CBB752b21 Quiz 1

1. Database (10pts)
[1] Which of the following is NOT a requirement for a table to be in First Normal Form (5pts):

(a) Each column name must be unique
(b) Each column value must be a single value only
(c) There should not be the case that a non-primary key column is determined by another non-primary key
(d) The order of the rows is insignificant

[2] List at least two reasons why the following table is not in Third Normal Form (5pts):

ID	**Name**	**Height(ft)**	**Weight(lb)**	**BMI**
1	Adam	6.0	164	22.2
3	Bob	5.8	140	20.7
3	Bob	5.8	140	20.7
4	Charlie	6.1	164	21.6
Five	Michael	5.6	124	19.4

All values for a given column must be of the same data type;
No two rows in a table can be identical;
There should not be the case that a non-primary key column is determined by another non-primary key (transitive dependency);
One for 2pts; two for 5pts

2. 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 (10pts)

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;

3. What's the probability of observing a new sequence TGCTAGG based on the PPM from the following given sequences? (10pts)

DNA 1: ACCTACG
DNA 2: AGCTACG
DNA 3: AGCTACG
DNA 4: TCCTAGG
DNA 5: ACCTACG

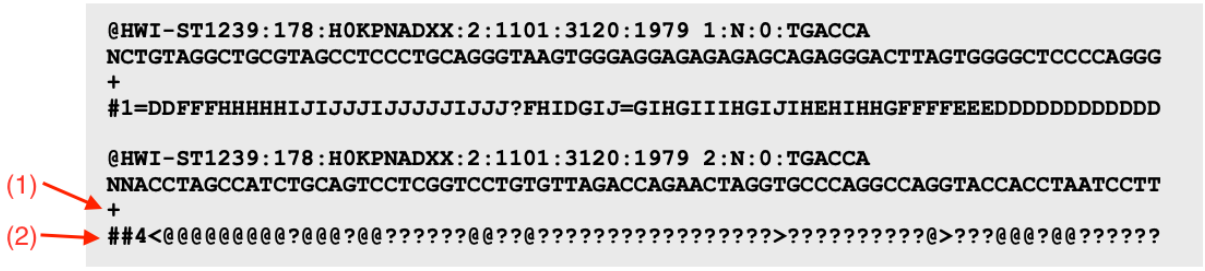
0.016 [10pts]
(The frequency of letter T, G and G in the 1^{st}, 2^{nd} and 6^{th} position is 0.2, 0.4, 0.2 respectively. Other positions are always the same, so have frequency of 1. Final result is 0.2*0.4*0.2)

4. What are the three major changes to make the global alignment algorithm into a local alignment? (10pts)

Penalty for miss-match: 3pts
Non-negative: 3pts
Trace back start from anywhere: 3pts
All correct: 1pt

5. Jim Grey considers data science as the fourth science paradigm, after the three earlier branches of science: empirical, theoretic and computational. The theoretical paradigm mainly involves: (10pts)
(a) description of natural phenomena
(b) generalization with models
(c) visualization of data
(d) simulation of complex phenomena

6. Below is a sample output from Illumina sequencing. Please briefly explain the meaning of the highlighted lines:



(1) The '+' sign (5pts): quality score identifier
(2) The symbols after the '+' sign (5pts): quality score

7. Proteomics (10pts)
[1] What is the m/z ratio in mass spectrometry? (5pts)
Mass/charge ratio
[2] Compared to sequencing of DNA, why is proteomic analysis more dependent on sample abundance? (5pts)
Protein samples cannot be amplified.

8. SILAC refers to "stable _(A)_ labeling with amino acids in cell culture". One common _(A)_ [same word as the first blank] used in proteomic study is __(B)__. (10pts)
(A) Isotope/isotopic (5pts)
(B) e.g., 13C, carbon 13 (lysine not accepted, but 13C lysine can be accepted) (5pts)

9. Choose the sequencing methods and their applications (NOTICE: you may reuse the options) (10pts):
Localization of transcription factors: B
Chromatin accessibility: C
Differential expression analysis: A
Determination of alternative splicing: A
(a) RNA-seq
(b) ChIP-seq
(c) DNase-Seq

10. Alignment (10pts)
Align the following two sequences using the Needleman-Wunsch global alignment algorithm. Upload a file showing [1] the complete dynamic programming matrix, [2] highlight the optimal traceback on the matrix and [3] write out the final alignment (e.g. AAA-TTCT and AAAGTT-T, where - represent gap).
(You could do it in microsoft office (excel/ppt tables/word tables) and highlight with cell background color, draw in photoshop or draw on a piece of paper and take a photo, etc.)

Sequence 1: ATACGG, Sequence 2: AACGTG
Use the following scoring scheme in the score matrix:
Match: +2
Mismatch: 0
Gap: -1

Final alignment 2pts
ATACG-G
A-ACGTG

There are 5 main matches along the trace back path (highlighted in yellow below)
<=3 correct: 1pt each
All 5 correct: 4pts
All other numbers correct 4pts
1 sporadic mistake -0.5pt
1 mistake that cause subsequent mistakes: -1pt for the whole set of errors

Matrix:

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

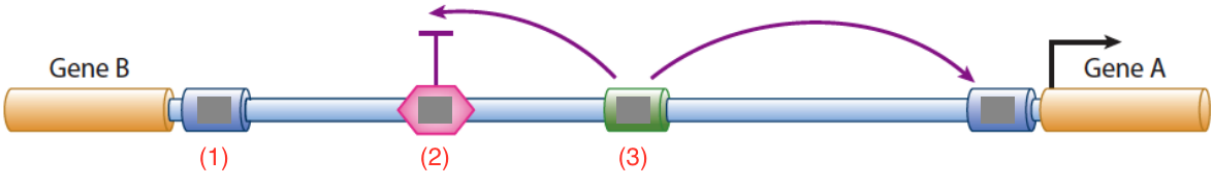
Or

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

Or

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

Bonus Question. (10pts)
[1] Name the regulatory elements in the diagram (6pts)



(1) Promoter
(2) Insulator
(3) Enhancer

[2] When processing of RNAseq reads, ____ RNA will dominate unless removed. (4pts)
Ribosomal/rRNA/ribosome