Shri Vile Parle Kelavani Mandal's

Dwarkadas J. Sanghvi College of Engineering

(Autonomous College Affiliated to the University of Mumbai)

Scheme and detailed Syllabus (DJ19)
of
Honors Degree Program
in
Computational Biology

With effect from the Academic Year: 2024-2025

Proposed Scheme for Final Year Undergraduate Program in Artificial Intelligence (AI) and Data Science: Semester VII (Autonomous)

Sr. No.	Course Code	Course	Teaching Scheme (hrs.)				Continuous Assessment (A) (marks)			Semester End Assessment (B) (marks)					Aggregate (A+B)	Total Credits
			Th.	P	T	Credits	Th.	T/W	Total CA (A)	Th.	O	P	O &P	Total SEA (B)		
SEM VII																
1	DJ19ADHN1C3	Bigdata in Bioinformatics	3	--	--	3	25	--	25	75	--	--	--	75	100	3
2	DJ19ADHN1L3	Bigdata in Bioinformatics Laboratory	--	2	--	1	--	25	25	--	--	--	--	--	50	1
SEM VIII																
3	DJ19ADHN1C4	Genomic Data Science	3	--	--	3	25	--	25	75	--	--	--	75	100	3
Total			6	2	-	4	50	25	75	150	0	0	0	150	250	07

Th	Theory	T/W	Termwork
P	Practical	O	Oral
T	Tuturial		

Prepared by Checked by Head of the Department Vice Principal Principal

Syllabus for Honors Program in Artificial Intelligence (AI) and Data Science (Autonomous)
(Academic Year 2024-25)

Program: Final Year B.Tech. in Artificial Intelligence(AI) & Data Science	**Semester : VII**
Course: Big Data in Bioinformatics	**Course Code: DJ19ADHN1C3**
Course: Big Data in Bioinformatics Laboratory	**Course Code: DJ19ADHN1L3**

Teaching Scheme (Hours / week)				**Evaluation Scheme**						
				Semester End Examination Marks (A)			**Continuous Assessment Marks (B)**			**Total marks (A+ B)**
Lectures	**Practical**	**Tutorial**	**Total Credits**	**Theory**			**Term Test 1**	**Term Test 2**	**Total**	
				75			**25**	**25**	**25**	**100**
				Laboratory Examination			**Term work**		**Total Term work**	
3	**2**	**--**	**4**	**Oral**	**Practical**	**Oral & Practical**	**Laboratory Work**	**Tutorial / Mini project / presentation/ Journal**		**25**
				--	--	--	--	--	**25**	

Objectives: To inculcate in-depth knowledge of processing and analyzing biological data

Outcomes:
The students will be able to:

1. Have a basic understanding of challenges in handling huge biological data
2. Apply tools for biological data analysis
3. Learn the basics of integrating the multi-omics data and the use of NoSQL databases for querying and storing & retrieval of biological data.
4. Use distributed computing architectures and cloud computing platforms for biological data analysis
5. Perform visualization on genomic epidemic data

Detailed Syllabus: (unit wise)		
Unit	**Description**	**Duration in Hrs.**
1	**Module 1: Big Data in Biological Data** • Overview of big data in the context of biological data analysis. • Challenges and opportunities of handling large-scale biological datasets. • Introduction to big data technologies and platforms for biological data analysis. • Case studies and examples of big data applications in genomics, transcriptomics, and other biological domains.	**6**
2	**Module 2: Tools Used for Big Data Analysis** • Introduction to commonly used tools and software packages for big data analysis in bioinformatics. • Hands-on sessions on data preprocessing, analysis, and visualization using popular tools such as Hadoop, Spark, and Python libraries. • Case studies demonstrating the use of big data tools in genomic data analysis, transcriptomic data analysis, and functional annotation.	**6**
3	**Module 3: Integrating Omics Data** • Techniques and methods for integrating multi-omics data from genomics, transcriptomics, proteomics, and metabolomics.	**6**

