ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION WINTER SEMESTER, 2022-2023

DURATION: 3 HOURS FULL MARKS: 150

CSE 4753: Bioinformatics

Programmable calculators are not allowed. Do not write anything on the question paper.
Answer **all 5 (five)** questions. Figures in the right margin indicate full marks of questions whereas corresponding CO and PO are written within parentheses.

---

1. a) TCGA gene expression dataset for a set of tumor samples were downloaded as a Summarized Experiment object in R environment. The object has six assays as in Table 1. 3x5 (CO3) (PO2)

Table 1: Object assays for Question 1.a)

unstranded	stranded_first	stranded_second
tpm_unstrand	fpkm_unstrand	fpkm_uq_unstrand

i. Which assay will you choose to find differentially expressed genes using DESeq method? Justify your answer.

ii. Which assay will you use to develop a tumor classification model? Explain your choice.

iii. How do you preprocess the assay data to be used in Question 1.a)i. and ii.?

b) A trained machine learning model generates a confusion matrix for a test dataset as in Table 2. 10 (CO3) (PO2)

Table 2: Confusion matrix for Question 1.b)

```
[[ 38  1   0   0  0 ]
 [  0  16  0   0  0 ]
 [  0  0   111 2  1 ]
 [  0  1   7   34 0 ]
 [  0  0   1   0  7 ]]
```

Generate classification report from the confusion matrix showing class-wise, macro average and weighted average of precision, recall and F1-score. Also generate model's accuracy over test dataset.

c) *Nature has developed a way to reduce the effects of random mutation in genes during translation process of central-dogma* – explain. 5 (CO2) (PO1)

2. a) DBSCAN is a density based algorithm for clustering. Explain how this method works along with defining the basic terms. 8 (CO3) (PO2)

b) A dataset is given for clustering. An analyst performs k-means algorithm five times that yields five different solutions. The analyst has to pick the best clustering outcome. How could (s)he pick the best solution from the five different outcomes? 8 (CO3) (PO2)

c) Similarity matrix for five data points are presented in Table 3 below. Cluster these dataset by applying agglomerative hierarchical clustering using MAX (complete linkage) approach. 10 (CO2) (PO1)

Table 3: Similarity matrix for Question 2.c)

	A	B	C	D	E
A	1.00	0.75	0.65	0.90	0.85
B		1.00	0.35	0.45	0.55
C			1.00	0.70	0.65
D				1.00	0.75
E					1.00

d) Describe InDel scenario lead by point mutation. 4 (CO2) (PO1)

3. a) *Micro RNA (miRNA), long-non-coding RNA (lncRNA), and pseudogenes RNA can affect the process of central-dogma by cross-talk* – explain. 10 (CO1) (PO1)

b) Explain how the imbalance in a dataset, i.e., unequal sample count for different classes affects performance of classification of samples by Machine Learning (ML) models. 5 (CO3) (PO1)

c) What are the ways to reduce the imbalanced property of a dataset before training ML models to reduce its effect on classification performance? 10 (CO3) (PO1)

d) Which information do global alignment and local alignment of biological sequences generate? 5 (CO2) (PO1)

4. a) Discuss various secondary RNA structures developed by RNA folding. 10 (CO1) (PO1)

b) Distance matrix for 4 hypothetical sequences- *S1, S2, S3* and *S4* is presented in Table 4. Build a phylogenetic tree for the sequences using Fitch-Margoliash Algorithm. 10 (CO3) (PO1)

Table 4: Distance matrix for Question 4.b)

	S1	S2	S3	S4
S1	-	10	8	11
S2		-	12	7
S3			-	9
S4				-

c) Explain gene expression regulation mechanism in terms of regulatory sites and transcription factors. 10 (CO2) (PO1)

5. a) Two sequences are given as follows: 13 (CO4) (PO3)

Sequence1: TATGCTAAC
Sequence2: GCATGCTAC

And the substitution matrix is in Table 5:

Table 5: Substitution matrix for Question 5.a)

	A	T	G	C	-
A	1	-2	-2	-2	-1
T	-2	1	-2	-2	-1
G	-2	-2	1	-2	-1
C	-2	-2	-2	1	-1
-	-1	-1	-1	-1	

Align these two sequences locally and explain the alignment result.

b) An amino acid chain is given below: 17 (CO5) (PO2)

T S P T A E L M R S T G

Using Chao-Fasman method, determine which secondary protein structure is likely to occur in the sequence. Propensity values for α-helix, β-sheet, and turn for different amino acids are shown in Table 6.

Table 6: Propensity values of amino acids for Question 5.b)

	Amino Acid			P(α-helix)	P(β-sheet)	P(turn)
1	Alanine	ala	A	1.42	0.83	0.66
2	Arginine	arg	R	0.98	0.93	0.95
3	Asparagine	asn	N	0.67	0.89	1.56
4	Aspartic Acid	asp	D	1.01	0.54	1.46
5	Cysteine	cys	C	0.70	1.19	1.19
6	Glutamine	gln	Q	1.51	0.37	0.74
7	Glutamic Acid	glu	E	1.11	1.11	0.98
8	Glycine	gly	G	0.57	0.75	1.56
9	Histidine	his	H	1.00	0.87	0.95
10	Isoleucine	ile	I	1.08	1.60	0.47
11	Leucine	leu	L	1.21	1.30	0.59
12	Lysine	lys	K	1.14	0.74	1.01
13	Methionine	met	M	1.45	1.05	0.60
14	Phenylalanine	phe	F	1.13	1.38	0.60
15	Proline	pro	P	0.57	0.55	1.52
16	Serine	ser	S	0.77	0.75	1.43
17	Threonine	thr	T	0.83	1.19	0.96
18	Tryptophan	trp	W	0.83	1.19	0.96
19	Tyrosine	tyr	Y	0.69	1.47	1.14
20	Valine	val	V	1.06	1.70	0.50