UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE : **COMP0082**

ASSESSMENT PATTERN : **COMP0082A7PD**

MODULE NAME : **COMP0082 - Bioinformatics**

LEVEL: : **Postgraduate**

DATE : **17-May-2022**

TIME : **10:00**

Controlled Condition Exam: 2 Hours exam

You cannot submit your work after the date and time shown on AssessmentUCL – 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
2021/22

Additional material	N/A
Special instructions	N/A
Exam paper word count	N/A

TURN OVER

UCL Computer Science Examination paper

Paper details

Academic year:	2021/22
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

Instructions

There are FOUR questions in total. Answer ALL questions from SECTION A and 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) Describe the two key processes by which an amino acid chain is produced, starting from its gene sequence. Your answer should include your own drawn diagrams and you should name the important components of the processes and any organelles involved.

[10 marks]

b) Draw a diagram of a typical worm cell and label its key components and features.

[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.

[4 marks]

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

[2 marks]

e) 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]

f) 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.

[10 marks]

[Total for Question 1: 34 marks]

Section B

Answer TWO questions from this section.

2)

1st Position	2nd Position U	2nd Position C	2nd Position A	2nd Position G	3rd Position
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**

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. 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′ – atgattaacattggcgaaaacgaaaagcgaacagtaa – 3′

[8 marks]

b) A scientist wishes to express a short human peptide “MFYACIL” using E. coli. Calculate how many possible coding gene sequences there are to choose from for this sequence.

[3 marks]

c) The scientist needs to choose one sequence from all the alternatives available. What approach would be used to allow the scientist to find the optimal gene 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 = 200 amino acids

C-value = 1.5×10^6

[3 marks]

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

[4 marks]

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

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

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

[10 marks]

[Total for Question 2: 33 marks]

3)

a) For an alignment of three sequences (*x,y,z*), the match score can be represented as $S(x_i, y_j, z_k)$ in the case of no gaps, or $S(x_i, -, z_k)$ in the case of a gap being placed in the second sequence and so on. Explain briefly how the standard NW algorithm can be modified to deal with the optimal alignment of three sequences, and using similar notation to that above, write out a recurrence formula for the three sequence NW global alignment algorithm, where *M* is the dynamic programming matrix, and the cost of inserting a single gap in one sequence is given by positive value *d*. Assume in your answer that the gap penalty will be applied for every individual gap position inserted i.e. the maximum total gap penalty that can be accumulated at any position in the dynamic programming matrix will be *2d*.

[5 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 the Smith-Waterman algorithm.

[3 marks]

d) Explain briefly how the Viterbi algorithm is used in the context of protein profile HMMs.

[3 marks]

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

[4 marks]

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

```
MIELSL
MNELTL
MLHLTL
MIHLTL
```

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 column.

[5 marks]

g) Explain how Sparse Inverse Covariance Estimation (SICE) can be used to predict protein tertiary structure by exploiting patterns of amino acid covariation in large protein multiple sequence alignments.

[10 marks]

[Total for Question 3: 33 marks]

4)

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) In addition to the sequence itself, give three different types of information that can be extracted from the header records of an entry in a protein sequence data bank.

[3 marks]

c) 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.

```
ATOM      1  N   MET     1      -5.655  30.682  -6.383  1.00
ATOM      2  CA  MET     1      -5.921  32.127  -6.684  1.00
ATOM      3  C   MET     1      -5.919  30.804  -5.949  1.00
ATOM      4  O   MET     1      -6.608  30.126  -5.631  1.00
ATOM      5  CB  MET     1      -6.678  32.268  -8.006  1.00
ATOM      6  SD  MET     1      -4.414  32.838  -9.497  1.00
ATOM      7  CE  MET     1      -3.147  31.785  -8.795  1.00
ATOM      8  CG  MET     1      -5.885  31.831  -9.226  1.00
ATOM      9  N   THR     2      -4.898  30.272  -5.568  1.00
ATOM     10  CA  THR     2      -4.661  29.030  -4.859  1.00
ATOM     11  C   THR     2      -4.211  29.596  -3.536  1.00
ATOM     12  O   THR     2      -3.488  30.229  -3.203  1.00
ATOM     13  CB  THR     2      -3.633  28.149  -5.593  1.00
ATOM     14  OG1 THR     2      -4.126  27.821  -6.898  1.00
ATOM     15  CG2 THR     2      -3.395  26.857  -4.825  1.00
ATOM     16  N   GLU     3      -4.758  29.316  -2.495  1.00
ATOM     17  CA  GLU     3      -4.504  29.730  -1.135  1.00
ATOM     18  C   GLU     3      -4.007  28.551  -0.298  1.00
ATOM     19  O   GLU     3      -4.256  27.570  -0.162  1.00
ATOM     20  CB  GLU     3      -5.767  30.330  -0.513  1.00
ATOM     21  CD  GLU     3      -7.479  32.186  -0.525  1.00
ATOM     22  CG  GLU     3      -6.240  31.610  -1.182  1.00
ATOM     23  OE1 GLU     3      -8.030  31.530   0.384  1.00
ATOM     24  OE2 GLU     3      -7.900  33.294  -0.920  1.00
```

[3 marks]

d) 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]

e) Explain the working of the ProtFun method for predicting the function of proteins without homology to already characterised proteins. What difficulties might you run into if you were to try to extend the method to the whole of the Gene Ontology (GO).

[10 marks]

f) Outline what procedures might be needed to develop a classifier capable of separating eukaryotic from prokaryotic amino acid sequences, based solely on amino acid sequence. In your answer identify which features and data resources (for training data and labels) you might use, and how you would run the experiment in order to estimate the accuracy of the method on unseen data.

[10 marks]

[Total for Question 4: 33 marks]

END OF PAPER