Indraprastha Institute of Information Technology Delhi (IIITD)

Department of Computational Biotechnology

BIO213 – Introduction to Quantitative Biology

END-SEM EXAM (May 03, 2024)

Time duration: 2 hours **Total marks**: 60

Question 1. Differentiate between **any 3** of the following: **(6 marks)**

(a) Endemic and Pandemic Lec 18, slide 5

(b) Character and distance-based methods of phylogenetic tree construction Lec 8, slide 4

(c) Transition and Transversion

Transition mutation is a type of point mutation in which one base (either purine or pyrimidine) is replaced by another base of the same class (purine or pyrimidine), whereas in transversion mutation a purine base is replaced by a pyrimidine base and pyrimidine base is replaced by a purine base.

(d) Computational complexity of dynamic programming and progressive alignment of multiple sequences Lec 6, slide 22 and 24

Question 2. Describe any three computational approaches used for prediction of protein-protein interactions. **(6 marks)**

Answers: Any three of the following methods are to be explained.

Gene cluster or gene neighborhood method

Rosetta stone method

Phylogenetic profile

Sequence-based co-evolution

Homology based inference

Association of structural motifs

Protein-protein docking

Machine learning-based methods

For explanation part refer to Lec19 (Biomolecular interactions). – 2 marks each

Question 3. State whether the following statements are correct or incorrect? In case of incorrect statement, justify your answer. **(7 marks)**

(a) Conformational search algorithm in *ab initio* protein structure modeling explores the potential energy surface and locate the local minimum. INCORRECT - Conformational search algorithm locates the global minimum. The native structure of the protein is believed to have the least potential energy, therefore a conformation representing the global minimum of the potential energy landscape. 1.5 marks

(b) Logs put positive and negative value of fold changes on a symmetric scale. CORRECT

1 mark

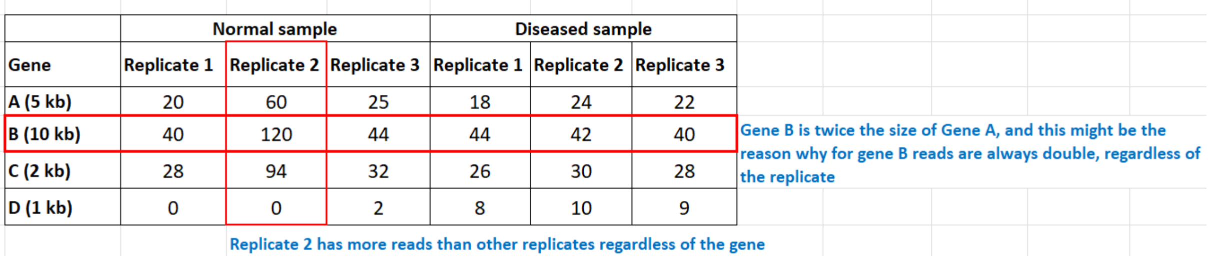
(c) The correct sequence of steps involved in RNA-seq analysis include library preparation, sequencing, quantification, read mapping, differential expression analysis. INCORRECT - Correct sequence is library preparation, sequencing, read mapping, quantification, differential expression analysis. 1.5 marks

(d) Technical replicates generally increase statistical power more than biological replicates. INCORRECT - Biological replicated contain both biological and technical variability, and therefore increase statistical power more than the technical replicates. 1.5 marks

(e) Total RNA extracted from the cells can directly be used for sequencing. INCORRECT - Ribosomal RNA removal is the major step before using RNA for sequencing. 1.5 marks

Question 4. Identify the problem associated with the following read count matrix. How can it be rectified? Elaborate. **(6 marks)**

	Normal sample			**Diseased sample**		
Gene	**Replicate 1**	**Replicate 2**	**Replicate 3**	**Replicate 1**	**Replicate 2**	**Replicate 3**
A (5 kb)	20	60	25	18	24	22
B (10 kb)	40	120	44	44	42	40
C (2 kb)	28	94	32	26	30	28
D (1 kb)	0	0	2	8	10	9



