University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

Exam#2 (Make-Up)

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 can equal the optimal local alignment score.

Question 2. Suppose you have two sequences S_1 and S_2 both of length N. How many global alignments are there with one indel?

Answer: _____

Question 3. True or False. ClustalW can find the optimal global alignment between N sequences.

Question 4. Consider the Position Weight Matrix. Circle the corresponding Frequency Logo.

$$\begin{bmatrix} A: & 0.1 & 0.8 & 0.1 & 0.2 & 0.9 \\ G: & 0.8 & 0.0 & 0.1 & 0.2 & 0.0 \\ C: & 0.1 & 0.1 & 0.4 & 0.2 & 0.0 \\ T: & 0.0 & 0.1 & 0.4 & 0.4 & 0.1 \end{bmatrix}$$

Question 5. Here is a table with information for building a BLOSUM substitution matrix. The last column has been left blank. For the first two entries "A to A" and "A to B" what will be the signs of their scores? In other words for each one, is it positive, negative or zero?

aligned pair	proportion observed	proportion expected	$2\log_2\left(\frac{\text{proportion observed}}{\text{proportion expected}}\right)$
A to A	26/60	196/576	
A to B	8/60	112/576	
A to C	10/60	168/576	
B to B	3/60	16/576	
B to C	6/60	48/576	
C to C	7/60	36/576	

Question 6. True or False. If a distance matrix M is not derived from a tree, then the neighbor joining method should not be applied.

Question 7. Cluster rows for a BLOSUM70 and then count the number of times A is paired with G.

Answer: _____

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

Question 8. How do you intepret a BLAST E-value of 3?

Question 9. Given a protein database and background probabilities $p_1, p_2, \ldots, p_{20}$, what is the null hypothesis probability that amino acid i is aligned with itself in any given position of an alignment.

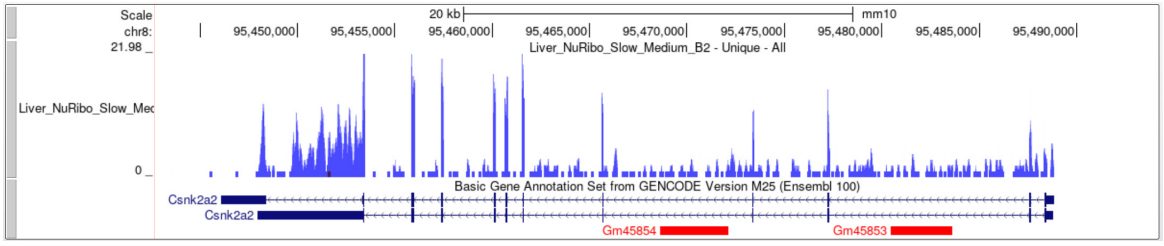
Question 10. BLAST has a threshold on the score of an ungapped (seed) alignment to trigger a gapped alignment. By what algorithm is that gapped alignment done?

Answer: _____

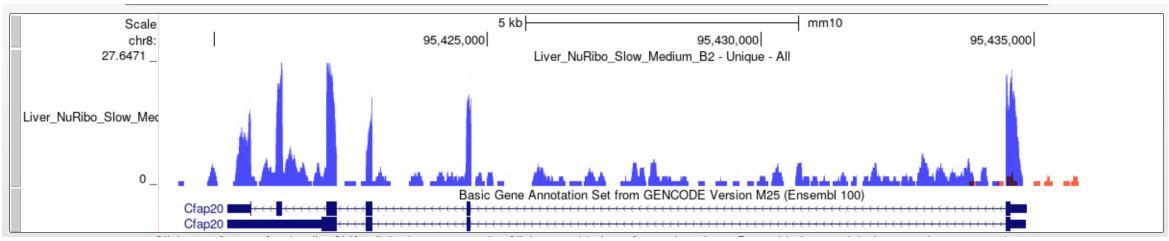
Question 11. Suppose you generate one DNA-Seq and one RNA-Seq assay. Assume both have the same number of reads. Which should have the greater *maximum* depth of coverage across the genome.

Answer: _____

Question 12. The gene shown Csnk2as has two single-exon genes living in its introns, Gm45854 and Gm45853. The first one Gm45854 is on the forward (plus) DNA strand and the other is one Gm45853 the reverse (minus) DNA strand. Suppose we have strand-specific RNA-Seq as shown in the coverage plot. Why can we more confidently conclude Gm45854 is not expressed that we can conclude that Gm45853 is not expressed?



Question 13. The gene shown Cfap20 has two isoforms. The figure shows an RNA-Seq coverage plot for this gene. It's possible that both isoforms are expressed. But which of the two isoforms can you be confident is definitely expressed?



Question 14. Draw lines from the things on the left to the corresponding things on the right. Things on the right connect to one thing on the left. But there is one thing on the left that connects to two things on the right.

ChIP-Seq	DNA methylation assay
RNA-Seq	Gene Expression
ATAC-Seq	H3K27me3
DNA-Seq	Anti-sense transcription
	Open Chromatin

Question 15. True or False. Transcription Factor Binding Assays tend to have narrow peaks and histone modification assays tend to have broader peaks.

Question 16. A q-value cutoff of 0.75 gives the entire list of all genes. How many genes do we expect to be DE?
(A) None of them
(B) 25% of them
(C) 50% of them
(D) 75% of them
(E) All of them

Question 17. Which is generally more conservative, the FWER or the FDR?

Answer: _____

Question 18. Suppose for a sequence of consecutive SNP's there is only one haplotype in the population. Suppose this population is an isolate, so no outbreeding. True or False, assuming no *de novo* mutations, a generation later the population must still have only one haplotype.

Question 19. True or False. FDR is a probability and FWER is an expected value.

Question 20. A "progressive" alignment means:
(A) A multiple sequence alignment that adds one base at a time.
(B) A multiple sequence alignment that adds one sequence at a time.
(C) A multiple sequence alignment that finds the best local alignment and extends progressively.
(D) A multiple sequence alignment that only aligns the related sequences.
(E) A multiple sequence alignment that progressively becomes more accurate.

Question 21. Explain how a DNA sequence is used by BLAST to search a *protein* database. We're not looking for an algorithm, just how protein is searched with a DNA input.

Question 22. What's the long term behavior of the Markov Chain given by this matrix?

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

Answer: _____

Question 23. Draw lines from the things on the left to their corresponding thing on the right

Compare different genes in same sample	Normalize for gene length
Compare same gene different samples	Normalize for depth of sequencing

Question 24. Suppose the genotype at a SNP is A/T in an individual. Suppose we perform DNA-Seq and a read covering the SNP has a 50% change of being from either parental chromosome. Suppose we get five reads that cover the SNP. What is the probability that all five reads have the same variant (so all A or all T)?

Answer: _____

Question 25. Consider the following CIGAR string: 20M3I37M2042N20M3D20M. What is the read length?

Answer: _____