UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE	:	**COMP0082**
ASSESSMENT Pattern:		**COMP0082A7PE**
MODULE NAME	:	**Bioinformatics**
LEVEL:	:	**Postgraduate**
DATE:	:	**12-May-2023**
TIME	:	**14:30**
DURATION	:	**02:15**

Late submission is permitted for Controlled Conditioned exams but late penalties will apply - any submissions that are up to 40 minutes late will be penalised, after which no submissions will be accepted under any circumstances.
You must ensure to allow sufficient time to upload and hand in your work

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

Year
2022-23

Duration	<< Exam Duration>>
Additional time for converting handwritten notes to PDF where applicable	**15 mins**
Upload window	**20**
Total time	**2 Hours 35 mins**

Additional material	N/A
Special instructions	Submit your answers as a single PDF file. Any handwritten answers should be scanned and compiled according to the guidance provided by the UCL Examinations Office. Any included diagrams should be your own original work.

TURN OVER

UCL Computer Science Examination paper

Paper details

Academic year:	2022/23
Module title:	Bioinformatics
Module code:	COMP0082
Exam period:	Main summer assessment period
Duration:	2 hours
Deliveries for which intended:	A7P (taught postgraduate, level 7) A7U (undergraduate, level 7)
Cohorts for which intended:	2021/22/23

Instructions

There are FOUR questions in total. Answer the 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.

Submit your answers as a single PDF file. Any handwritten answers should be scanned and compiled according the guidance provided by the UCL Examinations Office. Any included diagrams should be your own original work.

Section A

Answer the ONE question from this section.

1)

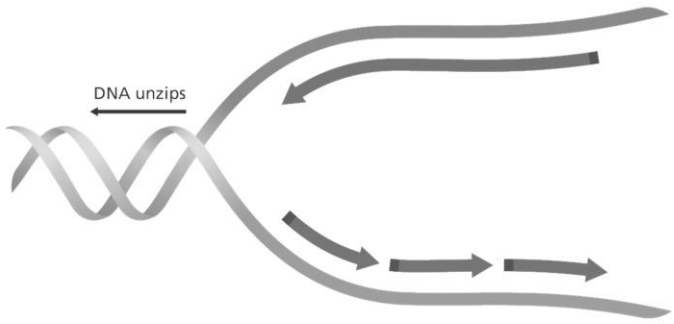
a) Read the following paragraph and identify the two factual errors it contains.

"The central dogma of molecular biology is a fundamental tenet of molecular biology that describes the flow of genetic information within a biological system. It states that genetic information stored in DNA is first transcribed into RNA, which is then translated into nucleotides. This flow of genetic information is bidirectional, meaning that information can flow from proteins back into DNA or RNA. The central dogma is an important concept in molecular biology because it helps us understand how genetic information is stored, accessed, and used to build the organs that are necessary for life. It also helps us understand how genetic mutations can affect the function of an organism, and how those changes can have downstream effects on an organism's health and function."

[2 marks]

b) Explain in what way your last answer might be different for some specific viruses, and how can this be useful in a molecular biology laboratory.

[3 marks]

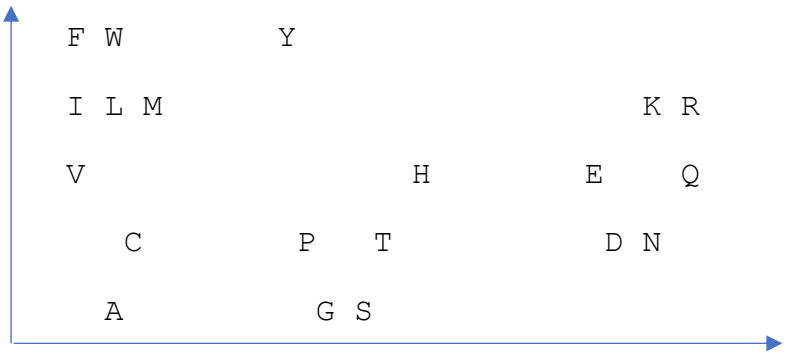


c) Reproduce the figure above with *all* occurrences of the following labels added: parent strand, leading strand, lagging strand, Okazaki fragment, 5', 3'. Note: some labels will need to be used more than once.

