INTERNAL ASSIGNMENT QUESTIONS
P.G. Diploma in Bio Informatics

Semester - II

ANNUAL EXAMINATIONS
2025

PROF. G. RAM REDDY CENTRE FOR DISTANCE EDUCATION
(RECOGNISED BY THE DISTANCE EDUCATION BUREAU, UGC, NEW DELHI)
OSMANIA UNIVERSITY
(A University with Potential for Excellence and Re-Accredited by NAAC with "A" + Grade)

DIRECTOR
Prof. G.B. Reddy
Hyderabad – 7 Telangana State

PROF.G.RAM REDDY CENTRE FOR DISTANCE EDUCATION
OSMANIA UNIVERSITY, HYDERABAD – 500 007

Dear Students,

Every student of PG Diploma in Bio Informatics semester I has to write and submit **Assignment** for each paper compulsorily. Each assignment carries **30 marks.** The marks awarded to the students will be forwarded to the Examination Branch, OU for inclusion in the marks memo. If the student fail to submit Internal Assignments before the stipulated date, the internal marks will not be added in the final marks memo under any circumstances. The assignments will not be accepted after the stipulated date. **Candidates should submit assignments only in the academic year in which the examination fee is paid for the examination for the first time.**

Candidates are required to submit the Exam fee receipt along with the assignment answers scripts at the concerned counter on or before **20-05-2025.** and obtain proper submission receipt.

ASSIGNMENT WITHOUT EXAMINATION FEE PAYMENT RECEIPT (ONLINE) WILL NOT BE ACCEPTED

Assignments on Printed / Photocopy / Typed will not be accepted and will not be valued at any cost. Only HAND WRITTEN ASSIGNMENTS will be accepted and valued.

Students are advised not to use Black Pen.

Methodology for writing the Assignments (Instructions) :

1. First read the subject matter in the course material that is supplied to you.
2. If possible read the subject matter in the books suggested for further reading.
3. You are welcome to use the PGRRCDE Library on all working days for collecting information on the topic of your assignments. (10.30 am to 5.00 pm).
4. Give a final reading to the answer you have written and see whether you can delete unimportant or repetitive words.
5. The cover page of the each theory assignments must have information as given in FORMAT below.

FORMAT

1. NAME OF THE STUDENT :
2. ENROLLMENT NUMBER :
3. NAME OF THE COURSE :
4. SEMESTER (I, II, III & IV) :
5. TITLE OF THE PAPER :
6. DATE OF SUBMISSION :

6. Write the above said details clearly on every subject assignments paper, otherwise your paper will not be valued.
7. Tag all the assignments paper wise and submit them in the concerned counter.
8. Submit the assignments on or before **20-05-2025** at the concerned counter at PGRRCDE, OU on any working day and obtain receipt.

DIRECTOR

INTERNAL ASSESSMENT

PAPER – I : COMPUTATIONAL APPROACHES TO MODERN BIOLOGY

ASSIGNMENT - I

UNIT – I : Answer the following questions (each question carries three marks) 5x3=15

1. Create a set of three publication – ready plots (Scatter histogram and bar chart) from a real or simulated dataset using ggplot2. Customize the aesthetics (color, labels, legend) to make them ready for publication. Explain why you made those choices.

2. Design a use-case where BioSQL is essential for managing bioinformatics data. Write sample SQL queries for loading sequences, retrieving metadata using JOINs, and explain how Biopython can be integrated for downstream analysis.

3. Use any small real-world dataset (e.g., marks, plant height, etc.) and perform a **t-test** or **linear regression** in both R and Python, Provide screenshots or code outputs and comment on any differences you observed in the results or interpretation.

4. Sketch (on paper or software) a **simple biological network** of at least five genes or proteins that interact in a pathway (e.g., insulin signalling or glycolysis) Write a few lines explaining how one change (like a mutation) might affect the network.

5. Ask a friend or family member to list what they eat in a typical day. Based on their diet, **predict what types of microbes** might be abundant in their gut (e.g., fiber-digesting, fat-metabolizing). Relate your answer to concepts from microbiome studies.

ASSIGNMENT - II

UNIT – II : Answer the following questions (each question carries three marks) 5x3=15

1. Write an R script that uses a for loop to count the number of vowels in a sentence of your choice (minimum 10 words). Comment each step in your code and explain why you used a loop instead of a vectorized function.

2. Write a Biopython script to read a **FASTA** file containing at least two sequences. Add your own annotation (like species name or gene function) as a comment in the code. Explain what challenges you faced while parsing or handling the file.

