CBB 752 Spring 2016 Midterm Exam 03/07/2016

Name: _____

1. 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 | TN |
| False | FP | FN |

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

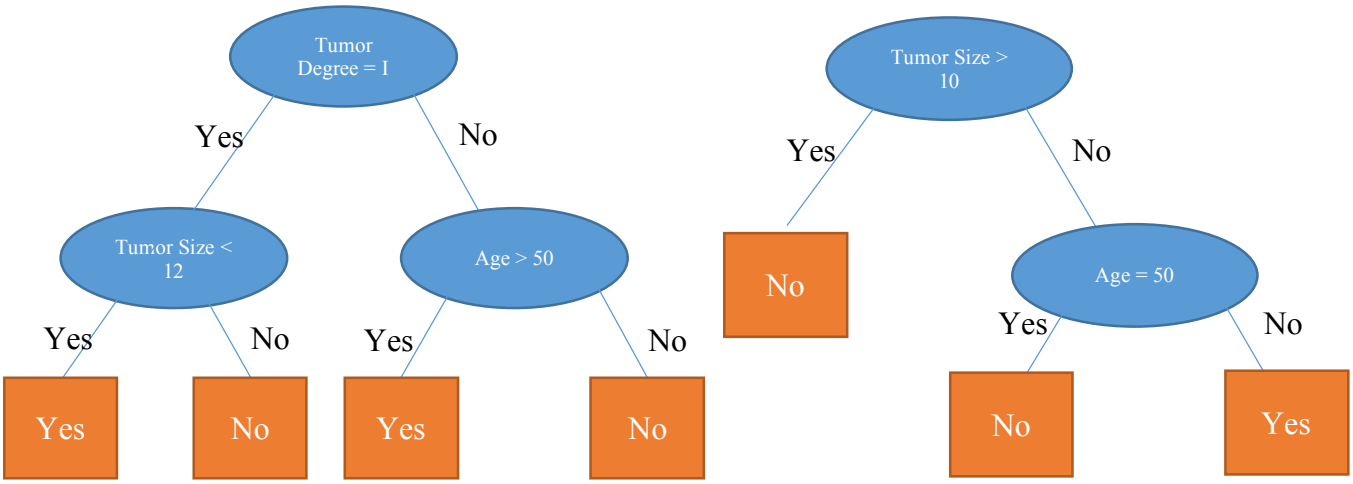
**A, B OR E**

(Both AB and E are considered as the correct answer, since the TN and FN labels were swapped in the given table.)

2. Decision Tree: [5 points]
Imagine that you are a computer. 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):

| 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:



Name: _____

3. Running the following SQL statement on the left generates the DB table on the right. Construct an SQL query that lists all patients whose age is 40 or above. [Optional for MCDB/MBB, 5 points]

```
CREATE DATABASE Patient_DB;

CREATE TABLE Patient_DB.Patient
(
    ID int,
    Name varchar (50),
    Address varchar (250),
    Age smallint
    Sex varchar (2)
);
```

| ID | Name | Address | Age | Sex |
|---|---|---|---|---|
| 1 | John Doe | XYZ | 40 | M |
| 2 | Jane Smith | ABC | 34 | F |
| 3 | Mary Queen | PQSRT | 46 | F |
| 4 | Mike Lee | DWQER | 60 | M |

```
SELECT ID, Name, Age, Sex
FROM Patient_DB.Patient
WHERE Age>=40
```

4. Describe at least one of the differences between X-ray crystallography and NMR. [Optional for CBB/CPSC, 5 points]

**X-ray:**
- **Direct detection of atom positions**
- **Crystals**

**NMR:**
- **Indirect detection of H-H distances**
- **In solution**

5. Place the following NGS steps in the correct order. [5 points]

A. Sequencing
B. Library preparation
C. Computational analysis
D. Isolation of sample

**D -> B -> A -> C**

Name: _____

6. The following steps describe an approach for discovering motifs using position weight matrix (PWM).

Step 1. Guess an initial weight matrix
Step 2. Use weight matrix to predict instances in the input sequences
Step 3. Use instances to predict a weight matrix
Step 4. Repeat 2 & 3 until satisfied

What is the name of algorithm used in this example? [5 points]