[5 marks]

d) Using mass spectrometry, it's possible to measure the abundances of different proteins in a cell, which can be monitored over time. List mechanisms by which protein abundances in a typical eukaryotic cell might be regulated and show where these mechanisms take place on a simple block diagram of the Central Dogma.

[6 marks]



e) Using two physicochemical features, the twenty amino acids are shown on the 2-D scatter plot shown above. Which feature would best label the x-axis and which the y-axis?

[2 marks]

f) Name the mechanism by which one gene can produce multiple different proteins, and explain the process using simple diagrams.

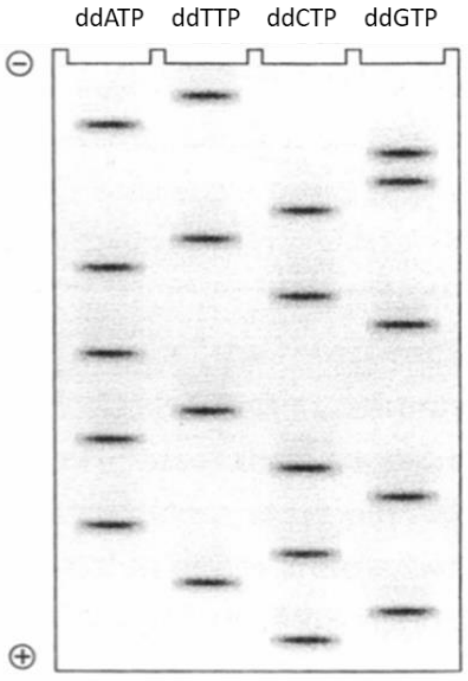
[2 marks]

g) Write a short definition of the term “protein domain” and outline one likely benefit of having domains to the proteins encoded in higher organisms specifically.

[3 marks]

h) Explain, with drawn diagrams, how short stretches of DNA can be sequenced in a lab, including the names of any reagents used. Briefly outline ways in which this basic lab technique can be adapted to increase throughput and produce longer raw reads of DNA sequence.

[8 marks]



i) The image of a Sanger sequencing gel above is that for a strand of cDNA derived from a short mRNA sequence. Write out the two possible mRNA sequences that could have produced this gel. Give your answers in single letter codes and correctly label the 5' and 3' ends of the two RNA sequences.

[3 marks]

[Total for Question 1: 34 marks]

Section B

Answer TWO questions from this section.

a) A short bacterial gene has been sequenced, giving the following DNA sequence. The sequencing gel was difficult to read and an extra base has ended up inserted into the sequence by mistake. Using the standard genetic code, write out the 6 possible reading frames for this sequence and indicate which is the most likely protein translation of this sequence. Explain your reasoning for producing the given translation, and show your working by writing all the translations in single letter amino acid code form.

5′ – atgaataacattggcgaaaacgaaaagcgaacagtaa – 3′

[8 marks]

b) Scientists wish to express a short human peptide "MYFACIL" using E. coli. Calculate how many possible coding gene sequences there are to choose from for this sequence. If the scientists find a gene sequence which expresses well in E. coli, why might the sequence need to be changed if they wished to express the same peptide in an actual human cell?

[4 marks]

c) The scientists need to choose one sequence from all the alternatives available. What approach could the scientists use to find the optimal DNA sequence to synthesise?

[3 marks]

d) Looking at the standard genetic code table, what observation can you make about the nucleotide occurring in the middle position, and the hydrophobic/polar nature of the encoded amino acid? Explain your answer briefly.

[2 Marks]

e) The genome of a newly discovered organism, found in an ocean sample, has just been sequenced and the resulting sequence data has been analysed. Use the following data to calculate the estimated fraction of the genome that is coding (show your working). Based on your result, is the organism more likely to be a eukaryote or prokaryote and justify your choice.

Number of genes predicted in the genome = 2200

Average translated protein length = 230 amino acids

C-value = 1.6×10^6

[3 marks]

f) Briefly discuss issues that would cause problems for finding human genes that would unlikely to be issues for prokaryotic genomes.

[4 marks]

g) Invertebrate mitochondria use a modified genetic code table, which differs from the standard code table as follows:

	Mito	Standard
AGA	Ser	Arg
AGG	Ser	Arg
AUA	Met	Ile
UGA	Trp	Stop

Draw a diagram of a low level state model representation of an exon which can correctly recognise such mitochondrial exons. Hint: you will need to adapt the standard VEIL exon model to handle the above changes to the genetic code.

