

25UAMT413	BIOINFORMATICS	Category	L	T	P	R	Credit	Year of Introduction
		PEC	3	0	0	0	3	2025

Preamble

The objective of this course is to provide students with a strong foundation in bioinformatics and computational approaches used for biological data analysis. The course covers biological databases, sequence alignment algorithms, similarity searching, phylogenetic analysis, pattern matching techniques, protein structure analysis, genomics, and proteomics. Students will learn to apply algorithmic methods and bioinformatics tools to analyze DNA, RNA, and protein sequences.

Prerequisite: Nil

Course Outcomes: After the completion of the course, the student will be able to:

CO	Course Outcomes	Knowledge Level
CO1	Explain the fundamental concepts of bioinformatics, biological databases (NCBI, GenBank), and sequence formats such as FASTA.	K2 Understand
CO2	Apply sequence alignment algorithms (Needleman-Wunsch, Smith-Waterman) and BLAST variants for biological sequence analysis.	K3 Apply
CO3	Analyze phylogenetic trees and implement combinatorial pattern matching techniques for biological sequences.	K4 Analyze
CO4	Apply computational techniques for protein structure prediction, functional genomics, and proteomics data analysis.	K3 Apply

Mapping of course outcomes with program outcomes and program specific outcomes

CO/PO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PSO1	PSO2
CO1	3	2	-	-	2	-	-	-	-	-	2	2	-
CO2	3	3	2	2	3	-	-	-	-	-	2	3	-
CO3	3	3	3	3	2	-	-	-	-	-	2	3	-
CO4	3	2	2	2	3	-	-	-	-	-	2	3	-
AVG	3	2.5	1.75	1.75	2.5	-	-	-	-	-	2	2.75	-

Mapping of course outcomes with Knowledge and Attitude Profile (WK)

	WK1	WK2	WK3	WK4	WK5	WK6	WK7	WK8	WK9
CO1	*	*	*	*				*	
CO2	*	*	*	*	*			*	
CO3	*	*	*	*	*			*	
CO4	*	*	*	*	*			*	

Wks Mapped with the course

WK1	WK2	WK3	WK4	WK5	WK6	WK7	WK8	WK9
*	*	*	*	*			*	

Mark Distribution:

Total Marks	CIE Marks	ESE Marks	ESE Duration
100	40	60	2.5 hrs

Continuous Internal Evaluation Pattern:

Attendance : 5 marks

Continuous Assessment Test : 20 marks

Continuous Assessment Assignment : 15 marks

SYLLABUS

Module 1 (11 hrs)

Introduction to Bioinformatics and AI: Nature, scope, and role of machine learning in biological data analysis. DNA, RNA, and Protein: structure and function. The Central Dogma of Molecular Biology: Messenger RNA, tRNA, rRNA, and the Genetic Code. Biological Databases: NCBI, GenBank, UniProt; automated data retrieval using APIs and Python (Biopython). Biological sequence formats: FASTA, FASTQ, GenBank format. Feature extraction from biological sequences for ML pipelines. Introduction to AI-driven biological data processing: data cleaning, normalization, and representation learning for genomic data.

Module 2 (11 hrs)

Sequence alignment: Global and Local Alignment. Dot Matrix Method, Dynamic Programming, Needleman-Wunsch and Smith-Waterman algorithms. Gap penalties and scoring. Amino Acid Scoring Matrices: PAM and BLOSUM. Database similarity searching: BLAST variants – BLASTN, BLASTP, BLASTX. Machine learning approaches to sequence alignment: deep learning-based alignment models, embedding biological sequences using word2vec and transformer-based models (ESM, ProtTrans). Neural network substitution matrices and learned gap penalties.

Module 3 (12 hrs)

Phylogenetic tree basics and construction methods: UPGMA, Neighbour Joining, Parsimony trees, Additive trees, and Bootstrapping. Machine learning for phylogenetics: neural network-based tree inference. Combinatorial Pattern Matching: Repeat finding, Keyword Trees, Suffix Trees. Heuristic similarity search algorithms and Approximate Pattern Matching. AI-based pattern recognition in genomic sequences: convolutional models for motif detection, recurrent neural networks for sequence pattern learning. Genome-scale ML applications: variant calling, regulatory element prediction, and gene finding using deep learning.

Module 4 (8 hrs)

Protein structure basics: visualization, comparison, and classification. AI for protein structure prediction: AlphaFold2 architecture, deep learning for structure-function relationship modelling. Functional Genomics: gene expression, EST, SAGE, and Microarrays. Analysing gene expression data using machine learning: clustering (k-means, hierarchical), dimensionality reduction (PCA, t-SNE, UMAP). Proteomics: technology of protein expression analysis. Deep learning for proteomics: mass spectrometry data analysis using neural networks, protein-protein interaction prediction.

