Indraprastha Institute of Information Technology Delhi (IIITD)

Department of Computational Biotechnology

BIO213 – Introduction to Quantitative Biology

END-SEM EXAM (May 09, 2023)

Name: ____ **Roll number:** ____________________

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

Question 1. Differentiate between **any 4** of the following: **(8 marks)**

(a) VNTR and STR Lec 10, slide 5-6
(b) Prognostic and Diagnostic biomarker Topic 10, slide 3
(c) E-value and p-value General explanation and significance
(d) UGMA and Neighbor-joining method for phylogenetic analysis Lec 5-II, slide 20
(e) Endemic and Epidemic Lec 16, slide 5
2 marks each

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

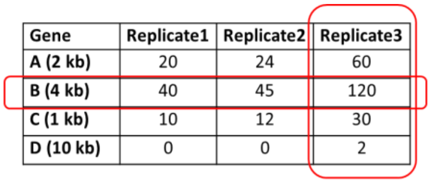
Answers: Any three of the following methods are to be explained.
Gene cluster or gene neighborhood method
Rosetta stone method
Phylogenetic profile
Sequence-based co-evolution
Homology based inference
Association of structural motifs
Protein-protein docking
Machine learning-based methods

For explanation part refer to Lec17 (Biomolecular interactions). – 2 marks each

<u>Question 3.</u> Identify the problem associated with the following read count matrix. How can it be rectified? **(6 marks)**

Gene	Replicate1	Replicate2	Replicate3
A (2 kb)	20	24	60
B (4 kb)	40	45	120
C (1 kb)	10	12	30
D (10 kb)	0	0	2

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



Gene	Replicate1	Replicate2	Replicate3
A (2 kb)	20	24	60
B (4 kb)	40	45	120
C (1 kb)	10	12	30
D (10 kb)	0	0	2

Replicate 3 has more reads than other replicates regardless of the gene

Normalization is required:

Method: RPKM

Gene	Replicate1	Replicate2	Replicate3
A (2 kb)	20	24	60
B (4 kb)	40	45	120
C (1 kb)	10	12	30
D (10 kb)	0	0	2
Total reads	70	81	212
Ten of reads	7	8.1	21.2

Gene	Replicate1 RPM	Replicate2 RPM	Replicate3 RPM
A (2 kb)	2.86	2.96	2.83
B (4 kb)	5.71	5.56	5.66
C (1 kb)	1.43	1.48	1.42
D (10 kb)	0.00	0.00	0.09

Gene	Replicate1 RPKM	Replicate2 RPKM	Replicate3 RPKM
A (2 kb)	1.43	1.48	1.42
B (4 kb)	1.43	1.39	1.42
C (1 kb)	1.43	1.48	1.42
D (10 kb)	0.00	0.00	0.01

OR

Method: TPM

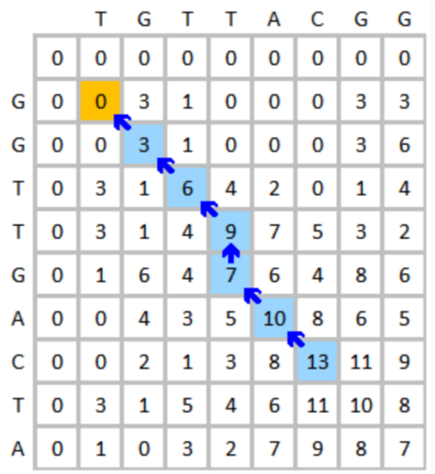
Gene	Replicate1 RPK	Replicate2 RPK	Replicate3 RPK
A (2 kb)	10	12	30
B (4 kb)	10	11.25	30
C (1 kb)	10	12	30
D (10 kb)	0	0	0.2

Gene	Replicate1 RPK	Replicate2 RPK	Replicate3 RPK
A (2 kb)	10	12	30
B (4 kb)	10	11.25	30
C (1 kb)	10	12	30
D (10 kb)	0	0	0.2
Total RPK	30	35.25	90.2
Tens of RPK	3	3.525	9.02

