Biomedical Data Science: Mining & Modeling | Spring 2020
Quiz 1

Name:

1. a) Which concept below does NOT help optimize database queries? (5 points)

A. Database index
B. Hashing functions
C. Un-normalized database
D. Normalized database

C.

b) Suppose Camila, a professor in the MB&B department, has recently moved to Ann Arbor for a one-year sabbatical to pursue research, changing zip codes from 06511 to 48103. The registrar made ***some*** update in the "ZIP Codes" table to reflect this move, but failed to make a complete update in the "Professor Information" table.

What is the error below in the "Professor Information" table? and what concept(s) from relational databases could have prevented it? (5 points)

Table 1: Professor Information

Professor	State	City	Zip_code
Camila	MI	New Haven	06511
Mark	CT	New Haven	06511
Yuzhen	CT	New Haven	06511

Table 2: ZIP Codes

Zip code	City
06511	New Haven
48103	Ann Arbor

Error:

Data redundancy, update not reflected in Table 1.

Preventative concept:

Normalization, avoidance of data redundancy, similar answers are acceptable.

2. Define “Proteomics.” (5 points)

“The study of the expression, location, interaction, function, and structure of all the proteins in a given cell, organelle, tissue, organ, or whole organism”.

Mention of protein: 1 pt
Two or more from “expression, location, interaction, function, and structure”: 3 pts
Two or more of “a given cell, organelle, tissue, organ, or whole organism”: 1 pt

3. a) Fill in the two remaining blanks to complete the list of main steps in X-Ray structure crystallization. (5 points)

Subcloning > _______________ > _______________ > Crystallization

Expression > Purification

b) Circle all methods to identify protein structures. (5 points)
A. X-ray crystallography
B. NMR
C. Mass Spectrometry
D. Cryo-EM
E. SILAC

A, B and D

2pts per right answer

-1 pt per wrong answer

4. What is a major advantage and a major disadvantage of long read sequencing? (10 points)

Main advantage: long reads to detect genomic structural variants
Main disadvantage(s): high sequencing error, high cost, low depth

5. What is the key difference between a PAM-50 and a PAM-500 substitution matrix? Human and mouse diverge at around 80 million years ago. Which one of these two matrix would you prefer to use to align a human protein and a mouse protein and why? (10 pts)

When building the matrix, it used different gold standard sets of sequences at different **evolutionary distances**. (4 pts)
PAM-50 (4 pts) , as it is closer to the diverge time of human and mouse (2 pts)

6. Position Probability Matrix (PPM) is commonly used to represent motifs in biological sequences. Given the following DNA sequences, calculate the position probability matrix profile. (10 pts)

DNA 1: AAGGTTAA
DNA 2: TACGTTCA
DNA 3: GACGTAGA
DNA 4: CCCGTTTA
DNA 5: AAGGTCAA

A	0.4	0.8	0	0	0	0.2	0.4	1
T	0.2	0	0	0	1	0.6	0.2	0
C	0.2	0.2	0.6	0	0	0.2	0.2	0
G	0.2	0	0.4	1	0	0	0.2	0

7. Order the following algorithms/schemes by running speed. Profiles, BWA, Smith-Waterman, HMM, Blast by increasing the speed, what is the tradeoff? (10 pts)

BWA > Blast > SW > Profiles > HMM

Sensitivity

8. Align the following two sequences using the Needleman-Wunsch global alignment algorithm. Show the complete dynamic programming matrix and arrows, and circle one optimal traceback on the matrix. (15 points)

Sequence 1: CTGCAG

Sequence 2: ACTCAG

Use the following scoring scheme in the score matrix:

Match: +2

Mismatch: 0

Gap: 0

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

Bottom-Up:

	C	T	G	C	A	G
A	8	6	6	4	4	0
C	10	6	6	6	2	0
T	6	8	6	4	2	0
C	6	4	4	6	2	0
A	2	2	2	2	4	0

G	0	0	2	0	0	2

Alternative 1 (Top-Bottom)

	C	T	G	C	A	G
A	0	0	0	0	2	0
C	2	0	0	2	0	2
T	0	4	2	2	2	2
C	2	2	4	6	4	4
A	0	2	4	4	8	6
G	0	2	6	4	6	10

Alternative 2 (Top-Bottom)

		C	T	G	C	A	G
	0	0	0	0	0	0	0
A	0	0	0	0	0	2	2
C	0	2	2	2	2	2	2
T	0	2	4	4	4	4	4
C	0	2	4	4	6	6	6
A	0	2	4	4	6	8	8
G	0	2	4	6	6	8	10