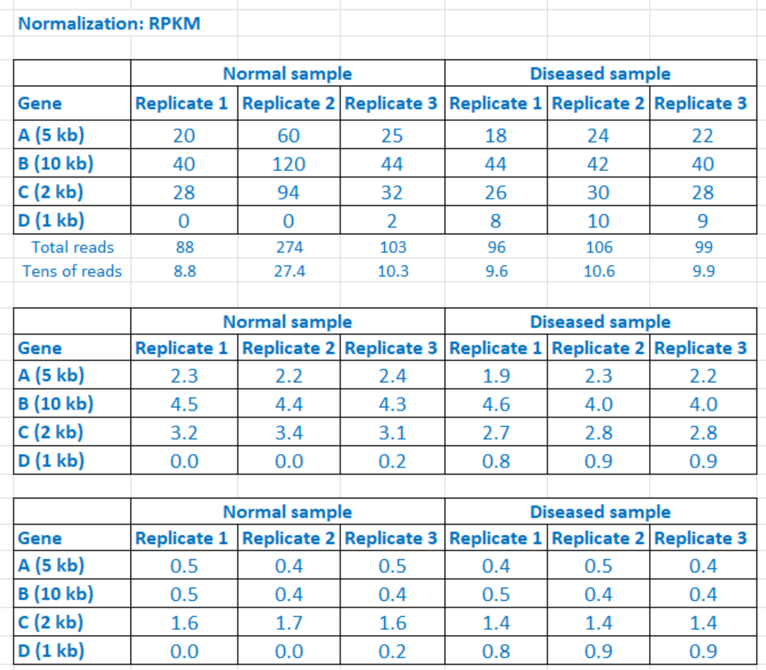
	Normal sample			Diseased sample		
Gene	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
A (5 kb)	20	60	25	18	24	22
B (10 kb)	40	120	44	44	42	40
C (2 kb)	28	94	32	26	30	28
D (1 kb)	0	0	2	8	10	9

Gene B is twice the size of Gene A, and this might be the reason why for gene B reads are always double, regardless of the replicate

Replicate 2 has more reads than other replicates regardless of the gene

Two correctly identified problems 2 marks

Any one of the following two methods is fine. Answers should not be in fractions, anything is fine – up to one decimal place or two. If all correct give 4 marks.

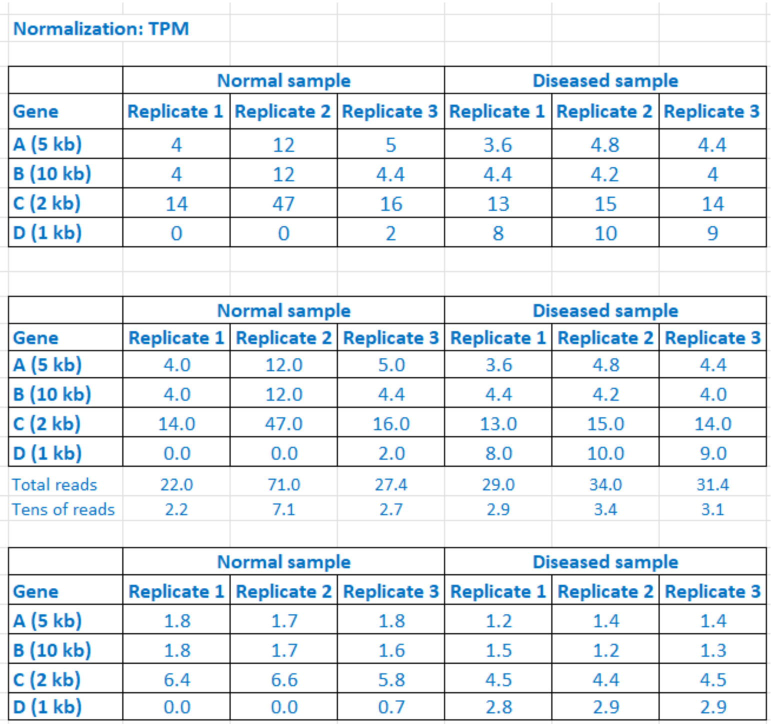


Normalization: RPKM

	Normal sample			Diseased sample		
Gene	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
A (5 kb)	20	60	25	18	24	22
B (10 kb)	40	120	44	44	42	40
C (2 kb)	28	94	32	26	30	28
D (1 kb)	0	0	2	8	10	9
Total reads	88	274	103	96	106	99
Tens of reads	8.8	27.4	10.3	9.6	10.6	9.9

	Normal sample			Diseased sample		
Gene	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
A (5 kb)	2.3	2.2	2.4	1.9	2.3	2.2
B (10 kb)	4.5	4.4	4.3	4.6	4.0	4.0
C (2 kb)	3.2	3.4	3.1	2.7	2.8	2.8
D (1 kb)	0.0	0.0	0.2	0.8	0.9	0.9

	Normal sample			Diseased sample		
Gene	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
A (5 kb)	0.5	0.4	0.5	0.4	0.5	0.4
B (10 kb)	0.5	0.4	0.4	0.5	0.4	0.4
C (2 kb)	1.6	1.7	1.6	1.4	1.4	1.4
D (1 kb)	0.0	0.0	0.2	0.8	0.9	0.9



