Name:
NetID:
Discussion Section:
Course Heading:

1. Genomics

(a) List 3 cis-regulatory DNA elements (5pt).

promoter, enhancer, insulator

-1.5 for each wrong
-0.5 if 3 right but wrote >3 things

(b) Name 2 types of long-read sequencing technologies. Then list 2 differences between long-read sequencing technology and normal sequencing technology (5pt).

-1 for each incorrect LR sequencing
-1.5 for each incorrect difference

(c) What is the typical number of single nucleotide polymorphism (SNP) in one person's typical genome with respect to the human reference genome? (5pt)

A. ~4,000
B. ~40,000
C. ~400,000
D. ~4,000,000
E. ~40,000,000

(d) Consider an European individual versus an African one, which person would you expect to have more SNVs relative to the reference genome? (5pt)

African
-5 if wrote nothing or just "European", -4 if wrote european but gave an (incorrect) explanation

(e) Choose the sequencing methods and their applications (NOTICE: you may reuse the options) (5pt):

Localization of transcription factors: b
Chromatin accessibility: c
Differential expression analysis: a
Determination of alternative splicing: a

-1 pt for each incorrect

(a) RNA-seq (b) ChIP-seq (c) DNase-Seq

2. Proteomics

(a) Circle all methods to identify protein structures (5pt)

A. X-ray crystallography
B. NMR
C. Mass Spectrometry
D. Cryo-EM

-1 pt for each incorrect answer (max of -4 pt in this case)

(b) Compared to sequencing of DNA, why is proteomic analysis more dependent on sample abundance? (5pt)

Needs to say that proteins cant be amplified

-5 if no answer
-4 or -3 if wrote something but wrong (discretionary)
-1 if answer is right but lacks detail

3. Sequence comparison.

(1) Compute the position probability profile matrix for the following nucleotide sequences (you just need to consider the simplest case): (5pt, -1pt for one mistake)
CTTCAG
CTGGCT
ATGCCT
ATGGCG

A	1/2				1/4	
T		1	1/4			1/2
G			3/4	1/2		1/2
C	1/2			1/2	3/4	

-1 pt for each mistake

(2) What is the probability of observing sequence ATTCCG, given the profile matrix in (1)? Note: Please write out the expression. You do not have to calculate the exact number. 5pts

½ x 1 x ¼ x ½ x ¾ x ½

-5 pts if wrote nothing
-4 pts if wrong expression
-1pt if minor mistake

4. A relational database has the following columns:

STUDENT ID	NAME	SUBJECT	STATE	COUNTRY	Zip_code
29	Jennifer	English	CT	USA	06510
33	Katrina	Music	CA	USA	06513
28	Zayed	Math	FL	USA	07511
79	Rameo	Chemistry	TX	USA	08911

(a) Explain why this table is not in third normal form (5pt)
(b) Explain and draw columns for a new table (or tables) for how you would normalize this database further (to third normal form): (5pt)

STUDENT ID	NAME	SUBJECT	Zip_code

Zip_code	COUNTRY	STATE

Also OK for 1st table:

STUDENT ID	NAME	SUBJECT	COUNTRY	Zip_code

(a) -5 pt if nothing written, -4 pt if written wrong answer, -1 pt for minor mistake
(b) -5 pt if nothing written, -4 pt if written wrong answer, -1 pt for minor mistake

5. Given the following confusion matrix, select ALL the statements that accurately define sensitivity and specificity using TP, TN, FP, and FN. (5pt)

	Predicted Positive	Predicted Negative
True	TP	FN
False	FP	TN

-2.5 pt for each incorrect

A. Sensitivity = TP / (TP + FN)
B. Specificity = TN / (TN + FP)
C. Sensitivity = TP / (TP + FP)
D. Specificity = TN / (TN + FN)
E. None of the above

6. In SVD, the data matrix A is decomposed as A = USV'. Suppose A is a 5x22 matrix. What are the dimensions of U, S, V respectively? (5pt)

