UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE :	**COMP0082**
ASSESSMENT PATTERN:	**COMP0082A7PF 001**
MODULE NAME :	**COMP0082 - Bioinformatics**
LEVEL :	**Postgraduate**
DATE: :	**10/05/2024**
TIME :	**14:30**
DURATION :	**2 Hours**

This paper is suitable for candidates who attended classes for this module in the following academic year(s):

Year
2023/24

EXAMINATION PAPER CANNOT BE REMOVED FROM THE EXAM HALL. PLACE EXAM PAPER AND ALL COMPLETED SCRIPTS INSIDE THE EXAMINATION ENVELOPE

Hall instructions	N/A
Additional materials	N/A
Standard Calculators	Yes
Non-Standard Calculators	No

TURN OVER

UCL Computer Science Examination Paper

Paper Details

Academic Year:	2023/24
Module Title:	Bioinformatics
Module Code:	COMP0082
Exam Period:	Central Assessment Period: Main Summer
Duration:	2 hours
Deliveries for which suitable:	A7P (Postgraduate Taught, Level 7) A7U (Undergraduate, Level 7)
Cohorts for which suitable:	2023-24 2022-23 2021-22

Instructions

There are FOUR questions in total.

Answer the ONE question from SECTION A and any TWO questions from SECTION B.

A maximum of 100 marks is available: 34 marks from SECTION A and 66 marks from SECTION B. The marks available for each part of each question are indicated in square brackets [n].

Standard calculators are permitted.

Section A

1.

(a) With suitable diagrams, describe the processes by which an amino acid sequence is ultimately produced from a typical eukaryotic gene. Make sure you correctly name the important components and correctly label the directions of nucleic acid chains involved.

[10 marks]

(b) Draw a diagram of a typical mammalian cell and label its key components and features. Indicate which aspects of the cell would be common to typical bacterial cells.

[5 marks]

(c) Write brief notes on human chromosomes. In your answer, include discussion of the numbers of chromosomes in typical human cells and discuss the importance of the X and Y chromosomes.

[3 marks]

(d) List and briefly define the 4 levels of protein structure.

[4 marks]

Q1 continued overleaf…

1st Position	2nd Position				3rd Position
	U	**C**	**A**	**G**	
U	Phe	Ser	Tyr	Cys	**U**
	Phe	Ser	Tyr	Cys	**C**
	Leu	Ser	STOP	STOP	**A**
	Leu	Ser	STOP	Trp	**G**
C	Leu	Pro	His	Arg	**U**
	Leu	Pro	His	Arg	**C**
	Leu	Pro	Gln	Arg	**A**
	Leu	Pro	Gln	Arg	**G**
A	Ile	Thr	Asn	Ser	**U**
	Ile	Thr	Asn	Ser	**C**
	Ile	Thr	Lys	Arg	**A**
	Met	Thr	Lys	Arg	**G**
G	Val	Ala	Asp	Gly	**U**
	Val	Ala	Asp	Gly	**C**
	Val	Ala	Glu	Gly	**A**
	Val	Ala	Glu	Gly	**G**

(e) A short insect gene has been sequenced, giving the following DNA sequence, and it is known that this gene includes a single unusually short intron of only 11-nt, but with typical spliceosome donor/acceptor motifs at both ends. Identify and remove the intron, and then write out the 3 possible forward reading frames for this sequence, indicating which is the most likely protein translation of this gene. Explain your reasoning for picking the given translation, and write all the translations in single letter amino acid code form.

5’ – tatggaaaacaccgcgctgatgttctctgaagtacctattaa – 3’

[8 marks]

(f) A rare mutation is known to occur in this gene.

The change is as follows: T>C at position 40.

What general effect do you think this mutation is likely to have on the final expressed protein for this gene and explain your answer briefly?

[2 marks]

(g) Calculate how many different short peptides of length 6 can theoretically be made out of standard genetically coded amino acids, and then calculate the total number of different mRNA sequences that could code for these same peptides. Assume that your peptides all include a standard start codon at position 1. Show your working.

[2 marks]

[Total for Question 1: 34 marks]

Section B

2.

(a) Briefly explain the difference between primary and secondary data resources in bioinformatics. Name ONE example each of data resources which contain the following types of biological information, and indicate whether they are primary or secondary data resources:

i) Nucleic acid sequences

ii) Protein sequences

iii) Families of aligned protein sequences in the form of HMMs.

[4 marks]

(b) Other that the sequence itself, give three different types of information that can be extracted from the initial header records of an entry in a protein sequence data bank. [3 marks]

(c) Using the ProtFun method as an example, explain how ML techniques can be used to predict the function of proteins without homology to already characterised proteins. [10 marks]

(d) Describe how a transformer encoder trained with BERT loss on a large protein sequence data bank might be used to predict GO terms for protein sequences. [10 marks]

(e) Other than the amino acid sequence itself, briefly describe 3 other sources of biological information that could be used to improve function prediction accuracy? [6 marks]

[Total for Question 2: 33 marks]

3.

(a) Briefly explain the difference between local and global protein sequence alignment. What feature of large proteins makes local alignment often preferable to global in detecting homology? [3 marks]

(b) Complete (and modify if necessary) the following pseudocode for the Smith-Waterman algorithm using a constant gap penalty G and scoring matrix B:

```
S[0,0] = 1
for i = 1 to M
     S[i,0] = 1
for j = 1 to N
     S[0,j] = 1

for i = 1 to M
     for j = N to 1
          S[i,j] = max(   )
```

If you do decide to make modifications to the code, briefly explain why the modification is necessary. Make sure you define any new variables or parameters that you add to the code.

[5 marks]

(c) Explain briefly how score matrices are computed for use in sequence alignment algorithms. [2 marks]

(d) Outline briefly how FASTA speeds up the searching of sequence data banks compared to the Smith-Waterman algorithm. [3 marks]

(e) Describe how HMMs can be used to identify likely homology between a protein and existing protein families. In your answer, name the key algorithms, and clearly describe the sources of data and procedures used. Include discussion on how model bias and overfitting are commonly avoided, and make use of diagrams where appropriate. [10 marks]

(f) Explain how machine learning methods can be used to derive 3-D structural information from patterns of amino acid covariation observed in multiple sequence alignments. [10 marks]

[Total for Question 3: 33 marks]

4.

(a) Name and describe 3 fields of contemporary -omics analysis.

[6 marks]

(b) RNAseq is an experimental method for measuring the concentration of mRNAs in a cell or tissue sample. Briefly describe how the RNAseq experimental method works. [6 marks]

(c) i. Your colleagues have completed a yeast-2-hybrid association experiment to measure the binding affinities of 4 proteins. They bring you a matrix of scaled, normalised (0 to 1.0) binding affinities for the 4 proteins. The values measure how strongly the proteins bind to one another. Using the UPGMA algorithm calculate and draw the dendrogram for this data.

[6 marks]

	COX12	**HSP87**	**OLI19**	**SCEI2**
COX12	1	0.8	0.9	0.1
HSP87	0.8	1	0.8	0.2
OLI19	0.9	0.8	1	0.1

SCEI2	0.1	0.2	0.1	1

ii. Given the matrix of affinities and the clustering you have calculated draw a putative Protein-Protein-Interaction graph for these four proteins.

[2 marks]

(d) Explain the main steps required in order to infer a biological network from -omics data. Where relevant, give examples of methods or algorithms that could be used for each step.

[9 marks]

(e) Describe 4 main steps in a typical spatial transcriptomics experiment.

[4 marks]

[Total for Question 4: 33 marks]