[9 marks]

[Total for Question 2: 33 marks]

3)

a) A student has written the following incorrect pseudocode for calculating the maximum alignment score between two sequences. What corrections should be made to the code to make it work properly.

```
PROCEDURE NeedlemanWunsch(s1, s2, gap_penalty)
    // Initialize the scoring matrix
    matrix ← ARRAY OF s1.length+1 BY s2.length+1
    FOR i <= s1.length
        matrix[i,0] ← i * gap_penalty
    END FOR
    FOR j <= s2.length
        matrix[0,j] ← j * gap_penalty
    END FOR

    // Fill in the scoring matrix
    FOR i <= s1.length
        FOR j <= s2.length
            match_score ← matrix[i-1,j-1] + (s1[i-1] == s2[j-1] ? -1 : 1)
            delete_score ← matrix[i-2,j] + gap_penalty
            insert_score ← matrix[i,j-1] + gap_penalty
            matrix[i][j] ← MAX OF match_score, delete_score, insert_score
        END FOR
    END FOR
END PROCEDURE
```

[2 marks]

b) Explain the difference between local and global alignments and discuss briefly how the amino acid score matrix should be adjusted to influence the global/local behaviour of the standard Smith-Waterman algorithm.

[3 marks]

c) Outline briefly how programs like FASTA speeds up the searching of sequence data banks compared to basic dynamic programming algorithms.

[3 marks]

d) Briefly describe the two main sources of bias in profile HMMs and what techniques can be used to overcome them.

[4 marks]

e) An alignment of four viral protein sequence motifs is shown below:

```
MIELSK
MNELTK
MLHLTK
MIHLTK
```

Calculate a sequence profile, formatted as 20 rows of 6 columns, for the above small sequence family using the Laplace rule (pseudocount=1) as needed. Give the resulting relative frequencies to 3 d.p. and order the rows in 3-letter amino acid code order (Ala, Arg, Asn … Val). Show your working for the first position.

[5 marks]

f) Look at the following section of a protein structure file in classic PDB format, and give counts of the number of amino acids, number of backbone atoms and the number of carbon atoms present in the data shown.

```
ATOM      1  N   LEU A   7      44.225  -5.302  18.243  1.00  0.00
ATOM      2  CA  LEU A   7      43.210  -4.356  18.822  1.00  0.00
ATOM      3  C   LEU A   7      43.055  -4.551  20.336  1.00  0.00
ATOM      4  O   LEU A   7      43.674  -3.817  21.110  1.00  0.00
ATOM      5  CB  LEU A   7      41.848  -4.516  18.121  1.00  0.00
ATOM      6  CG  LEU A   7      41.894  -4.284  16.606  1.00  0.00
ATOM      7  CD1 LEU A   7      40.507  -4.562  16.000  1.00  0.00
ATOM      8  CD2 LEU A   7      42.398  -2.859  16.315  1.00  0.00
ATOM      9  N   GLY A   8      42.237  -5.510  20.769  1.00  0.00
ATOM     10  CA  GLY A   8      42.096  -5.728  22.203  1.00  0.00
ATOM     11  C   GLY A   8      40.873  -6.472  22.747  1.00  0.00
ATOM     12  O   GLY A   8      40.688  -6.563  23.975  1.00  0.00
ATOM     13  N   GLY A   9      40.050  -7.018  21.857  1.00  0.00
ATOM     14  CA  GLY A   9      38.857  -7.713  22.295  1.00  0.00
ATOM     15  C   GLY A   9      37.662  -6.805  22.034  1.00  0.00
ATOM     16  O   GLY A   9      37.819  -5.581  21.841  1.00  0.00
ATOM     17  N   LEU A  10      36.470  -7.400  22.054  1.00  0.00
ATOM     18  CA  LEU A  10      35.203  -6.709  21.795  1.00  0.00
ATOM     19  C   LEU A  10      34.732  -5.664  22.839  1.00  0.00
ATOM     20  O   LEU A  10      34.237  -4.573  22.479  1.00  0.00
ATOM     21  CB  LEU A  10      34.118  -7.788  21.534  1.00  0.00
ATOM     22  CG  LEU A  10      34.401  -8.654  20.263  1.00  0.00
ATOM     23  CD1 LEU A  10      33.673 -10.043  20.339  1.00  0.00
ATOM     24  CD2 LEU A  10      34.019  -7.822  18.967  1.00  0.00
```

[3 marks]

g) From the same structure, give the distance in nanometres between the first alpha-carbon and the second alpha-carbon atom, showing your working. How much would you expect this distance to vary e.g. for different types of amino acid? Justify your answer.

