Name:

NetID:

Discussion Section:

Course Heading:

Keep your answers concise and to the point.

Long responses won't earn extra credit.

NetID: Score:

Genomics:

1) The most raw form of data from an Illumina sequencer is a series of _____. (4pt)

Images

2) sequencing technique application: (8pt)

_____-Seq is used to identify where transcription factors bind to DNA.

_____-Seq measures DNase I hypersensitive sites to identify regulatory elements.

_____ is a technique used to measure chromatin conformation and 3D genome organization.

_____-Seq can identify open chromatin regions.

ChIP

DNase

Hi-C

ATAC / DNase / FAIRE

3) In Illumina sequencing, what is the purpose of cluster amplification? (4pt)
(A) To increase the overall length of DNA fragments
(B) To separate individual molecules and give each a spatial address while avoiding overlaps
(C) To convert RNA to DNA
(D) To remove adapter sequences
B

4) What are potential sources of bias during library preparation using ligation? (4pt)
(A) Selective PCR amplification
(B) Size selection
(C) Enzyme specificities
(D) All of the above
D

NetID: Score:

5) What is the approximate size of the human haploid genome? (4pt)
(A) 300 Mb
(B) 3 Gb
(C) 30 Gb
(D) 300 Gb

B

6) Spatial transcriptomics methods are valuable because they: (3pt)
(A) Preserve information about where transcripts are located within tissue
(B) Provide higher sequencing depth
(C) Are cheaper than standard RNA-seq
(D) Require fewer cells

A

7) What is the output from an Illumina sequencing experiment? List 4 components in one read (fastq format). (8pt)

1. _____
2. _____
3. _____
4. _____

Answer:
1. Read identifier
2. Sequence
3. Quality score identifier "+"
4. Quality score

Each wrong answer -2,

8）What is the primary reason proteomics cannot achieve the same depth of coverage as genomics? (3pt)
A) Mass spectrometers are too expensive for widespread use
B) There is no equivalent to PCR amplification for proteins
C) Proteins are too small to detect accurately
D) The human genome has not been fully sequenced

Answer: B

9) What was a major biological conclusion from systematic protein-protein interaction mapping in both yeast and humans? (4pt)

A) Most proteins function independently without interacting with other proteins
B) Cellular proteins are organized into complexes, and this organization is conserved across species
C) Protein interactions only occur during cell division
D) Human cells have fewer protein interactions than yeast cells

Answer: B

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

ABD

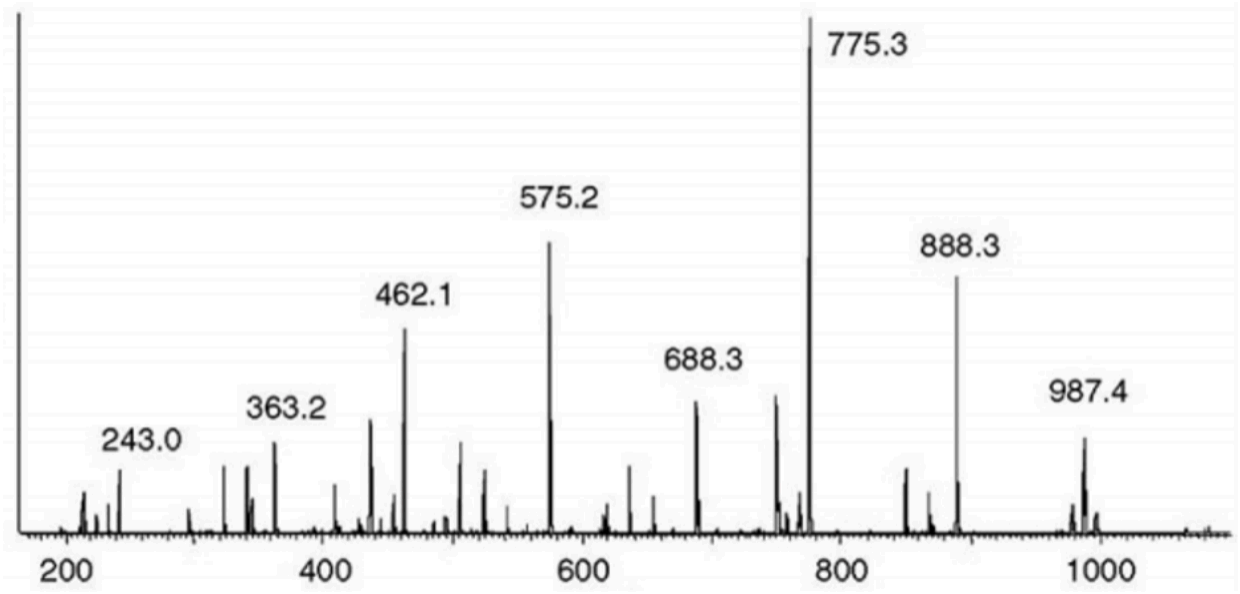
11) Fill in the two remaining blanks to complete the list of main steps in X-Ray structure crystallization. (5 points)

