ఆంధ్రప్రదేశ్ కేంద్రీయ విశ్వవిద్యాలయం
आंध्रप्रदेश केंद्रीय विश्वविद्यालय
Central University of Andhra Pradesh
Jnana Seema, Ananthapuramu

School of Interdisciplinary and Applied Sciences

Department of M.Sc. Molecular Biology

Vidya Dadati Vinayam
(Education Gives Humility)

PG Diploma in Bioinformatics and Data Science

w.e.f. Academic Year 2025-26

CONTENTS

Sl. No.	Section	Page No.
1	Introduction to the Programme	1-2
2	Programme Structure	3
3	Credits Distribution	4
4	Important Information to Students	4-5
5	Syllabus	6-17

PG Diploma in Bioinformatics and Data Science

Introduction to the Programme

The Department of Molecular Biology, Central University of Andhra Pradesh, offers a one-year online PG Diploma certification programme in Bioinformatics and DataScience from the academic year 2025-26. For undergraduate (UG) students, even those with only a basic foundation in life sciences, mathematics, or computer science, a PG diploma provides substantial benefits. It enhances computational and analytical skills that are often not covered in a typical UG syllabus and introduces students to essential bioinformatics tools. The diploma also prepares students for undergraduate research projects, mini-projects, or summer internships, allowing them to efficiently analyze biological datasets including DNA, RNA, and protein sequences, or interpret experimental results. In terms of career exposure, it opens opportunities for entry-level roles in biotechnology, pharmaceuticals, and healthcare startups, while making students more competitive for postgraduate entrance exams and fellowship programs. For postgraduate (PG) students, who often already have specialized knowledge in life sciences, biotechnology, or statistics, the diploma offers advanced analytical skills that enable handling large-scale biological datasets such as genomics, proteomics, and metabolomics. The diploma also facilitates collaboration on multi-disciplinary projects combining biology, computer science, and statistics. For Ph.D. students, who are engaged in advanced research, the diploma provides high-level computational support for analyzing large datasets generated from next-generation sequencing, microarrays, or clinical trials.

Objectives

- To provide foundational and advanced knowledge in bioinformatics, including sequence analysis, structural biology, genomics, and proteomics.
- To develop strong computational and analytical skills for efficient analysis of biological data using modern tools and programming languages.
- To introduce key data science concepts, including statistics, machine learning, data visualization, and predictive modelling.
- To train students in major bioinformatics databases, software tools, and pipelines used in biological and biomedical research.
- To equip learners with problem-solving abilities to manage and analyze large-scale omics datasets using computational approaches.
- To prepare students for careers or advanced research in biotechnology, pharma, healthcare, agriculture, and computational biology through integrated theory and hands-on training.

Learning Outcomes:

On successful completion of the programme students should be able to:

- Apply bioinformatics tools for DNA, RNA, and protein sequence analysis, annotation, and structural prediction.
- Use Python to process, analyze, and visualize biological datasets.
- Perform statistical and machine learning analyses on biological and clinical data to derive meaningful insights.
- Design and implement bioinformatics pipelines for genomics, transcriptomics, proteomics, and systems biology studies.
- Interpret and integrate multi-omics datasets to solve biological problems in research or industry.
- Communicate scientific findings effectively, prepare technical reports, and work independently or collaboratively in interdisciplinary data-driven projects.

Programme Structure

S. No.	Course Code	Title of the Course	Credit Points	Credit Distribution		
				L*	T*	P*
Semester-I						
1	PDBDS 101	Basics of Bioinformatics	4	3	1	-
2	PDBDS 102	Fundamentals of Data Science	4	3	1	-
3	PDBDS 103	Genomics and Proteomics	4	3	1	-
4	PDBDS 104	MOOC/NPTEL/SWAYAM	3	3	-	-
5	PDBDS 105	MOOC/NPTEL/SWAYAM	3	3	-	-
6	PDBDS 106	Lab	2	1	-	1
		Total	**20**	**16**	**3**	**1**
Semester-II						
1	PDBDS 201	Bioinformatics Algorithms and Applications	4	3	-	1
2	PDBDS 202	Next Generation Sequencing Technologies: Data Analysis and Applications	4	3	-	1
3	PDBDS 203	MOOC/NPTEL/SWAYAM	3	3	-	-
4	PDBDS 204	MOOC/NPTEL/SWAYAM	3	3	-	-
5	PDBDS 205	Dissertation	6	3	-	3
		Total	**20**	**15**	**-**	**5**

Note: Project Work is compulsory. Faculty members of the department will be assigned as the supervisors for students.