[3 marks]

h) Describe how from just the multiple sequence alignment of a protein can be used to predict its tertiary and quaternary structure.

[10 marks]

[Total for Question 3: 33 marks]

4)

a) Fill in the blanks in the following passage, choose only six of the following phrases,

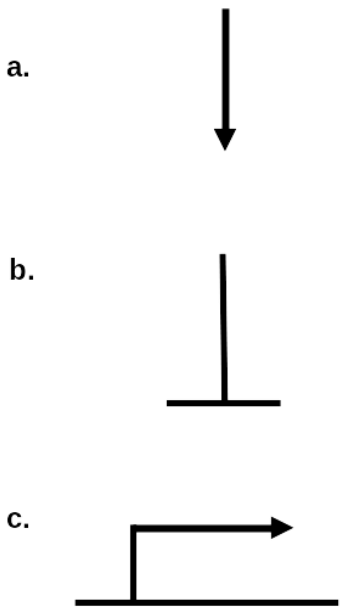
1. "biological"
2. "protein-protein interaction networks"
3. “micro RNA”
4. "metabolic pathways"
5. "transcriptomics"
6. "signalling pathways"
7. “gene regulation networks"
8. “enzymes”
9. “chemistry”

Analysis of _____ networks is necessary when we want to understand complex cellular systems such as metabolism or cell signalling. Four common types of biological network are; _____, _____, _____ and _____. When modelling biological networks we frequently analyse -omics data such as _____ data .

[6 marks]

b) Some gene network diagrams often use the following symbols (a, b and c). Name each and briefly explain what each symbol represents.

[6 marks]



c) You are given the following table, by your colleagues, of normalised expression levels expressed as fold changes of mRNA concentration. You're told this data is derived from a single-cell RNA-Seq experiment. The researchers want to characterise a small gene regulation network to find any transcription factors that control some genes in the organism *Examius examplium*. At time point 0 the researchers take a cell sample and measure the mRNA concentrations. As an experimental treatment they then immediately heatshock the cells. A cell samples is then taken every 5 minutes for 20 minutes and mRNA concentrations are measured. Fold level changes are recorded in the table below:

Genes	0 minutes	5 minutes	10 minutes	15 minutes	20 minutes
COX12	1	1.2	1	8	15
HSP87	1	3	3.2	2.9	3.1
OLI19	1	0.9	2	2.1	1.8
SCEI2	1	1.1	4	4	8
COB00	1	0.8	1	1	0.9

i. Given an empty distance matrix (see below), calculate the pairwise euclidean distance between each gene product. You only need to calculate the lower half of the matrix. Report values to two decimal points.

[4 marks]

	COX12	HSP87	OLI19	SCEI2	COB00
COX12	0	-	-	-	-
HSP87		0	-	-	-
OLI19			0	-	-
SCEI2				0	-
COB00					0

ii. Euclidean distance is not usually used when calculating distance in transcriptomics experiments, name two alternative distance measure that would be better.

[2 marks]

iii. You're asked to hierarchically cluster the gene products. Using single linkage, agglomerative clustering construct and draw a dendrogram of the genes based on the distances in your distance matrix.

[4 marks]

iv. Your colleagues would like to help identify any putative transcription factors in the data. Looking over the expression data do you believe any of the genes in the set may be a transcription factor, which activates the transcription of other genes? Explain your reasoning and name which gene or genes you believe may be transcription factors.

[2 marks]

v. Are there any genes in the set of 5 that you do not believe are part of the gene regulation network that is being studied? Explain your reasoning and name as many genes as you believe are not involved.

[2 marks]

d) MIAME, MAGE-OM and MAGE-TAB are data standards for transcriptomic. Describe what these are, and the purpose each standard. What key types of information are captured by each standard and how they are interrelated?

[7 marks]

[Total for Question 4: 33 marks]

END OF PAPER