PRIFYSGOL ABERYSTWYTH UNIVERSITY

SEMESTER ONE 2025

COMPUTER SCIENCE

CSM6820 : COMPUTATIONAL BIOINFORMATICS

Time allowed: **2 HOURS**

Answer all **THREE** questions

All questions carry equal marks.

YOU MAY BRING THE FOLLOWING INTO THE EXAMINATION:
UP TO 3 SHEETS OF NOTES ON A4 PAPER (WHICH MAY BE DOUBLE-SIDED). THERE ARE NO ADDITIONAL RESTRICTIONS ON CONTENT OR FORMATTING OF THESE NOTES.
NO ELECTRONIC DEVICES.

CALCULATORS ARE PERMITTED (CASIO FX991, FX83 OR FX85 MODELS ONLY)

STUDENTS MAY SUBMIT ANSWERS TO THIS PAPER IN EITHER WELSH OR ENGLISH

Once seated, students may complete the front cover of the answer book(s) and the attendance slip ONLY.

DO NOT TURN THE EXAMINATION QUESTION PAPER OVER UNTIL INSTRUCTED TO DO SO.

1. **This question is related to genome assembly.**

You are given the following two DNA reads:

Read-1: TGCATGACCTGAACTGA

Read-2: CCTGAACTGATTCGGA

a) Generate a set of 6-mers for each read and explain how you generate them. [6 marks]

b) Using your set of 6-mers from Read-1 and Read-2, construct the DeBruijn graph and explain how you construct it. [10 marks]

c) Identify the start and end nodes and determine whether the graph has an Eulerian path and explain how you check for the existence of Eulerian path. [4 marks]

d) Reconstruct the contig (the largest sequence that can be constructed using the De Bruijn graph) and verify that the reconstructed contig contains both Read-1 and Read-2. Explain how you reconstruct the contig. [7 marks]

e) Briefly discuss how choosing k=3 instead of k=6 would affect graph quality. [6 marks]

2. **This question is related to open reading frames and sequence similarity.**

You are given the following two DNA sequences:

Seq-1: ATGGCTAACGTTTAATGTAA

Seq-2: ATGGCTAACTGACATGTAAC

a) Write down all possible open reading frames in each sequence in the forward strand and report the strand they are found. Explain how you search for the ORFs. [4 marks]

b) Calculate the pairwise sequence similarity scores between the ORFs of Seq-1 and Seq-2 using Levenshtein distance. Explain how you calculate the Levenshtein distance. [12 marks]

c) Calculate the pairwise sequence similarity scores between the ORFs of Seq-1 and Seq-2 using Manhattan distance on 2-mer vectors. Explain how you calculate the Manhattan distance. [12 marks]

d) Compare and contrast alignment-based and alignment-free approaches for measuring sequence similarity. In your answer, explain the underlying principles of how each method works and discuss their respective advantages and limitations. Provide at least one key point for each. [5 marks]

3. **This question is related to DNA sequencing and ethical issues surrounding human genome editing.**

a) Compare and contrast Sanger sequencing, Illumina sequencing, and Oxford Nanopore sequencing in terms of underlying principles, read length, accuracy, and cost. Provide one key point for each metric. [10 marks]

b) Prodigal is a widely used bioinformatics tool for predicting protein-coding genes in prokaryotic genomes. Describe the underlying algorithmic principles used by Prodigal to identify open reading frames (ORFs) and distinguish true genes from random sequences. Provide at least five key points. [10 marks]

c) Describe one potential benefit and two potential harms of using gene editing in humans [6 marks].

d) Explain how international regulations and guidelines (e.g., WHO and UNESCO) address the ethical use of gene editing technologies. [7 marks]

END OF PAPER