	• Dimensionality reduction techniques for visualizing and analyzing multi-omics data. • Network-based analysis methods for constructing gene regulatory networks and protein-protein interaction networks. • Case studies demonstrating the integration of omics data to study complex biological phenomena and diseases.	
4	**Module 4: NoSQL Databases in Biological Data** • Introduction to NoSQL databases and their applications in storing and querying biological data. • Comparison of different types of NoSQL databases (e.g., document-oriented, graph-based, key-value stores) and their suitability for biological data. • Hands-on sessions on setting up and using NoSQL databases such as MongoDB, Cassandra, and Neo4j for storing and querying biological datasets. • Case studies demonstrating the use of NoSQL databases in genomic data storage, metadata management, and data integration.	**7**
5	**Module 5: Distributed and Cloud-Based Environments for Biology** • Overview of distributed computing architectures and cloud computing platforms for biological data analysis. • Hands-on sessions on deploying bioinformatics pipelines on cloud computing platforms such as AWS, Google Cloud, and Microsoft Azure. • Best practices for optimizing performance, scalability, and cost-effectiveness of bioinformatics workflows in distributed and cloud-based environments. • Case studies demonstrating the use of distributed computing and cloud-based platforms for large-scale genomic data analysis and collaborative research.	**7**
6	**Visualizing Genomic Epidemiology Data** • Techniques for visualizing genomic data in epidemiological studies, including single nucleotide polymorphisms (SNPs), genetic variants, and phylogenetic trees. • Case studies demonstrating the use of genome browsers and phylogenetic tree visualization tools to analyze the spread and evolution of infectious diseases, such as HIV, influenza, and SARS-CoV-2. • Visualization methods for transcriptomic and proteomic data in epidemiological research, including expression heatmaps, pathway analysis, and protein interaction networks.	**7**

List of experiments:

1. **Analysis of Public Genomic Datasets**: Access public genomic datasets (e.g., from NCBI or ENCODE) and analyze their size, structure, and complexity, gaining an understanding of the scale of biological big data.
2. **Simulated Data Generation**:Use Python libraries like NumPy and SciPy to generate simulated biological datasets of varying sizes and characteristics, exploring the challenges of handling large-scale data.
3. **Introduction to Hadoop**: Set up a Hadoop cluster (either locally or on a cloud platform) and perform basic data processing tasks using Hadoop MapReduce, such as word count on biological text data.
4. **Exploration of Spark**: Explore Apache Spark through hands-on exercises, analyzing biological datasets using Spark RDDs (Resilient Distributed Datasets) and DataFrame APIs, and comparing performance with traditional Hadoop MapReduce.
5. **Data Visualization with Python Libraries**: Use Python libraries like Matplotlib, Seaborn, and Plotly to visualize biological big data, creating plots, histograms, and heatmaps to explore patterns and trends in genomic and transcriptomic datasets.

6. **Introduction to Bioinformatics Databases**: Learn about popular bioinformatics databases (e.g., GenBank, UniProt, TCGA) and retrieve data using APIs or SQL queries, exploring the challenges of handling heterogeneous data sources.
7. **Case Studies in Big Data Applications**: Analyse case studies of big data applications in genomics, transcriptomics, and other biological domains, discussing challenges, methodologies, and insights gained from large-scale data analysis projects.
8. **Data Compression Techniques**: Explore data compression techniques such as gzip and bzip2 and apply them to compress large genomic datasets, comparing compression ratios and trade-offs in storage and processing speed.
9. **Parallel Computing with Python**: Learn parallel computing concepts using Python libraries like multiprocessing and Dask, parallelizing data processing tasks on multi-core CPUs and comparing performance with serial processing.
10. **Data Mining and Machine Learning**: Apply data mining and machine learning techniques (e.g., clustering, classification, regression) to analyze biological big data, identifying patterns, biomarkers, and predictive models from large-scale datasets.

Books Recommended:

1. Integrative Cluster Analysis in Bioinformatics, Author(s):Basel Abu-Jamous, Rui Fa, Asoke K. Nandi First published:20 March 2015

2. NoSQL for Dummies by Adam Fowler - Cloud Computing for Data-Intensive Applications by Jiaheng Lu, Lizhe Wang, and Rajiv Ranjan
3. Cloud Computing in Bioinformatics edited by Shui Qing Ye

Evaluation Scheme:

Semester End Examination (A):

Theory:

- Question paper will be based on the entire syllabus summing up to 75 marks.
- Total duration allotted for writing the paper is 3 hrs.

Continuous Assessment (B):

Theory:

- Two term tests of 25 marks will be conducted during the semester.
- Total duration allotted for writing each of the paper is 1 hr.
- Average marks of the two tests will be considered for final grading.

Prepared by Checked by Head of the Department Vice Principal Principal