**Expectation–maximization algorithm**

7. The first principal component produced by Principal Components Analysis (PCA) maximizes _____. [5 points]

**Variance**

8. What is the key difference between a PAM-50 and a PAM-500 substitution matrix? [5 points]

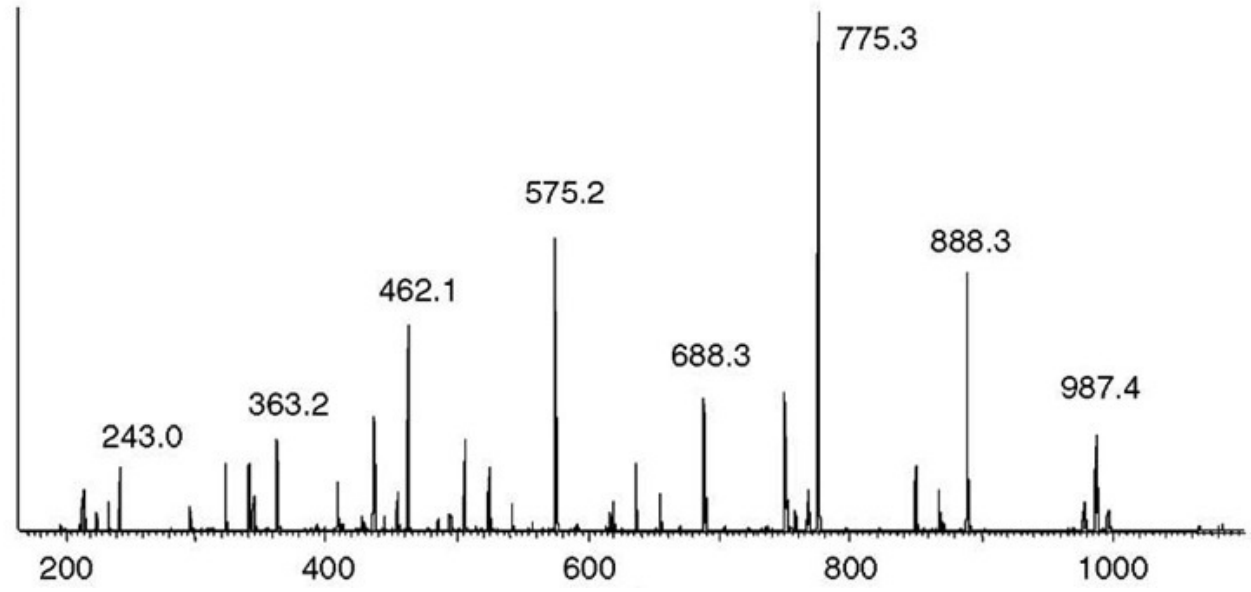
**The PAM-500 matrix represents tolerated amino acid substitutions at ten times the evolutionary distance of the PAM-50 matrix.**

9. SILAC uses _____ to label cells grown in different conditions. (Hint: what does SILAC stand for?) [5 points]

**(non-radioactive) stable isotope**

Name: _____

10.




Mass spectrometry is a powerful technique to differentiate distinct proteins. A sample mass spectrum is shown above. What does X-axis and Y-axis of mass spectrum denote? [5 point]

**X-axis:** **The mass-to-charge ratio (m/z)**

**Y-axis:** **Relative abundance**

11. Illumina reads are generally 50nt - 250nt in length. What limits from producing longer read? [5 points]

**Base quality score tends to drop towards the end. Later base calls are more error-prone.**

12. List at least two things that can be studied with RNA-seq. [5 points]

**Any two of the following:**
- **Gene expression level (quantification)**
- **Alternative splicing**
- **Novel transcripts**
- **Gene fusions**
- **RNA editing**

Name: _____

13. List at least two limitations of ChIP-seq? [5 points]

**Any two of the following:**
- **Cross linking efficiency is not necessarily uniform.**
- **Enrichment is dependent on the quality of antibody. e.g., Site and degree of histone modifications.**
- **Enrichment is dependent on the accessibility of the epitope.**
- **Output is descriptive. Hard to infer function without more experimentation.**

14. Position Weight Matrix: [10 points]

Given the following DNA sequences, complete the corresponding position probability matrix profile. Using the profile, calculate the probability of the sequence *S* = GAGGT being observed.

```
DNA 1: GAGGT
DNA 2: TCCGT
DNA 3: CAGGT
DNA 4: ACAGT
DNA 5: TAGGT
DNA 6: TAGGT
DNA 7: ATGGT
DNA 8: CAGGT
```

| | Pos 1 | Pos 2 | Pos 3 | Pos 4 | Pos 5 |
|---|---|---|---|---|---|
| A | 0.25 | **0.625** | **0.125** | **0** | **0** |
| C | 0.25 | **0.25** | **0.125** | **0** | **0** |
| G | 0.125 | **0** | **0.75** | **1** | **0** |
| T | 0.375 | **0.125** | **0** | **0** | **1** |