3. Download a small sample of RNA-seq FASTQ data (e.g., from SRA or a public dataset). Try running a basic quality check using **FastQC** and record the report. Interpret **any one warning** or **low- quality score** in your own words-what might it mean for downstream analysis?

4. Choose any plant, animal, or bacterial species of interest and research if its genome has been assembled **de novo** or using a reference. Provide a brief explanation of what method was used, and why that method might have been chosen.

5. Take a real or hypothetical SNP variant (e.g., "G>A in gene BRCA1") and describe **what functional effect it could have**, based on its location (e.g., coding region, intron). Support your answer with reasoning-not a database search.

INTERNAL ASSESSMENT

PAPER – II : COMPUTER AIDED DRUG DESIGNING AND DEVELOPMENT

ASSIGNMENT - I

UNIT – I : Answer the following questions (each question carries three marks) 5x3=15

1. You are given a protein sequence and three potential templates for homology modeling. Based on sequence identity, alignment quality, and functional similarity, how would you prioritize the templates? Justify your choice with reference to at least two structural features.

2. You are designing a drug targeting a known protein-protein interaction site. List two key considerations when evaluating docking results and explain how they influence drug efficacy

3. Given two ligands that bind the same target with similar scores but different binding modes, propose a method to determine which one might be a better drug candidate and justify your reasoning

4. You developed a QSAR model that performs well on training data but poorly on test data. Identify and explain two likely reasons for this issue and suggest how to address them

5. Compare the use of CNNs and RNNs in analyzing molecular structure versus biological sequence data. In what scenario would each be preferable?

ASSIGNMENT - II

UNIT – II : Answer the following questions (each question carries three marks) 5x3=15

1. Given a plot of free energy vs. reaction coordinate from metadynamics, identify the most stable conformational state and explain how this information could guide ligand optimization.

2. A newly discovered disease has minimal pathway annotation available. Suggest a computational workflow for target identification using network analysis, starting from transcriptomic data.

3. You are given RMSD and RMSF plots from a molecular dynamics simulation. Interpret the stability of the protein and its flexible regions and propose what insights this offers into drug binding.

4. You trained a machine learning model to predict hepatotoxicity but it performs poorly on new data. Identify and explain two possible causes related to training data and model choice.

5. You are comparing two ligands in terms of binding affinity and selectivity. How would structural flexibility of the target influence your choice during lead optimization?

INTERNAL ASSESSMENT
PAPER – III : AI / ML FOR BIOINFORMATICS

ASSIGNMENT - I

UNIT – I : Answer the following questions (each question carries three marks) 5x3=15

1. Take a small example of **DNA or RNA sequence data** (even a fake/simulated one). Describe how you would manually **clean and normalize** the data before applying a Machine Learning model. Write a step-by-step process.

2. Imagine you have RNA-seq data from 1000 unknown samples. Describe **how you would apply K-means clustering** to group the samples. Write the steps you would perform manually before and after clustering (no code needed, only logic).

3. Assume you are asked to predict a missing part of a DNA sequence using a Neural Network. Would you use a CNN or an RNN? Explain your choice with an example in your own words (no definitions, just application reasoning)

4. Design a **multi-omics project** by choosing **one genomics, one transcriptomics**, and **one proteomics** dataset you would integrate to study cancer progression. Explain how AI/ML could help you in the integration.

5. You built an ML model for disease prediction, but you find that its accuracy is very high while precision is low. Explain (in your own understanding) why this might happen in biological data and **suggest two ways to improve model reliability.**

ASSIGNMENT - II

UNIT – II : Answer the following questions (each question carries three marks) 5x3=15

1. Choose either **Depth-first search (DFS) or Breadth-first search (BFS)** and design a **simple real-world example** where this search method could be applied in a biological experiment or bioinformatics workflow. Draw a diagram if needed.

2. Imagine you are designing an **AI system** for predicting mutations in COVID-19 virus genomes. Which **AI technique** (CNN, RNN, Deep Learning, etc.) would you choose and why? Justify your answer with your own reasoning, not a Google search.

3. Pick a real biological problem (e.g., predicting plant disease, finding cancer markers). Explain which type of learning (Supervised, Unsupervised, Reinforcement) you would use and why. Justify your choice using your own example.

4. Imagine your AI model predicts which proteins interact in a disease pathway. Choose either **SHAP** or **LIME** and **explain how you would use it** to explain your model's predictions to biology research team.

5. Suppose you are developing an AI tool for predicting genetic disease risks. List **two ethical concerns** you think are most important when using patient genomic data, and explain why?