Normalization: TPM

	Normal sample			Diseased sample		
Gene	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
A (5 kb)	4	12	5	3.6	4.8	4.4
B (10 kb)	4	12	4.4	4.4	4.2	4
C (2 kb)	14	47	16	13	15	14
D (1 kb)	0	0	2	8	10	9

	Normal sample			Diseased sample		
Gene	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
A (5 kb)	4.0	12.0	5.0	3.6	4.8	4.4
B (10 kb)	4.0	12.0	4.4	4.4	4.2	4.0
C (2 kb)	14.0	47.0	16.0	13.0	15.0	14.0
D (1 kb)	0.0	0.0	2.0	8.0	10.0	9.0
Total reads	22.0	71.0	27.4	29.0	34.0	31.4
Tens of reads	2.2	7.1	2.7	2.9	3.4	3.1

	Normal sample			Diseased sample		
Gene	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
A (5 kb)	1.8	1.7	1.8	1.2	1.4	1.4
B (10 kb)	1.8	1.7	1.6	1.5	1.2	1.3
C (2 kb)	6.4	6.6	5.8	4.5	4.4	4.5
D (1 kb)	0.0	0.0	0.7	2.8	2.9	2.9

Question 5. Briefly describe the function of **any 5** of the following: **(5 marks)**

(a) FASTQC – It is a widely used quality control tool. It assesses the quality of high-throughput raw sequencing data.

(b) STAR - Its primary function is to align RNA-seq reads (short sequences of RNA molecules obtained through high-throughput sequencing) to a reference genome.

(c) DEseq2 – It is used for differential gene expression analysis in RNA-sequencing (RNA-seq) data. Its primary function is to identify genes that are differentially expressed between different experimental conditions or sample groups.

(d) FASTA - FASTA can search large sequence databases using a query sequence to identify homologous sequences.

(e) I-TASSER – It is a widely used computational method for protein structure prediction and modeling by integrating threading/fold recognition and ab initio modelling.

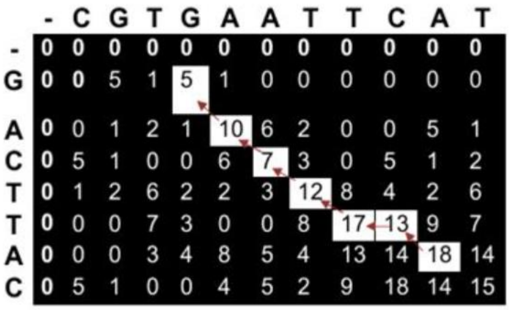
(f) Combined DNA Index System (CODIS) - Developed by the FBI, CODIS enables the sharing and comparison of DNA profiles. It uses primers to amplify 13 unique STR's in DNA using PCR. Odds that two individuals will have the same 13-loci DNA profile are more than one in a billion.

1 mark each. Look for the main point.

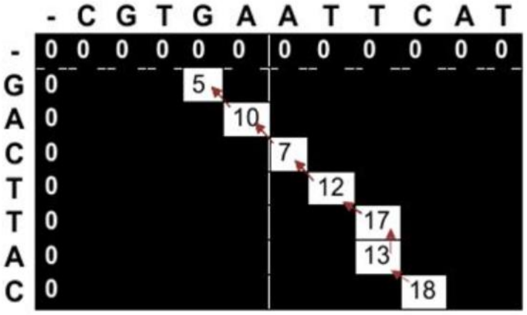
Question 6. What does this representation mean - <A-x-[ST](2)-x(0,1)-V-{ED}? **(2 mark)**

(a) C terminal- Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val-{any except Glu or Asp}
(b) N terminal-Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val-{any except Glu or Asp}
(c) C terminal- Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val-{Glu or Asp}
(d) N terminal-Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val-{Glu or Asp}

<u>Question 7.</u> Find the local regions of similarity between the following DNA sequences using dynamic programming and the given scoring scheme. **(5 marks)**
DNA sequences: (1) CGTGAATTCAT and (2) GACTTAC
Scoring function: Match = +5, Mismatch = -3, Gap = -4.