***L:** Lectures, ***T**: Tutorials, ***P**: Practical

Note: 1. MOOCs are chosen by the student based on the availability of the courses offered on SWAYAM & other related platforms as suggested/recommended by the Department.

2. The desired changes may be made by the Department in the programme structure as and when necessary with the prior approval of the BOS.

Credit Distribution

Semester	**Total Credits**
I	20
II	20
Total	**40**

Important Information to the Student

1. Eligibility:
   - i. CUAP/Non-CUAP students pursuing/completed any PG/PhD Programmes can enroll for PG Diploma Programme offered by the University.
   - ii. Non-CUAP students shall have to appear for an Entrance Examination conducted by the University.
   - iii. A student can enroll for as many PG Diploma Programmes as he/she wishes.

2. The minimum duration for completion of any Postgraduate Diploma Programme is two semesters (one academic year).
3. Maximum duration of completion of programme is two years.
4. A student should have minimum 75% attendance in classes, seminars, practical/ lab in each course of study without which he/she will not be allowed for the Semester -end examination.
5. All theory courses in the programme shall have Continuous Internal Assessment (CIA) component of 40 marks and a Semester-end component of 60 marks. The minimum pass marks for a course is 50%.
6. The student has to appear 3 CIA tests of 15 marks each per semester in each course from which the best 2 performances shall be considered for the purpose of calculating the marks. A record of the continuous assessment is maintained by the department. The remaining 10 marks are awarded based on participation and performance in:
   - Assignments
   - Class presentations
   - Seminars
   - Quizzes
7. A student should pass separately in both CIA and the Semester-end Examination.
8. Semester-end examination shall consist of objective type questions, descriptive type questions, short answer questions and case studies or any others.
9. A student failing to secure the minimum pass marks in the CIA is not allowed to take the semester-end examination of that course. She/He has to redo the course by attending special classes for that course and get the pass percentage in the internal tests to become eligible to take the semester-end examination.
10. Students failing a course due to lack of attendance should redo the course.

Semester - I

Course Code: PDBDS101 **No. of Credits: 4** **Total Hours: 60**	**Basics of Bioinformatics**

Course Objectives

- Teach the basic ideas and importance of bioinformatics.
- Help students learn about biological databases and how to access information from them.
- Explain how to compare sequences, model molecules, and study evolutionary relationships.
- Introduce important computational tools and methods used in bioinformatics.
- Show students how to use bioinformatics to study genes, proteins, and molecular structures.

Learning Outcomes

- Understand the basic ideas, history, and uses of bioinformatics in biology.Find and use major biological databases like NCBI, EMBL, and UniProt to get information.
- Compare biological sequences and interpret results using tools. Build and study phylogenetic trees to explore evolutionary relationships. Use computational methods to predict gene and protein structures.
- Apply bioinformatics tools in areas like genomics, proteomics, and drug development and Analyze and solve biological problems using bioinformatics software.

Unit-I **13 Hours**

Definition and Scope of Bioinformatics, Historical Development and Key Milestones, Major Applications in Biology, Medicine, and Agriculture, Interdisciplinary Nature: Integration of Biology, Computer Science, and Statistics. Major Branches of Bioinformatics: Metabolomics.

Unit -II **13 Hours**

Classification of Databases: Primary, Secondary, and Composite, Nucleotide Sequence Databases: GenBank, EMBL, DDBJ, Protein Sequence Databases: SWISS-PROT, TrEMBL, PIR, Specialized Databases: KEGG (Pathways), Pfam (Protein Families), PDB (Protein Structures), Data Formats, Access Methods.