Final alignment:
-CTGCAG
ACT-CAG

10 pts for matrix arrows and values (-2pts on every alignment mistake, i.e. wrong arrow)
5 pts for alignment (-1 for each mistake in translating arrows to alignment)

9. a) Given the following confusion matrix, select **all** the statements that accurately define sensitivity and specificity using TP, TN, FP, and FN. (5 points)

	Predicted Positive	Predicted Negative
True	TP	FN
False	FP	TN

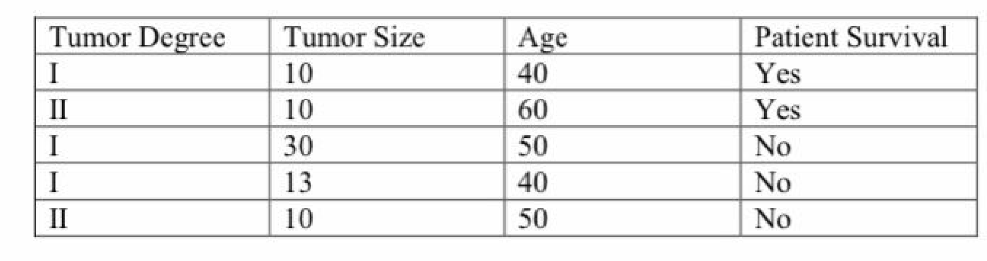
A. Sensitivity = TP / (TP + FN)
B. Specificity = TN / (TN + FP)
C. Sensitivity = TP / (TP + FP)
D. Specificity = TN / (TN + FN)
E. None of the above

A and B
5 pts or 0 pts

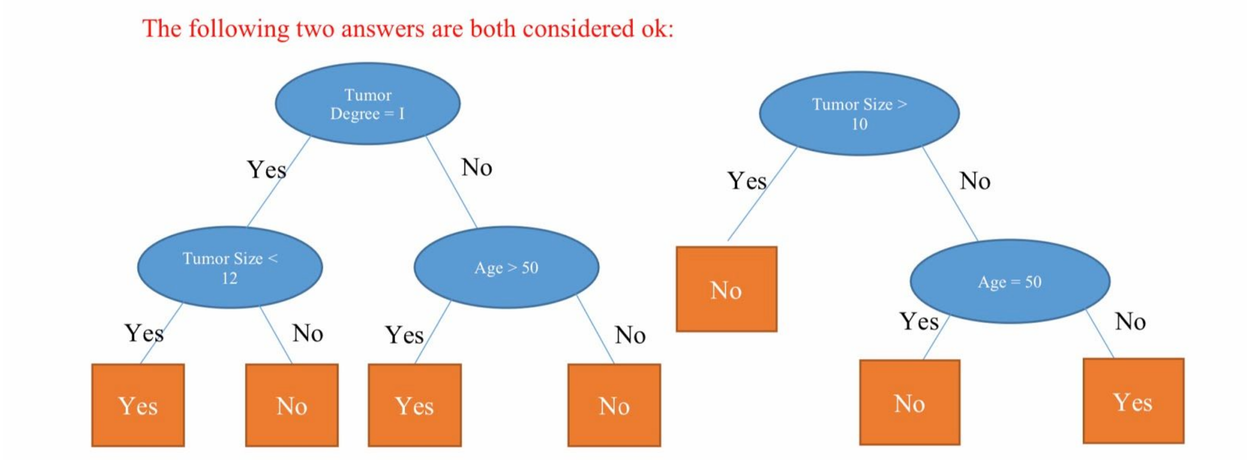
b) A group of biomedical researchers are designing a model to predict tumor state—benign (label 0) or malignant (label 1). The priority of the model is to avoid missing a malignant tumor in predictions. Does that indicate the model must have high sensitivity or specificity? (5 points)

Sensitivity

10. 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 pts)



Tumor Degree	Tumor Size	Age	Patient Survival
I	10	40	Yes
II	10	60	Yes
I	30	50	No
I	13	40	No
II	10	50	No



The following two answers are both considered ok:

Bonus Question. When doing RNA-seq, the basic unit is transcript or gene. In ChIP-seq, the basic unit is peak. What is the “basic unit” in Hi-C experiments and what’s the important measurement we used to find these units from the connecting network? (5 points)

TAD (Topologically Associating Domain)
Network modularity