Indraprastha Institute of Information Technology Delhi (IIITD)

Department of Computational Biotechnology

BIO213 – Introduction to Quantitative Biology

END-SEM EXAM (May 06, 2022)

Time duration: 1.5 hours **Total marks:** 60

Question 1. 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 2. Construct a simple dot plot using the following two nucleotide sequences. Does your plot reveal any regions of similarity? Comment on the same. **(5 marks)**
Sequence 1: **GCTAGTCAGATCTGACGCTA**
Sequence 2: **GATGGTCACATCTGCCGC**

Answers:

	G	C	T	A	G	T	C	A	G	A	T	C	T	G	A	C	G	C	T	A
G	O				O				O					O			O			
A				O				O		O					O					O
T			O			O					O		O						O	
G	O				O				O					O			O			
G	O				**O**				O					O			O			
T			O			**O**					O		O						O	
C		O					**O**					O				O		O		
A				O				**O**		O					O					O
C		O					O					O				O		O		
A				O				O		**O**					O					O
T			O			O					**O**		O						O	
C		O					O					**O**				O		O		
T			O			O					O		**O**						O	
G	O				O				O					**O**			O			
C		O					O					O				O		O		
C		O					O					O				**O**		O		
G	O				O				O					O			**O**			
C		O					O					O				O		**O**		

```
GTCAGATCTGACGC
||||  |||||  |||
GTCACATCTGCCGC
```

This is the region of similarity with two mismatches

Question 3. 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. For explanation part refer to Lec18 (Biomolecular interactions). – 2 marks each

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

Question 4. (i) What sites in the following alignment would be informative for a parsimony analysis? How many sites are invariant? **(4 marks)**

	1----------10	11---------20	21---------30	31---------40
1	GAATGCTGAT	ATTCCATAAG	TCACGAGTCA	AAAGTACTCG
2	GGATGGTGAT	ACTTCGTAAG	TCCCGAGTCG	AAAGTACTCG
3	GGATGATGAT	ACTTCATAAG	TCTCAAATCA	AAGGTACTTG
4	GGATGCTGAC	ACTTCATAAG	TCGCGAGTCA	AAAGTACTTG
5	GGATGCTGAC	ACTCCGTAAG	TCCCGAGTCA	AATGTACTCG

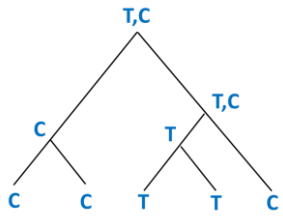
Informative sites: 10, 14, 16, 39 (2 marks)

Invariant sites: 28 sites − 1, 3, 4, 5, 7, 8, 9, 11, 13, 15, 17, 18, 19, 20, 21, 22, 24, 26, 28, 29, 31, 32, 34, 35, 36, 37, 38, 40 (2 marks)

(ii) How many rooted trees can be constructed to describe the possible relationships between 5 taxa? **(1 mark)**

105

(iii) Draw any one possible rooted tree for five taxa whose nucleotides at the position under consideration are C, C, T, T, C. Label each of the internal node with the most likely candidate for the inferred ancestral sequence. What are the minimum number of substitutions required by your tree topology? **(5 marks)**



Minimum number of substitutions will be 1 (with C at the root node).

This is one of the numerous possibilities of making the tree

Question 5. Match the following: **(5 marks)**

(A) Quality check of raw reads	(a) STAR
(B) Read mapping	(b) Illumina
(C) Counting of aligned reads	(c) FastQC
(D) Differential expression analysis	(d) featureCounts
(E) Sequencing platform	(e) edgeR

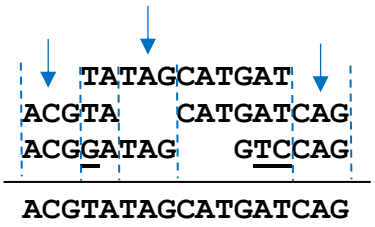
Answer: A-c, B-a, C-d, D-e, E-b (1 mark each)

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

(a) Coverage and Linkage
(b) Prognostic and Diagnostic biomarker
(c) E-value and p-value
(d) Motif and domain
(e) Linear and Affine gap penalty

Different answers can be there. Just make sure that the main points of differences are present. (2 marks each)

Question 7. The following layout shows the assembly of five fragments into a contig. Is this assembly valid according to the MULTICONTIG model with ε= 0.4 and t = 3? Justify your answer.



(6 marks)

t=3, therefore minimum linkage should be 3 (shown above by arrows). (2 marks)

ε=0.4, therefore 4 insertions/deletions/substitutions are allowed for a fragment length of 10.
ds (ACGGATAG, ACGTATAGCATGATCAG) = 1, ε| f | = 0.4 X 8 = 3.2
this ds <= ε| f |, so this output is OK for ε=0.4 (2 marks)
Similarly, ds (GTCCAG, ACGTATAGCATGATCAG) = 2, ε| f | = 0.4 X 6 = 2.4
Here also ds <= ε| f |, so this output is OK for ε=0.4 (2 marks)

OR

Question 7. Find the local regions of similarity between the following DNA sequences using dynamic programming and the given scoring scheme.
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

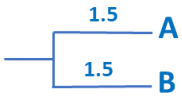
G T T - A C
| | | | | | |
G T T G A C

Question 8. Use weighted UPGMA to reconstruct a phylogenetic tree using the following distance matrix. **(6 marks)**

Species	A	B	C	D
B	3	-	-	-
C	6	5	-	-
D	9	9	10	-
E	12	11	13	9

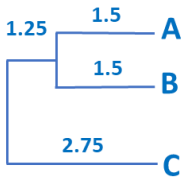
I.

	A	B	C	D
B	3	-	-	-
C	6	5	-	-
D	9	9	10	-
E	12	11	13	9



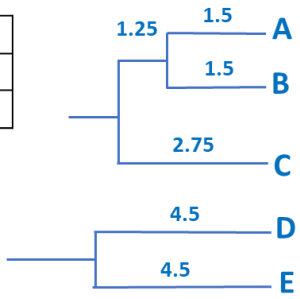
II. d(AB,C)=[d(AC)+d(BC)]/2 = [6+5]/2 = 5.5
d(AB,D)=[d(AD)+d(BD)]/2 = [9+9]/2 = 9
d(AB,E)=[d(AE)+d(BE)]/2 = [12+11]/2 = 11.5

	AB	C	D
C	5.5	-	-
D	9	10	-
E	11.5	13	9



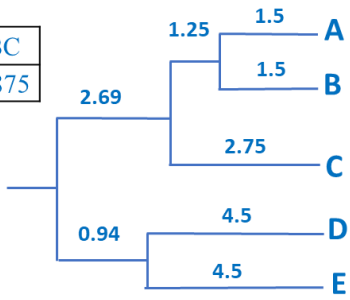
III. d(ABC,D)=[d(AB,D)+d(C,D)]/2 = [9+10]/2 = 9.5
d(ABC,E)=[d(AB,E)+d(C,E)]/2 = [11.5+13]/2 = 12.25

	ABC	D
D	9.5	-
E	12.25	9



IV. d(ABC,DE)=[d(ABC,D)+d(ABC,E)]/2 = [9.5+12.25]/2 = 10.875

	ABC
DE	10.875



Question 9. 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

Question 10. Which of the following statements are incorrect? Justify your answer. **(4 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 (0.5 marks)

(c) 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)

(d) The first search against the sequence database in PSI-BLAST uses PAM 250 substitution matrix. – INCORRECT - It uses BLOSUM62. (1.5 marks)