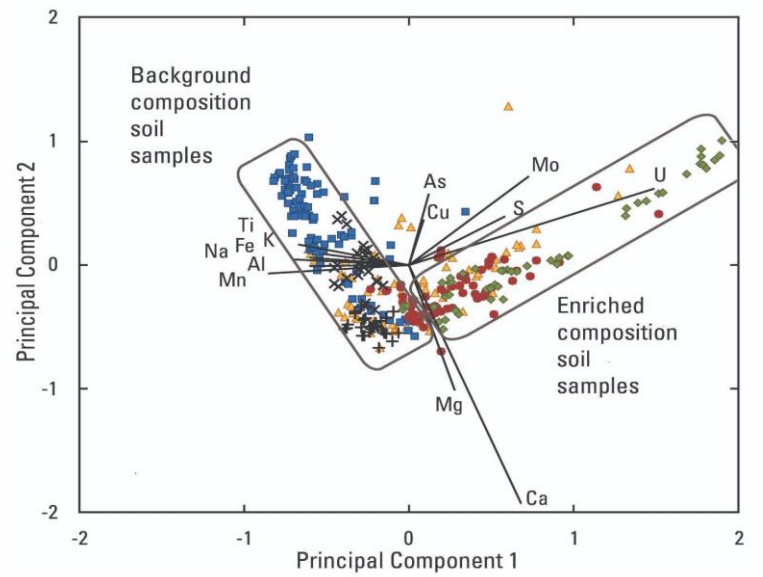
U 5x5
S 5x22
V 22x22

OR

U 5xn
S nxn
V 22xn

-5 pt if nothing written,
-1 pt for each dimension mistake, capped at 5 pts (6 total dimensions; 2 each for U,S,V)

7.



For the above biplot, assume that the first 2 components account for 95% of the variance in the original dataset. Answer **True or False**: (2.5pt x4)

(a) As is strongly correlated with Mn False
(b) As projects on to PC2 stronger than Mo False
(c) Ca is strongly correlated with Mg True
(d) As strongly projects onto principal component 1 False

-2.5 pt for each incorrect

8. Align the following two sequences using the Needleman-Wunsch global alignment algorithm (Smith-Waterman algorithm will also be

acceptable, if you choose Smith-Waterman, you could draw a new alignment table by yourself). Show the complete dynamic programming matrix, and circle one optimal traceback on the matrix (15pt).
Sequence 1: ACTGCA
Sequence 2: ACATGA

Use the following scoring scheme in the score matrix:
Match: +3
Mismatch: 0
Gap: -2

	A	C	T	G	C	A
A						
C						
A						
T						
G						
A						

SW:

	A	C	T	G	C	A
A	1 1	5	4	6	1	3
C	7	8	2	1	6	0
A	8	7	3	1	1	3
T	2	2	7	3	1	0
G	1	1	1	4	3	0

A	3	0	0	0	0	3

NW:

	A	C	T	G	C	A
A	3	1	-1	-3	-5	-7
C	1	6	4	2	0	-2
A	-1	4	6	4	2	3
T	-3	2	7	6	4	2
G	-5	0	5	10	8	6
A	-7	-2	3	8	10	11

15 pts
7.5pt for matrix: -1 pt for minor mistakes in filling matrix
7.5pt for traceback: -1 pt for minor mistakes in traceback

9. Position weight matrix (PWM) is commonly used to represent motifs (patterns) in biological sequences. Describe the main steps of using EM algorithm to update position weight matrix (10pt)

1. Guess an initial weight matrix

2. Use weight matrix to predict instances in the input sequences
3. Use instances to predict a weight matrix
4. Repeat 2 [E-step] & 3 [M-step] until satisfied

Key points:
Initial 2pts;
E-step 3pts;
M-step 3pts;
Repeat E and M 1pt;
End when satisfied 1pt;

-8 pt wrote something wrong
-5 pt only did half the problem
-1 pt for each minor mistake

10. Explain why the kernel function is important for non-linear SVM (5pt)?

important to include something about non separable by a linear fx/plane and requires nonlinear

SVM can be extended to solve nonlinear classification tasks when the set of samples cannot be separated linearly. By applying kernel functions, the samples are mapped onto a high-dimensional feature space, in which the linear classification is possible.

-5 pt if written nothing
-4 pt if written wrong answer
-1 pt if not detailed or minor mistake

11. Construct an optimal decision tree (with depth of 2) based on the following input data to predict patient survival. Partial credit for any reasonable decision tree. (10 pt)

Tumor degree	Tumor Size	Age	Patient survival
I	10	40	Yes
II	10	60	Yes
I	13	40	No
II	50	50	No

-5 pt if wrote a tree that doesn't work

12. One primary reason that makes single-cell RNA-seq analysis challenging is doublet. What is “doublet” in single-cell analysis? How do people usually overcome the doublet problem (5pt)?

-5 pt if nothing written
-4 if incorrect answer
-1 if answer mostly correct but minor mistake

13. Describe two clustering methods and the pros/cons of each method? (5pt)

1 pt each for clustering method named (tsne pca and other dim reduc methods NOT clustering method)
1.5 pt for pros/cons list for each method

14. Bonus question
Describe the workflow of how to find differentially expressed genes in cell types starting from the count matrix in single-cell experiments (10pt)
3 pts if given general single cell workflow but no highlight of DE genes
5pts given if general workflow described- needs to have something about DE genes in cell types
10 pts given if very detailed and includes something like deseq or a way to get DE genes