**0.125 * 0.625 * 0.75 * 1 * 1 = 0.05859375**

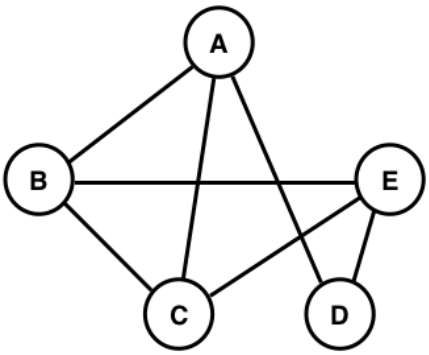
**or**

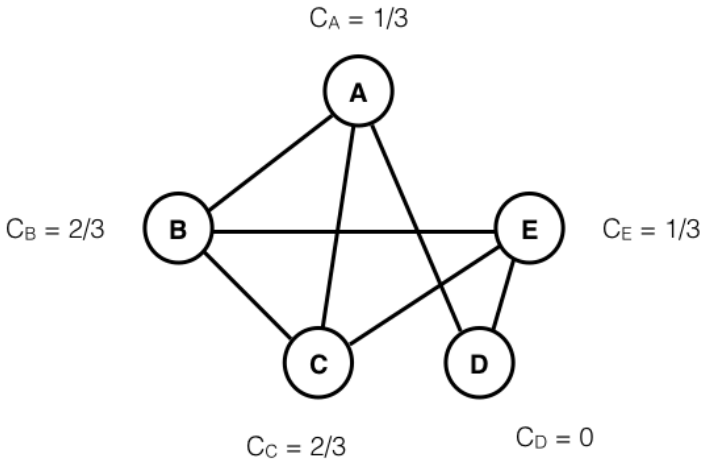
**1/8 * 5/8 * 6/8 * 1 * 1 = 30 / 512 = 15 / 256**

CBB 752 Spring 2016 Midterm Exam 03/07/2016

Name: _____

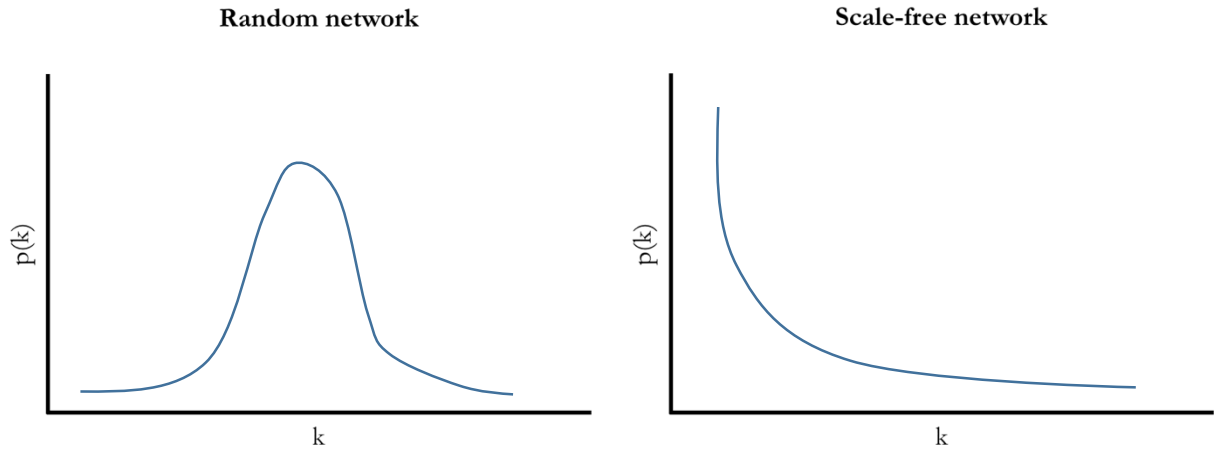
15. Calculate the average clustering coefficient of the following network. [5 points]




$C_{avg}$ = 1/3+2/3+2/3+0+1/3 = 2/5 or 0.4

16. Given k is degree and p(k) is frequency, complete the degree distribution curves for random network and scale-free network. [5 points]



CBB 752 Spring 2016 Midterm Exam 03/07/2016

Name: _____

17. Use the appropriate dynamic programming algorithm to LOCALLY align the sequences below. Use +2 for a match, -2 for a mismatch, and -1 for a gap. Complete the matrix shown below and show ALL optimal alignments. [20 points]

| | | C | T | G | T | T |
|---|---|---|---|---|---|---|
| | | | | | | |
| G | | | | | | |
| C | | | | | | |
| T | | | | | | |
| G | | | | | | |
| C | | | | | | |

| | | C | T | G | T | T |
|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 2 | 1 | 0 |
| C | 0 | 2 | 1 | 1 | 0 | 0 |
| T | 0 | 1 | 4 | 3 | 3 | 2 |
| G | 0 | 0 | 3 | 6 | 5 | 4 |
| C | 0 | 2 | 2 | 5 | 4 | 3 |

```
 CTGTT
 |||
GCTGC
```