Subcloning > _____ > _____ > Crystallization

Expression > Purification

1 wrong -3

12) Below is a mass spectrum result. What are the x-axis and y-axis? (5pt)



Answer:

X = m/z (mass-to-charge ratio)
Y = Relative abundance or Intensity

-1 if miss “relative”

Database

13) The SQL command used to combine rows from two or more tables based on a related column is called _____ (3pt)
Join

14) _____ is the process of organizing data in a relational database to minimize data redundancy and inconsistent dependency (3pt)
Normalization
(1NF/2NF/3NF -2pt)

15) Link each ACID property to its description (8pt):

Property	Description
Atomicity	Concurrent transactions do not interfere with each other

Consistency	A committed transaction's effects are permanent, even after system failure
Isolation	If any part of a transaction fails, the entire transaction is rolled back
Durability	After a transaction completes, all data rules (e.g., no negative bank balance) must hold

A - 3
C - 4
I - 1
D - 2

Each wrong link - 2

16) Choose False and True for each statement (8pt)

(a) [False / True] An EHR system where physicians enter medication orders in real time is an example of an OLTP system.

(b) [False / True] A data warehouse used for analytical queries and business reports is an example of an OLTP system.

(c) [False / True] Referential integrity means that a foreign key column can only contain values that exist in the referenced column of another table.

(d) [False / True] A primary key can contain NULL values as long as no two rows have the same key.

T,, F, T, F

NetID: Score:

17) Sum Matrix Cell Calculation
Given this partial dot plot matrix (1 = match, 0 = mismatch):

	C	A	T
C	1	0	0
A	0	1	0
G	0	0	0
T	0	0	1

Using the recurrence relation from lecture (no gap penalty):

```
new_value_cell(R,C) = cell(R,C) + Max{
        cell(R+1, C+1), {diagonal}
        cells(R+1, C+2 to C_max), {horizontal}
        cells(R+2 to R_max, C+1) {vertical}
}
```

(A)The bottom two rows of the sum matrix are already computed, fill in the rest of the cells: (5pt)
Your Answer:

	C	A	T
C			
A			
G	1	1	0
T	0	0	1

	C	A	T
C	3	1	0
A	1	2	0
G	1	1	0
T	0	0	1

Each wrong number -1

(B) Write down the alignment results (4 pt)

Your Answer:

```
C A  -  T
|  |     |
C A G T
```

No gap - 2
Wrong gap position -2
Missed G - 2

NetID: Score:

18) Multiple Sequence Alignment

(A) Write the E-step and M step in Expectation-Maximization (EM) algorithm (5pt)

1. Guess an initial weight matrix
2. Use weight matrix to _____ in the input sequences
3. Use _____to _____
4. Repeat 2 [E-step] & 3 [M-step] until satisfied.

Predict Instance
Instances
Predict a weight matrix

(B) Compute the position probability 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					
T		1				
G						
C	1/2					

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

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	

ATTCCG
½ * 1 * ¼ * ½ *¾ * ½