	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	5	1	0	0	0	0	0	0
A	0	0	1	2	1	10	6	2	0	0	5	1
C	0	5	1	0	0	6	7	3	0	5	1	2
T	0	1	2	6	2	2	3	12	8	4	2	6
T	0	0	0	7	3	0	0	8	17	13	9	7
A	0	0	0	3	4	8	5	4	13	14	18	14
C	0	5	1	0	0	4	5	2	9	18	14	15



	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0				5							
A	0					10						
C	0						7					
T	0							12				
T	0								17			
A	0								13			
C	0									18		

There are two possible alignments through these two pathways

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

+ + - + + - +    + + - + + - +
5 5 3 5 5 4 5    5 5 3 5 5 4 5
```

3 marks for correctly filled matrix. 1 mark for each of the correct alignment.

<u>Question 8.</u> What are the different steps involved in homology modelling of protein structures? Describe the **major** challenges associated with any two of these steps, and also discuss the possible solutions to those problems. **(6 marks)**

Answers: Steps involved in homology modelling of protein structures: (2 marks)
1- Finding the best template/homologous protein with known structure
2- Correct sequence alignment
3- Generating the backbone
4- Loop modeling
5- Side chain modeling
6- Model optimization and structure refinement
7- Validation of the developed model

<u>Some of the challenges</u>: (any two of these - 2 marks each)

a) Experimentally derived structure of homologous protein is essential - major limitation.

b) All the missing residues are assigned a loop structure, which is difficult to model - Loop modeling is done by knowledge-based method where PDB is searched for known loops, or by Energy based method where long chains are built by sampling Ramachandran conformations randomly.

c) Side chains are flexible and can adopt multiple conformations - Rotamer libraries are used.

d) Wrong backbone affects the side chain building process - Template that generates a backbone with least errors is chosen. Further, alignment that leads to smallest gap is used for backbone assignment.

Question 9. What is the smallest value of ε such that the layout below is valid under the Reconstruction model? **(3 marks)**

```
F= (ACCCT, CGTGC, TTAC, TGCCCA)      --ACCCT--
                                     ----CGTGC
                                     TTAC-----
                                     -TGCCCA--
                                    ___________
                                     TTACCCTGC
```

There exists one error between the second fragment and the consensus sequence.
So, d_s(CGTGC, TTACCCTGC) = 1

Now, we know that d_s(TGCCGT, TTACCGTGC) ≤ ε |CGTGC|

Therefore, 1 ≤ ε 5 => ε >= 1/5 (1 mark)

Then there exists two errors between the last fragment and the consensus sequence.
So, d_s(TGCCCA, TTACCCTGC) = 2

Now, we know that d_s(TGCCGT, TTACCGTGC) ≤ ε |CGTGC|

Therefore, 2 ≤ ε 5 => ε >= 2/5 (1 mark)

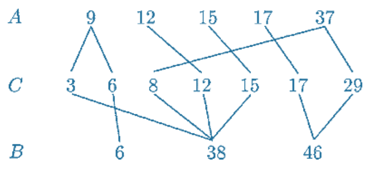
So, the smallest value for ε = 2/5 (1 mark) This final value has to be choosen

Question 10. Write a short note on **any 4** of the following: **(8 marks)**

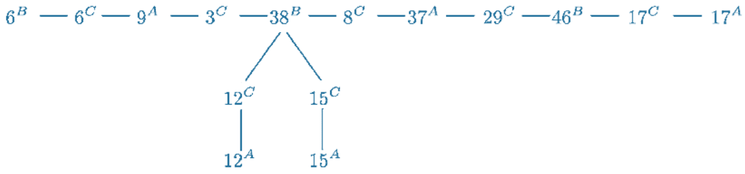
(a) Importance of relative solvent accessibility in characterizing interaction interface of proteins Lec19, slide 18
(b) *de novo* genome assembly Lec16-17, slide 36
(c) Importance of computational methods for disease forecasting Lec18, slide 2-3
(d) Threading for protein structure prediction Lec14-15, slide 54-56
(e) Applications of DNA fingerprinting Some imps from Lec-11-I
(f) Protein folding funnel Lec14-15, slide 61

Question 11. Create the physical map of DNA using the following information of fragment sizes obtained from an enhanced double digestion experiment using restriction enzymes A and B.

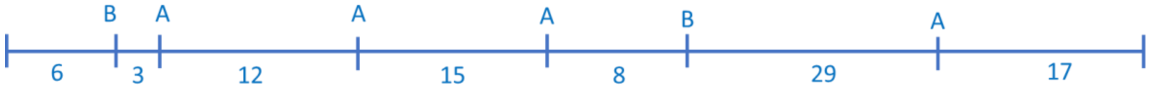
A = 9, 12, 15, 17, 37
B = 6, 38, 46
AB_1 = 3, 6, 8, 12, 15, 17, 29 **(6 marks)**



(a)


6^B — 6^C — 9^A — 3^C — 38^B — 8^C — 37^A — 29^C — 46^B — 17^C — 17^A
12^C
15^C
12^A
15^A

Fragment sequence: 6, 3, 12, 15, 8, 29, 17 (5 marks up to this part, spanning tree is a must)



(1 mark for this site map)