Unit- III **10 Hours**

Overview of the Central Dogma: DNA - RNA - Protein, Structure and Function of Nucleic Acids(DNA & RNA) and Proteins, Mechanisms of Gene Expression and Regulation, Genetic Code:Features and Biological Significance, Types of Mutations and Their Molecular Consequences.

Unit -IV **14 Hours**

Pairwise Sequence Alignment Techniques: Dot Matrix Method, Dynamic Programming

Course Code: PDBDS 102 **No. of Credits: 4** **No.of Hours : 60**	**Fundamentals of Data Science**

Approaches: Needleman-Wunsch (Global), Smith-Waterman (Local), Multiple Sequence Alignment (MSA)

Unit- V **10 Hours**

Algorithms and Tools: ClustalW, MUSCLE, Scoring Matrices: PAM and BLOSUM Series, Database Similarity Searching: BLAST (Basic Local Alignment Search Tool), FASTA Algorithm.

References :

- Mount, D. W. – Bioinformatics: Sequence and Genome Analysis
- Baldi, P., & Brunak, S. – Bioinformatics: The Machine Learning Approach
- Lesk, A. M. – Introduction to Bioinformatics
- Baxevanis, A. D., & Ouellette, B. F. F. – Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins
- Alberts et al. – Molecular Biology of the Cell
- Watson et al. – Molecular Biology of the Gene
- Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. – Biological Sequence Analysis
- Gibas, C., & Jambeck, P. – Developing Bioinformatics Computer Skills
- Mount, D. W. – Bioinformatics: Sequence and Genome Analysis

Course Objectives:

- Understanding the Data Science Process: Introduce students to the complete data science lifecycle, from defining the problem to generating actionable insights.
- Data Management and Preparation: Learn how to collect, store, manage, and preprocess data, including cleaning and transforming it for analysis.
- Statistical and Algorithmic Techniques: Explore various statistical methods and machine learning models commonly used in data science.
- Data Visualization: Develop skills in using visualization tools and techniques to analyze data and communicate findings effectively.
- Practical Applications: Apply data science concepts and tools to solve real-world problems efficiently.

Learning Outcomes:

- Understand the Significance of Data Science: Explain the role and impact of data science across different domains. Manage Data: Implement methods for effectively collecting, organizing, and storing data for data science projects.
- Prepare Data: Perform data cleaning, integration, and transformation to make data ready for analysis and modelling. Analyze and Visualize Data: Apply statistical techniques and visualization methods to uncover and interpret data patterns.
- Apply Data Science Models: Choose and implement suitable models, including machine learning algorithms, to solve practical problems. Use Data Science Tools: Develop proficiency in tools and libraries such as Python, SQL, and visualization platforms for data analysis and manipulation.

Unit -I **13 Hours**

Definition and historical background of Data Science, Significance, and real-world applications (healthcare, finance, marketing, etc.), Components of Data Science, Roles in a data science project: data analyst, data engineer, data scientist, Overview of the Data Science lifecycle: data acquisition, cleaning, exploration, modeling, visualization, communication.

Unit-II **13 Hours**

Types of data: structured, semi-structured, unstructured, Common data sources: files, databases, APIs, web scraping (basics), Introduction to data quality and preprocessing, Data cleaning: Missing values, outliers, duplicates, Data transformation: normalization, standardization, Encoding categorical variables, Basics of exploratory data analysis (EDA), Descriptive statistics And correlation.

Unit- III **10 Hours**

Introduction to Python programming, Data types, variables, operators, control structures (if-else,loops), Functions, basic file handling, Introduction to libraries: NumPy for numerical computing, Pandas for data handling, Reading and manipulating data using Pandas, DataFrames and Series operations, Data filtering, sorting, grouping, and aggregation

Unit- IV **14 Hours**

Importance of data visualization, Data visualization using Matplotlib and Seaborn: bar plots, histograms, scatter plots, box plots, customizing plots with titles, labels, legends.

