BBL434 Major

2019-2020 Semester 2

September 2nd 2020

Maximum marks:25

Time: 120 minutes

Please submit A PDF OF HAND-WRITTEN ASNWERS in the google form below:
https://forms.gle/cS4ZD9PXz3CuoQzy5

1. You have deteremined the genome sequence of a bacterium. How can you use BLAST to identify protein coding genes in the bacterium if you only have access to protein databases ***(1)***

2. If you BLAST the same protein in July and August, you got an E-value of the top hit as 0.1 and 0.2 respectively, while the scores remained the same. Can you explain why ? ***(1)***

3. Detail the 8 steps followed to perform computer aided drug discovery ? ***(1)***
What value is used to determine if the docking of a potential chemical inhibitor in a target protein is strong and valid after performing Molecular Dynamic simulations ? ***(1)***

4. While downloading the chromsomes 1, 13, 22 from the human genome, I made a mistake of naming them A, B, C and forgot. Can you use your knowledge about the human genome to map the following files to the correct chromosome. Explain ***(1)***

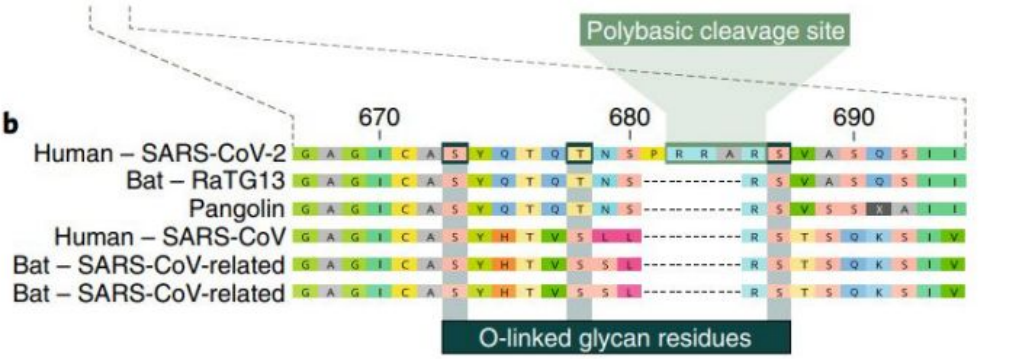
```
[ishaan] ➢ ls -ltrh *.fa.gz
-rwx------  1 ishaa  UsersGrp  30.1M Sep  1 21:45 A.fa.gz
-rwx------  1 ishaa  UsersGrp  10.8M Sep  1 21:45 B.fa.gz
-rwx------  1 ishaa  UsersGrp  70.4M Sep  1 21:49 C.fa.gz
```

5. N50 of DNA from old human fossil bones is much less than the N50 of DNA obtained from fresh human bone samples. Given the above observation, what can you say about the physical properties of DNA obtained from a Neaderthal vs a Modern human ? Which one of the assemblies will have more contigs ? ***(2)***

7. The corona virus is about 4000 bp long, recently CSIR-IGIB New Delhi released a genome assembly with average 10x coverage using single end 100 base pairs DNA sequencing. What is the minimal number of reads used for this genome assembly ? ***(2)***

8. Detail the difference between paired end sequencing and single end sequencing ?
If you are a startup company that provides personal genome sequencing (seqeuncing the genome of a person) who has to decide between the following options from a sequencing provider. Which provider will you use and why ? ***(2)***
1. Gives PE 100 bp for 1000 INR per Gigabases (10^9 nucleotides) or Gb
2. Gives SE 150 bp for 1000 INR per Gb
3. Gives PE 50 for 750 INR per Gb

9. Make a maximum Parisony Tree from the following alignment of Coronavirus strains from sequences 670 nucleic acids to the 675+x nucleotides, where x = Number of vowels (AEIOU) in your first and last name combined. ***(5)***



OR

The following is matrix of distances between a gene found in 7 organisms based on a multiple sequencing alignment for the gene's nucleic acid sequence.

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	2.00

a. There is an issue with this matrix, point it out and correct it. Explain ***(1)***
b. What are the assumptions of the UPGMA method ? ***(1)***
c. If the last digit of your ID is odd then use the above matrix to make a phylogenetic tree for A, B, C, D, E, F by the UPGMA method . If the last last digit of your ID is even then make a phylogenetic tree using A, B, C, D, E, G by the UPGMA method. ***(3)***
Example: If your ID is bb1170013 you will make a phylogenetic tree of A, B, C, D, E, F

10. Which of the statements are True/False about results from the Human genome Project paper published in 2001. ***(1 mark each)***
Explain in detail with calculation/exact statements from the Human genome project paper. No marks for no explanation. Link to paper:
https://drive.google.com/file/d/1jjTfbbrUhzP4z8Sj2z6ljq-rOA3qniSv/view?usp=sharing
a. The average ratio between the length of introns and exons in a gene is 0.03
b. N90 of the Human genome assembly is less than 100,000 bp
c. The average length of contigs was greater than the average length of scaffolds in the human genome assembly.

11. What part of the genome does the 1% of the studied under the ENCODE project represent? Briefly explain the 2 findings of the project in your own words ***(2)***
Link to paper:
https://drive.google.com/file/d/1hoVn6jjMTEeCcclU4ld8GVU1yscQNQz7/view?usp=sharing

12. Construct a de Bruijn graph for the genome "TAATGCCATGGGATGTT" with k=3. Identify all possible assemblies from the graph. ***(3)***

If the resulting assemby is not unique, then what can you do to obtain a unique assembly ? Represent the new assembly as a graph ***(1)***