Text Books:

1. Compeau, P., & Pevzner, P. (2018). Bioinformatics algorithms: An active learning approach (3rd ed.). Active Learning Publishers.
2. Lesk, A. M. (2019). Introduction to bioinformatics (4th ed.). Oxford University Press.
3. Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). Biological sequence analysis. Cambridge University Press.
4. Pevsner, J. (2015). Bioinformatics and functional genomics (3rd ed.). Wiley-Blackwell.

Reference Books:

1. Compeau, P., & Pevzner, P. (2018). Bioinformatics algorithms (3rd ed.). Active Learning Publishers.
2. Lesk, A. M. (2019). Introduction to bioinformatics (4th ed.). Oxford University Press.
3. Durbin, R., et al. (1998). Biological sequence analysis. Cambridge University Press.

END SEMESTER EXAMINATION – MODEL QUESTION PAPER

Course Code	25UAMT413
Course Name	BIOINFORMATICS
Max. Marks: 60	Duration: 2.5 Hours

Part A (Answer all questions. Each question carries 3 marks)				
Qn No.	Question	Marks	CO	BL
1.	Explain the Central Dogma of Molecular Biology. Describe the roles of mRNA, tRNA, and rRNA in protein synthesis.	3	CO1	L2
2.	What is the FASTA format? How does it differ from FASTQ? What information does each format convey about a biological sequence?	3	CO1	L2
3.	Apply the Needleman-Wunsch algorithm to globally align the sequences: GATTACA and GCATGCU. Use match=+2, mismatch=-1, gap=-2.	3	CO2	L3
4.	Explain the BLOSUM and PAM scoring matrices. How are they derived and when is each matrix most appropriate?	3	CO2	L2
5.	Explain the UPGMA method for phylogenetic tree construction with a suitable distance matrix example.	3	CO3	L3

6.	Explain Suffix Trees for pattern matching in biological sequences. What is the time complexity of a suffix tree search?	3	CO3	L2
7.	Explain the AlphaFold2 architecture and its significance in protein structure prediction.	3	CO4	L2
8.	Describe the use of microarrays in functional genomics. How is gene expression data analyzed using machine learning.	3	CO4	L2

Part B (Answer any one full question from each Module. Each carry 9 marks)				
Module I				
Qn No.	Question	Marks	CO	BL
9.1	Explain the structure of DNA, RNA, and proteins. How does the genetic code translate a DNA sequence into a protein?	5	CO1	L2
9.2	Describe the NCBI and UniProt biological databases. How are they used for automated data retrieval using Biopython?	4	CO1	L2
OR				
10.1	Explain feature extraction from biological sequences for ML pipelines. What representations are used for genomic sequences in deep learning models?	5	CO1	L2
10.2	Compare BLASTN, BLASTP, and BLASTX. In what scenarios is each variant used? Explain the E-value and bit score in BLAST results.	4	CO2	L2
Module II				
Qn No.	Question	Marks	CO	BL
11	Apply the Smith-Waterman local alignment algorithm to align the sequences: ACACACTA and AGCACACA. Use match=+2, mismatch=-1, gap=-2. Show the complete scoring matrix and traceback. Compare the result with global alignment.	9	CO2	L3
OR				
12.1	Explain transformer-based models (ESM, ProtTrans) for biological sequence embedding. How are they used to improve sequence alignment?	5	CO2	L2
12.2	Explain the Dot Matrix method for sequence alignment. What are its advantages and limitations compared to dynamic programming approaches?	4	CO2	L2
Module III				
Qn No.	Question	Marks	CO	BL

13	Explain the Neighbour Joining method for phylogenetic tree construction. Given the following distance matrix, construct the phylogenetic tree: [A-B:5, A-C:9, A-D:11, B-C:10, B-D:8, C-D:6]. Discuss the limitations of distance-based methods.	9	CO3	L4
OR				
14.1	Explain Keyword Trees and Suffix Trees for combinatorial pattern matching. Construct a keyword tree for the patterns: ACG, ACGT, CGT.	5	CO3	L3
14.2	Describe AI-based approaches for variant calling and regulatory element prediction in genomic sequences.	4	CO3	L2
Module IV				
Qn No.	Question	Marks	CO	BL
15.1	Describe the AlphaFold2 model for protein structure prediction. What deep learning innovations enabled AlphaFold2 to achieve near-experimental accuracy.	5	CO4	L3
15.2	Illustrate the use of graph neural networks (GNNs) for protein-protein interaction prediction and drug target identification.	4	CO4	L3
OR				
16	Explain functional genomics using microarray analysis. Describe how PCA, t-SNE, and clustering (k-means, hierarchical) are used to analyze gene expression profiles. Discuss how dimensionality reduction aids in biological insight.	9	CO4	L3