Unit- V **10 Hours**

Introduction to machine learning: basic concepts, supervised vs. unsupervised learning, overview of simple algorithms: Linear Regression, KNN, Train-test split, Evaluation metrics: accuracy, Hands- on implementation using Scikit-learn.

References :

- W. McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd ed.,O'Reilly Media, 2017.
- J. Grus, Data Science from Scratch: First Principles with Python, 2nd ed., O'Reilly Media, 2019.
- B. Downey, Think Stats: Exploratory Data Analysis in Python, 2nd ed., O’Reilly Media, 2014.
- C. Müller and S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists, O’Reilly Media, 2016.
- P. Bruce, A. Bruce, and P. Gedeck, Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python, 2nd ed., O’Reilly Media, 2020.

Course Code: PDBDS 103 **No. of Credits: 4** **No.of Hours: 60**	**Genomics and Proteomics**

Course Objectives:

- Gain a comprehensive understanding of the fundamental concepts, principles, and scope of genomics and proteomics, including the organization of genomes, gene structure and regulation, protein structure.
- Describe, compare, and practically apply advanced experimental and computational techniques in Genome sequencing, functional genomics, transcriptomics, and proteomics.
- Develop the ability to analyze, interpret, and visualize large-scale genomic and proteomic datasets. Using modern bioinformatics tools, databases, and statistical methods, enabling insights into gene. Expression patterns, protein networks, and molecular pathways.
- Explore and evaluate the diverse applications of genomics and proteomics in medicine, agriculture, biotechnology, pharmacology, and environmental sciences, with emphasis on disease diagnosis, personalized medicine, crop improvement, and biomarker discovery.

Learning Outcomes:

- Understand and explain the structure, function, and regulation of genomes and proteins, and describe key concepts in genomics, transcriptomics, and proteomics.
- Apply experimental and computational techniques such as genome sequencing, microarrays, mass spectrometry, and bioinformatics tools to analyze genomic and proteomic data.
- Interpret and integrate large-scale data to identify gene expression patterns, protein
- interactions, and molecular pathways, and relate these to biological processes and disease mechanisms.
- Critically evaluate scientific literature and experimental results to design experiments
- formulate hypotheses, and propose solutions in genomics and proteomics research.

Unit I **13 Hours**

Definition and scope of genomics. Types of genomes: Prokaryotic vs Eukaryotic. Structural genomics and functional genomics. Genome assembly and annotation strategies. Comparative genomics: principles and applications. Applications of genomics in medicine, agriculture, and evolutionary biology.

Unit II **13 Hours**

Overview of transcriptomics and gene expression profiling. Techniques for gene expression analysis: DNA microarrays and RNA-Seq. Gene regulatory elements: promoters, enhancers, silencers. Epigenetic modifications: DNA methylation and histone modification.

Unit III **10 Hours**

Non-coding RNAs and their functions: microRNAs (miRNA), small interfering RNAs (siRNA). Introduction to genome editing technologies: CRISPR-Cas9 and related tools. Definition, scope, and types of proteomics (structural, functional, and expression proteomics).

Unit IV **10 Hours**

Structure and function of proteins. Protein–protein interaction networks and their biological Significance. 2D Gel Electrophoresis (2D-PAGE).

Unit V **14 Hours**

SDS-PAGE, Western Blotting, Introduction to mass spectrometry (MS) and its role in proteomics. Methods for protein purification and separation.

References:

- Bioinformatics: Sequence and Genome Analysis" – David W. Mount
- Essential Bioinformatics – Jin Xiong
- NCBI Tutorials and EBI Resources
- Discovering Genomics, Proteomics, and Bioinformatics" by A. Malcolm Campbell & Laurie J. Heyer
- Genomics and Proteomics: Functional and Computational Aspects" by Suhai S.
- Functional Genomics: A Practical Approach" by Hunt S.P.
- Principles of Genome Analysis and Genomics" by Primrose S.B. & Twyman R.H.

