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

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Proposed scheme for Final Year Undergraduate Program in Artificial Intelligence (AI) & Data Science with honors in Computational Biology:
Semester VIII (Autonomous)
(Academic Year 2024-2025)

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	4	--	--	4	25	--	25	75	--	--	--	75	100	4
2	DJ19ADHN1L2	Bigdata in Bioinformatics Laboratory	--	2	--	1	--	25	25	--	--	--	--	--	25	1
SEM VIII																
3	DJ19ADHN1C4	Genomic Data Science	4	--	--	4	25	--	25	75	--	--	--	75	100	4
Total			6	2	-	4	50	25	75	150	0	0	0	150	250	09

Th.	Theory	T/W	Term work
P	Practical	O	Oral

Prepared by

Checked by

Head of the Department

Vice Principal

Principal

Program: B.Tech. in Artificial Intelligence(AI) & Data Science with Honors in Computational Biology	Semester : VII
Course: Big Data in Bioinformatics	Course Code: DJ19ADHN1C3
Course: Big Data in Bioinformatics Laboratory	Course Code: DJ19ADHN1L2

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	25
4	2	--	5	Oral	Practical	Oral & Practical	Laboratory Work	Tutorial / Mini project / presentation/ Journal			
				--	--	--	--	--		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. • Dimensionality reduction techniques for visualizing and analyzing multi-omics data.	6

	• 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. NoSQL For Dummies by Adam Fowler, 1st Edition, Wiley, 2015.
2. Cloud Computing for Data-Intensive Applications by Jiaheng Lu, Lizhe Wang, and Rajiv Ranjan, 1st Edition, Springer, 2014.
3. Bioinformatics: A Practical Approach edited by Shui Qing Ye, 1st Edition, Chapman & Hall/CRC, 2007

Web Links:

1. Big Data in Biology | Freshman Research Initiative: https://chatgpt.com/c/67400f97-2fd4-800e-ac3f-353696700cf1
2. MongoDB Courses and Trainings | MongoDB University: https://learn.mongodb.com/
3. Neo4j Graph Database & Analytics | Graph Database Management System: https://neo4j.com/

Online Resources:

1. National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/
2. Coursera | Online Courses & Credentials From Top Educators. Join for Free: https://www.coursera.org/
3. MongoDB Courses and Trainings | MongoDB University: https://learn.mongodb.com/

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

Program: B. Tech. in Artificial Intelligence(AI) & Data Science with Honors in Computational Biology	Semester : VIII
Course: Genomic Data Science	Course Code: DJ19ADHN1C4

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

Prerequisite: -Basics of Computational Biology courses

Course Objectives:

1. To provide an understanding of the foundational concepts in genomics, including DNA, RNA, genes, and genomes, and the role of the central dogma of molecular biology in gene expression.
2. To introduce data science applications in genomics, covering common data formats, sequencing technologies, alignment, mapping, variant analysis, and the interpretation of genomic data for applications in medicine, research, and biotechnology.

Course Outcomes:

On completion of the course, learner will be able to,

1. Explain Fundamental Concepts in Genomics and Data Science.
2. Analyze DNA Sequencing Technologies and Data Characteristics.
3. Apply Sequence Alignment and Mapping Techniques
4. Perform Variant Calling and Genomic Data Annotation
5. Investigate Population Genomics and Trait Associations
6. Evaluate Ethical, Legal, and Social Implications of Genomics

Genomic Data Science (DJ19ADHN1C4)		
Unit	**Description**	**Duration**
1.	**Overview of Genomics and Data Science**: Introduction to Genomics- DNA, RNA, genes, genomes, Central Dogma of molecular biology, Applications of genomics in medicine, research, and biotechnology, Role of data science in genomics, Common data formats (FASTA, VCF, BAM)	**05**
2.	**DNA Sequencing Technologies**: Types of Sequencing Technologies- Sanger sequencing, Next-Generation Sequencing (NGS), Single-molecule sequencing (e.g., PacBio, Oxford Nanopore), Data Generation and Characteristics- Sequencing depth, quality scores, Raw vs. processed genomic data	**06**
3.	**Sequence Alignment and Mapping**: Basics of Sequence Alignment- Pairwise and multiple sequence alignment (Needleman-Wunsch, Smith-Waterman), Alignment algorithms (BLAST, BWA, Bowtie)	**07**
4.	**Variant Calling and Analysis**: Detecting Genetic Variations- Single Nucleotide Polymorphisms (SNPs), Insertions/Deletions (INDELs) **Genomic Data Annotation**: Gene prediction methods, Annotating coding vs. non-coding regions	**07**

5.	**Population Genomics and GWAS**: Allele frequency, population structure, Hardy-Weinberg equilibrium, GWAS- Methods for linking genetic variants to traits **Transcriptomics and RNA-Seq Data Analysis**: RNA sequencing and gene expression analysis, Differential expression analysis	07
6.	**Epigenomics and Regulatory Genomics**: DNA methylation, histone modification, Regulatory elements in the genome (promoters, enhancers) **Introduction to Structural Genomics and Metagenomics:** Structural Variations in Genomics, Copy number variations (CNVs), translocations, inversions, Study of microbial communities from genomic data, Shotgun sequencing, 16S rRNA sequencing **Ethical, Legal, and Social Issues in Genomics:** Privacy, data sharing, and consent, Implications of genomic data in personalized medicine	07
	TOTAL	**39**

Books Recommended:

Textbooks:

1. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, by Durbin et al., Cambridge University Press.
2. Understanding Bioinformatics, by M. Zvelebil and J.O. Baum. Published by Garland Science, 2008.
3. Bioinformatics: Sequence and Genome Analysis by David Mount
4. Computational Genome Analysis: An Introduction by Richard C. Deonier, Simon Tavare, Michael S. Waterman, Springer India

Reference Books:

1. Bioinformatics algorithms: an active learning approach, by Phillip Compeau and Pavel Pevzner. Published by Active Learning Pub.
2. Inferring Phylogenies, by Joseph Felsenstein.
3. Genome-Scale Algorithm Design, by V. Makinen, D. Belazzougui, F. Cunial and A. Tomescu, Cambridge University Press, 2015.

Web Links:

1. https://www.google.co.in/books/edition/Big_Data_Analytics_in_Genomics/5_xRDQAAQBAJ?hl=en&gbpv=1&dq=Data+Science+for+Genomics&printsec=frontcover
2. https://www.sciencedirect.com/book/9780323983525/data-science-for-genomics

Online Courses:

1. https://www.classcentral.com/course/bioinformatics-the-university-of-california-san-d-8962
2. https://www.coursera.org/specializations/genomic-data-science
3. https://onlinecourses.nptel.ac.in/noc24_bt03/preview

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