication.

Molecular Structure Prediction

Course credit: 3 (3–0–0-0)

Course Outcomes

Upon completion of this course, the student will be able to:
1. Learn about the computational techniques related to biomolecular structure predictions and well-skilled for their utilization.
2. Get well versed with the principle of predicting authentic structures of biomolecules for many health and core biological applications.
3. Acquire fundamental knowledge of applications of systems engineering principles (optimization, systems modeling and simulation, machine learning etc.) to reach industry level readiness in Information Technology applications to predict molecular structure.

Course Contents:

Fundamental aspects of molecular structure and their various physicochemical energy driven stability. Fundamentals of molecular structures, their representations and the experiments to determine their structures. Different hierarchy levels of structures.
Principles of simulations of molecular structure – random walk and molecular-dynamics based simulations. Applicability of different simulation methods for different types of molecular conformations.
Principles of Artificial Intelligence and Machine Learning based prediction of molecular structures and their applicability for different types of molecular conformations.
Principle of knowledge driven prediction of molecular structure using comparative modelling.
Optimization and validation of molecular structure.
Details on utilization of frequently used Industry Tools: Visualization using PyMOL, validation using UCLA SAVES, Simulation using Gromacs, overall structure prediction using Modeller, RoseTTAFold and Google AlphaFold.

Text Book & Reference Books:

1. Introduction to Protein Structure: Carl Branden, John Tooze (Garland)
2. Proteins: Structures and Molecular Properties: Thomas E. Creighton (Freeman)
Guidelines for practical’s if applicable: A one credit lab is to be conducted by covering them relevant and useful topics from aforementioned syllabus.
3. Molecular Modelling: Principles and Applications (2ndEdition): Andrew R. Leach (Prentice Hall)

Cheminformatics for Engineers

2. LTP structure of the course: : L-3, T-0, P-0hrs

3. Objective of the course: Students should be able to build up QSAR models
4. Outcome of the course: Efficiency in prediction of Biological Activity
5. Course Plan: As per the below format only

Unit 1: Fundamentals of organic chemistry,Detail Introduction to Cheminformatics in Drug Discovery. 2D Databases and Database searching: Substructure search, Virtual Screening, property searching, similarity searching, Representation and manipulation of 2D Molecular Structures, 3D Databases: experimental data sources, database searching, Representation and manipulation of 3D Molecular Structures, Selecting Diverse Sets of Compounds.

Unit 2 QSAR: Quantitative Structure and Activity Relationship, Historical Development of QSAR, Hammett Equation, Hansch Equation, Kubinyi bilinear model, Tools and Techniques of QSAR: Biological Parameters, Statistical Methods: Linear Regression Analysis.

Unit 3 Parameters used in QSAR: Electronic Parameters, Hydrophobicity Parameters, Steric Parameters, Molecular Structure Descriptors, Quantitative Models: Linear Models, Nonlinear Models, Free- Wilson Approach, Applications of QSAR: Isolated Receptor Interactions, Interactions at the Cellular Level Interactions in-Vivo, Comparative QSAR: Database Development, Software: GRID, CoMFA.

Unit 4 QSPR : Quantitative Structure and Property Relationship, Octonal Water Partition Coefficient, Quantum Chemical Descriptor, HUMO/LUMO, Predictive Quantitative Structure–Activity Relationships Modeling: Data Preparation and General Modeling Workflow, Reaction network Generation, Open Source Chemoinformatics Software and Database Technologies, Machine Learning based Bioinformatics Algorithms-Applications to Chemicals.

Combinatorial Library Designing: Diverse and Focussed Libraries, Monomer Selection, Product based Library Design, Structure Based Library design, Parallel Synthesis, Microfluidics,Computational fluid dynamics.

6. Text Book: Mandatory for UG core courses
7. References:

1. Chemoinformatics, Concepts, Methods & Tools for Drug Discovery; Ed. Jurgen Bajorath (Humana Press)
2. Chemoinformatics Ed by Johann Gasteigen, Thomas Engel, Wiley-VCH
3. Molecular Modeling, Principles & Applications, Andrew R. Leach
4. Bioinformatics from Genomes to Drugs ; Vol I & 2
5. An Introduction to Chemoinformatics, Andrew R. Leach, Valerie J. Gillet.

Next Generation Sequencing

LTP Structure: 3-0-0-0, Credits: 3

Objectives & Outcomes
1. Develop foundational understanding of nextgeneration sequencing principles, biological data representation, and challenges of largescale datasets.
2. Gain competence in data handling and analysis, including preprocessing, organization, and interpretation of sequencing data with appropriate computational and statistical approaches.
3. Cultivate critical thinking and research orientation to evaluate current tools, identify gaps, and design innovative strategies for genomic data analysis and bioinformatics research.

Unit 1: DNA Sequencing, Strings, and Matching: Introduction to DNA sequencing technologies and their working principles; representation of DNA as a computational string; parsing and manipulation of real genome sequences and sequencing data; fundamentals of sequence comparison including exact matching, homology detection, and pairwise alignment; statistical evaluation of alignment scores; strategies for efficient database searches; applications in metabolomics and pathway modeling.

Unit 2: Preprocessing, Indexing, and Approximate Matching
Methods for preprocessing and indexing genomic data; grouping and ordering strategies; concepts of kmers and their indexing applications; approaches for approximate matching and error tolerance; principles of edit distance and overlaps in sequence analysis; dynamic programming for alignment; distinction between global and local alignment; introduction to de novo assembly and overlap graph construction.

Unit 3: Algorithms for Genome: Assembly Strategies for genome assembly including superstring construction and greedy approaches; challenges posed by repetitive DNA sequences; graphbased assembly frameworks; walkbased approaches for genome reconstruction; insights into the functioning of modern assemblers; discussion on emerging trends and future directions in genome assembly research.

Unit 4: Data Variability and Genomic Analysis: Exploration of variability and replication in sequencing data; techniques for data transformation, preprocessing, and normalization; clustering methods and dimension reduction approaches for highdimensional datasets.

Text/ Reference books:
1. Computational Methods for Next Generation Sequencing Data Analysis by Alexander Zelikovsky and Yi Pan , Wiley
2. Genome-Scale Algorithm Design: Bioinformatics in the Era of High-Throughput Sequencing by Alexandru I. Tomescu, Djamal Belazzougui, and Fabio Cunial, Cambridge University Press