Course Code: **PDBDS106** No. of Credits: **2** No.of Hours: **60 hrs**	**Practical: Bioinformatics and Data science**

1.Exploration of Major Bioinformatics Databases (NCBI, UniProt, KEGG, PDB)

2.Retrieval and Analysis of DNA/Protein Sequences

3.Use of BLAST and FASTA for Sequence Similarity Search

4.Building a linear regression model using Scikit-learn

5.Clustering data using K-Means

6.Evaluating classification results with precision, recall, accuracy

7.Running ML notebooks on Google Colab

8.Setting up Python and Jupyter Notebook

9.Writing simple Python programs (data types, input/output)

10.Pairwise and Multiple Sequence Alignment using ClustalW and MUSCLE

Semester - II

Course Code: **PDBDS201** No. of Credits: **4** No of Hours: **60**	**BioInformatics: Algorithms and Applications**

Course Objectives:

- Understand the fundamentals of bioinformatics and the role of algorithms in analyzing biological data.Builds the foundation to interpret large-scale biological data accurately.
- Learn key algorithms for sequence alignment, genome analysis, and protein structure prediction.Equips students with tools to solve real-world problems in genomics and proteomics.
- Apply computational tools and databases to analyze biological sequences and functional genomics data. Enables practical skills for research and biotechnology applications.
- Explore applications of bioinformatics in medicine, agriculture, and drug discovery. Demonstrates the impact of bioinformatics on improving health, crops, and therapeutics.

Learning Outcomes:

- Explain the principles of bioinformatics and the importance of algorithms in biological data analysis.
- Apply algorithms for sequence alignment, gene prediction, and protein structure analysis.
- Use bioinformatics tools and databases to analyze and interpret biological.
- Evaluate and discuss real-world applications of bioinformatics in research, medicine, biotechnology.

Unit-I **15 Hours**

Concepts and importance of Bioinformatics, Complexities in biological systems, DNA sequence analysis, Sequence based parameters, Database, Database categories. Protein structure and function, Protein sequence databases, Pairwise alignment, Pairwise alignment II, Uniprot Demo.

Unit-II **10 Hours**

Sequence alignment, Sequence alignment II, Sequence alignment: Online resources. Sequence alignment: Online resources II, Conservation score, Conservation score II,Blast Demo. Phylogenetic trees, Protein sequence analysis, Protein sequence analysis II, Hydrophobicity profiles, Patterns and PSSM profiles.

Unit-III **10 Hours**

Construction of Non-redundant datasets I, Non-redundant datasets II,Protein secondary structure.Protein tertiary structure, Protein structure analysis, Demo: PDB or Pymol or PDB Param

Unit- IV **15 Hours**

Protein structure analysis – IV ,Protein structure prediction, Protein stability – II, Demo: Homology Modelling. Stabilizing residues, Thermodynamic database, Stability of proteins upon mutations, Stability of proteins upon mutations II, Demo: Pro Therm

Unit-V **10 Hours**

Protein folding rate Protein interactions, Computer aided drug design. Computer aided drug design – II, Virtual screening – I, Virtual screening – II,QSAR – I, QSAR – II.Demo: Autodock, awk programming, Development of algorithms. Applications of bioinformatics

Practicals :

1.Interpretation of Alignment Scores and Phylogenetic Trees
2.File Format Conversions and Data Handling (FASTA, GenBank, GFF, PDB)
3.Exploring datasets in Excel and CSV format
4.Computing descriptive statistics in Pandas
5.Creating visualizations using Matplotlib and Seaborn

References

- Arthur Lesk – Introduction to Bioinformatics
- David W. Mount – Bioinformatics: Sequence and Genome Analysis
- Jonathan Pevsner – Bioinformatics and Functional Genomics
- Andreas D. Baxevanis & B. F. Francis Ouellette – Bioinformatics: A Practical Guide to the Analysis of Genes and Protein.
- T. K. Attwood & D. J. Parry-Smith – Introduction to Bioinformatics Rastogi et al. – Bioinformatics: Methods and Applications

Course Code: **PDBDS202** No. of Credits: **4** No of Hours : **60**	**Next Generation Sequencing Technologies: Data Analysis and Applications**

Course objectives :

- Understand the basics and principles of next-generation sequencing (NGS) technologies.
- Learn how to process and analyze NGS data using computational tools.
- Explore applications of NGS in genomics, medicine, and biotechnology.
- Develop skills to interpret NGS results for research and practical use.

