University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

Exam#2

November 1st, 2023 Name: _____

25 Questions, 4 points each

Question 1. Suppose you have two sequences S_1 and S_2. True or False, the optimal global alignment score is always higher than the optimal local alignment score.

Question 2. True or False. Suppose you have two sequences S_1 and S_2 and a local alignment that involves the same number of bases of S_1 as it does S_2. Then the alignment does not have any indels.

Question 3. True or False. An indel can align with another indel in a **multiple sequence alignment**.

Question 4. In the context of DNA, connect the things on the left with what could be an of it example on the right.

Substitution Matrix

$$\begin{bmatrix} 0 & 7 & 8 & 3 \\ 7 & 0 & 1 & 11 \\ 8 & 1 & 0 & 8 \\ 3 & 11 & 8 & 0 \end{bmatrix}$$

Markov Transition Matrix

$$\begin{bmatrix} 4 & 2 & -3 & -1 \\ -1 & 2 & 0 & -6 \\ -3 & -1 & 4 & -2 \\ -4 & -2 & 0 & 5 \end{bmatrix}$$

Position Weight Matrix

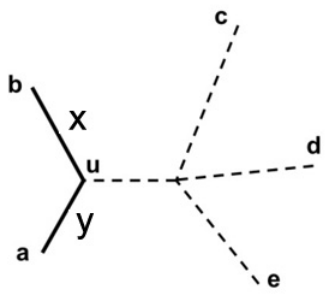
$$\begin{bmatrix} .2 & .5 & .2 & .1 \\ .1 & 0 & .7 & .2 \\ .5 & .2 & .2 & .1 \\ .1 & .1 & .8 & 0 \end{bmatrix}$$

Distance Matrix

$$\begin{bmatrix} .3 & .7 & .2 & .4 & .2 & 0 \\ .1 & 0 & .1 & .2 & .2 & .3 \\ .5 & .2 & .2 & .4 & .2 & .3 \\ .1 & .1 & .5 & 0 & .4 & .4 \end{bmatrix}$$

Question 5. Consider the following distance matrix and unrooted tree. Suppose $x = 3$. What is y?

	a	b	c	d	e
a	0				
b	5	0			
c	9	10	0		
d	9	10	8	0	
e	8	9	7	3	0



Question 6. True or False. In a SAM file, the strand is encoded in the bitflag.

Question 7. If we are clustering rows of the following block to construct a BLOSUM70 matrix, how many clusters will there be?

Answer: _____

Answer the same question for BLOSUM80:

Answer: _____

$$\begin{vmatrix} A & A & C & C \\ A & A & C & G \\ A & G & C & G \\ G & G & C & G \end{vmatrix}$$

Question 8. The BLAST null hypothesis is modeled by alignment to a database of random sequence generated according to the background frequencies. Explain where the "background frequencies" come from.

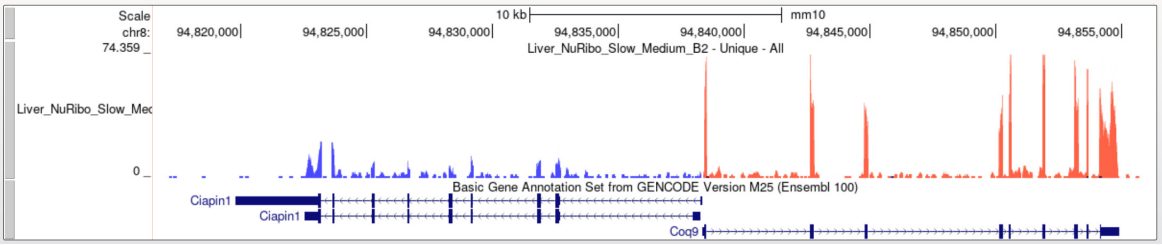
Question 9. True or False. The BLAST "two-hit" method is more efficient because the Smith-Waterman step is applied to shorter sequences.

Question 10. True or False, the *total running accumulated score* of the random walk must be positive at the position where a max-excursion happens. In other words the max excursion can take place entirely below the x-axis.

Question 11. Give one reason why RNA-Seq *aligment* is a more difficult problem than DNA-Seq.

Question 12. When doing RNA-Seq quantification, explain why isoform level quantification is more difficult than gene level.

Question 13. Three Questions: (a) What do the colors represent? (b) Which of the two isoforms of Ciapin1 appear to be the primary one that is expressed? (c) Which of the two genes shown has a higher intron signal to exon signal ratio?



Question 14. Draw lines from the things on the left to the corresponding things on the right. This is many-to-many association, meaning more than one line can emanate from the same entry.

	Isoform-Level Quantification
	BWA
ChIP-Seq	Phasing
RNA-Seq	STAR
ATAC-Seq	SNP-Calling
DNA-Seq	Peak Calling
	Differential Expression

Question 15. What are the two tasks for ChIP-Seq that we've discussed?
(A) Transcription Factor Binding
(B) Gene Differential Expression Analysis
(C) SNP Calling
(D) Histone Modification
(E) Open Chromatin

Question 16. A q-value cutoff of 0.18 gives 50 DE genes. How many do we expect to be *true*-positives?
(A) 0
(B) 41
(C) 18
(D) 36
(E) 9

Question 17. You are performing 20 tests. You want to use the Bonferroni correction to control the FWER at the level 0.01. What cutoff should you use?

Answer: _____

Question 18. True or False. You should never use a q-value cutoff greater than 0.05.

Question 19. FWER stands for
(A) False Without Error Rate
(B) Fragment-Wide Expected Rate
(C) Fragment Witheld Exogenous Rate
(D) Formally Well-Established Rate
(E) Family-Wise Error Rate

Question 20. Consider three consecutive SNPs. Suppose a population has two haplotypes at these SNPs: AGC and TCT. True or False. If an individual is homozygous at any one of the three SNPs, then they are heterozygous at the other two.

Question 21. What's the long term behavior of the Markov Chain given by this matrix?

$$\begin{array}{c} \\ \mathsf{A} \\ \mathsf{B} \\ \mathsf{C} \end{array} \begin{array}{c} \begin{array}{ccc} \mathsf{A} & \mathsf{B} & \mathsf{C} \end{array} \\ \begin{bmatrix} .2 & .5 & .3 \\ 0 & .5 & .5 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

Answer: _____

Question 22. We know the BLAST algorithm is heuristic, it cannot guarantee it returns the local alignment with the highest score. True or False, if we could implement an efficient enough algorithm so that it did return the local alignment with the highest score, then we would not need the BLAST E-value.

Question 23. True or False. The multiple-testing problem gets worse (more severe) in an RNA-Seq differential expression analysis as the number of DE genes decreases.

Question 24. Suppose an individual is heterozygous at a SNP location and suppose we have DNA-Seq from that individual, but we only got two reads that cover that SNP location. Assuming no errors in the reads and equal probability of a read coming from either parental chromosome, what is the probability that the reads identify both of the SNP variants? (*Hint: It might help to model the problem on flipping a coin*)

Question 25. Consider the following CIGAR string: 3S99M2042N47M1S. What is the read length?

Answer: _____