Gene	Replicate1 TPM	Replicate2 TPM	Replicate3 TPM
A (2 kb)	3.33	3.40	3.33
B (4 kb)	3.33	3.19	3.33
C (1 kb)	3.33	3.40	3.33
D (10 kb)	0.00	0.00	0.02

Give full marks if any of the two methods is given.
2 marks for describing the problem and 4 marks for the normalized values
Value in fractions or up to one decimal place are also acceptable

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



		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

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

Question 5. 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

Some of the challenges: (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 6. Which of the following statements are incorrect? Justify your answer. **(6 marks)**

(a) 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. (2 marks)

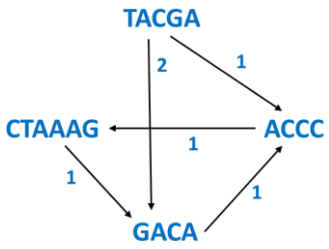
(b) Technical replicates generally increase statistical power more than biological replicates.
INCORRECT - Biological replicated contain both biological and technical variability, and therefore increase statistical power more than the technical replicates. (2 marks)

(c) The first search against the sequence database in PSI-BLAST uses PAM 250 substitution matrix. INCORRECT - It uses BLOSUM62. (2 marks)

Question 7. List the major factors that contribute to the function for potential energy calculations. **(2 marks)**
Topic 9/Lec 13, Slide 33 (formula not required)

Question 8. Why is it important to study RNA even though all the instructions that a cell follows are encoded in its genomic DNA? **(4 marks)**
Topic 10, Slide 8

Question 9. Construct an overlap graph for F= (TACGA, ACCC, GACA, CTAAAG). Find a shortest common superstring for this collection. **(4 marks)**



```
TACGA
   GACA
      ACCC
         CTAAAG
_______________
TACGACACCCTAAAG
```

Give 2 marks for the graph and 2 marks for the fragment assembly. The alignment layout given should result in the shortest string.

OR

What is the smallest value of ε such that the layout below is valid under the Reconstruction model?

```
F= (ACCGT, CGTGC, TTAC, TGCCGT)       --ACCGT--
                                      ----CGTGC
                                      TTAC-----
                                      -TGCCGT--
                                     ___________
                                      TTACCGTGC
```

There exists one error between the last fragment and the consensus sequence.
So, $d_s(\text{TGCCGT, TTACCGTGC}) = 1$

Now, we know that $d_s(\text{TGCCGT, TTACCGTGC}) \leq \varepsilon \ |\text{TGCCGT}|$

Therefore, $1 \leq \varepsilon \ 6$.

So, the smallest value for $\varepsilon = 1/6$

Question 10. Match the following: **(3 marks)**

(a) Protein data bank	i. Protein clustering on sequence similarity
(b) GenBank	ii. Dot plot
(c) CODIS	iii. Mutations and crossover
(d) BLOCKS	iv. Biomolecular structural database
(e) Genetic algorithm	v. Short tandem repeats
(f) FASTA	vi. Nucleotide database

(a)-iv, (b)-vi, (c)-v, (d)-i, (e)-iii, (f)-ii
0.5 mark for each of the correct answer

Question 11. Write a short note on **any 5** of the following: **(10 marks)**

(a) Importance of relative solvent accessibility in characterizing interaction interface of proteins Lec17, slide 18
(b) *de novo* genome assembly Topic 10, slide 31
(c) Use of biomarkers for cancer General importance along with an example
(d) Threading for protein structure prediction Lec13, slide 23-35
(e) Correction measure for generating position specific scoring matrices to account for lack/bias in data Lec 11, slide 8
(f) Drug repurposing with a suitable example Lec 10, slide 8-9 (other appropriate examples are also acceptable)

2 marks each