Learning Outcomes:

- Explain how NGS technologies work and their importance in modern biology.
- Process and analyze NGS data using software and bioinformatics tools.
- Apply NGS data to study genes, genomes, and biological systems.
- Interpret and communicate NGS results for research or practical applications.

Unit-I **17 Hours**

Introduction, Next Generation Sequencing Technologies - 454 Sequencing, Illumina Sequencing. By Synthesis (SBS),Single Molecule Real Time (SMRT) Sequencing, I on Torrent and Nanopore Sequencing, Sequencing Coverage, Quality Score and Experiment Design, Data Formats, Data Quality, Data QC and Trimming. Hands on: Setting up the system, Basic Shell Commands, Data Download and Exploration, Hands-on 1 - Data exploration and QC, Hands-on 1 - Data QC and Trimming. Read Mapping, Mapping Algorithms, Suffix tree-based mapping algorithm, Burrows-Wheeler Transform (BWT), Read Mapping with BWT.

Unit-II **10 Hours**

Bowtie2 tool, Mapping reads with Bowtie2,Bowtie2 output, SAM and BAM format, SAM format: Alignment section, Variant Calling, Calling SNP/SNVs and Indels, Hands-on analysis : Variant Calling, VCF Files. Variant Annotation, Analysis of CNVs and SVs, Introduction to RNA sequencing, RNA-seq data processing pipeline, Transcriptome Assembly & Quantification, Transcript Abundance Quantification.

Unit-III **10 Hours**

Biases in RNA-seq experiments, Data Normalization Methods, Data Normalization Methods (Contd.), Differential Gene Expression (DGE) Analysis, DGE analysis results and visualizations. Multiple hypothesis testing correction, FDR correction and interpretation of DGE analysis results, Functional Enrichment Analysis, RNA-seq data analysis. Hands-on 2: Preliminary Data Analysis, Sample Specific Bias Correction, Differential Gene Expression Analysis I, DGE Analysis with spike-ins, DGE Analysis Results and Functional Enrichment Analysis.

Unit -IV **10 Hours**

Genome Assembly, Shortest Common Superstring (SCS) assembly, Overlap-Layout-Consensus (OLC) approach, de Bruijn Graph (DBG) based assembly, Assembly and Quality Control.

Unit-V **13Hours**

Applications of NGS in Epigenomics, Detecting DNA Methylations, Genome-wide Transcription Factor (TF) Binding Sites, Chromatin Accessibility, Genome Organization in 3D.

Practicals:

1.Analyzing relationships using correlation matrix and heatmaps

2.Plotting probability distributions

3.Loading and inspecting datasets usin

g Pandas

4.Data cleaning (removing nulls, fixing data types)

5.Normalizing and encoding data

Reference :

- Bioinformatics and Functional Genomics – Jonathan Pevsner
- Next-Generation DNA Sequencing Informatics – Stuart M. Brown
- Genomes 4– T.A. Brown
- Understanding Bioinformatics – Marketa Zvelebil & Jeremy O. Baum
- Research papers and manuals for tools like Bowtie2, SAMtools, GATK, DESeq2.

Course Code: PDBDS205 No. of Credits: 6	Project Work

- Each student will work on a project towards a dissertation by applying the knowledge acquired in molecular biology and biochemistry. The project maybe selected based on a literature survey and available resources as suggested by the respective supervisors.
- The students may conduct the project work either at the CUAP or laboratory of their choice in India or abroad upon obtaining approvals from the competent authority. Student is required to submit detailed project report on the selected topic for their project as per the guidelines decided by the department.
- The project work is to be evaluated through presentation and viva-voce during the semester and the final evaluation will be done at the end of the semester as per the guidelines decided by the department from time to time.
- Candidate may visit research labs/institutions with the due permission of the head on the recommendation of the supervisor concerned.

  Dissertation/Project report: Evaluation – 60 Marks

  Viva-Voce – 40 marks