# University of Pennsylvania

## BIOL4536 Fall 2023

**Professor: Gregory R. Grant**

# Exam#2 (Make-Up) (SOLUTIONS)

**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.

**ANSWER:** True.

**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:** Zero, if the sequences have the same length, then if you have an indel in one sequence somewhere, you must have one in the other sequence too, somewhere else to have a global (end-to-end) alignment.

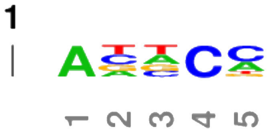
**Question 3.** True or False. ClustalW can find the optimal global alignment between $N$ sequences.

**ANSWER:** True, it is a heuristic algorithm, but it could luck into the optimal.

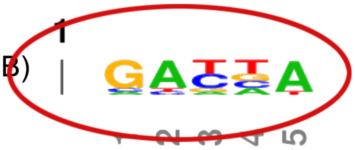
**Question 4.** Consider the Position Weight Matrix. Circle the corresponding Frequency Logo.

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

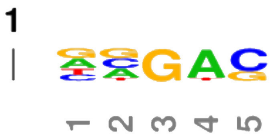
(A)



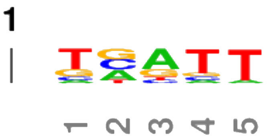
(B)



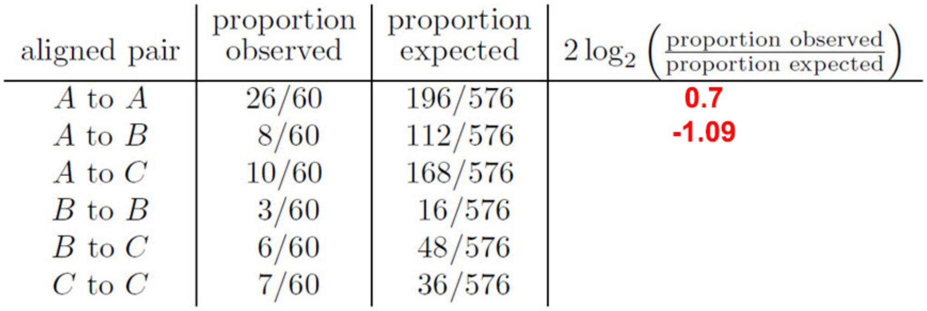
(C)



(D)



**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 $A$ | $26/60$ | $196/576$ | 0.7 |
| $A$ to $B$ | $8/60$ | $112/576$ | -1.09 |
| $A$ to $C$ | $10/60$ | $168/576$ | |
| $B$ to $B$ | $3/60$ | $16/576$ | |
| $B$ to $C$ | $6/60$ | $48/576$ | |
| $C$ to $C$ | $7/60$ | $36/576$ | |

**ANSWER:** The first one is positive and the second one is negative.

**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.

**ANSWER:** False, it will give a good approximate tree.

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

$$\left|\begin{array}{cccc} A & A & C & C \\ A & A & C & G \\ A & G & C & G \\ G & G & C & G \\ A & A & A & A \end{array}\right|$$

**ANSWER:** 1/4 + 2/4 + 3/4 = 3/2.

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

**ANSWER:** It means you should expect three alignments with that score or higher just by chance against a random database.

**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.

**ANSWER:** $p_i p_i$.

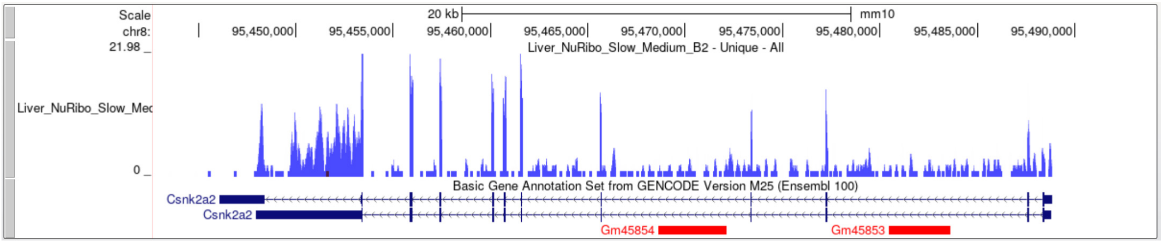
**Question 10.** BLAST has a threshold on the score of an ungapped (seed) alignment to trigger an gapped alignment. By what algorithm is that gapped alignment done?

**ANSWER:** Smith-Waterman

**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.

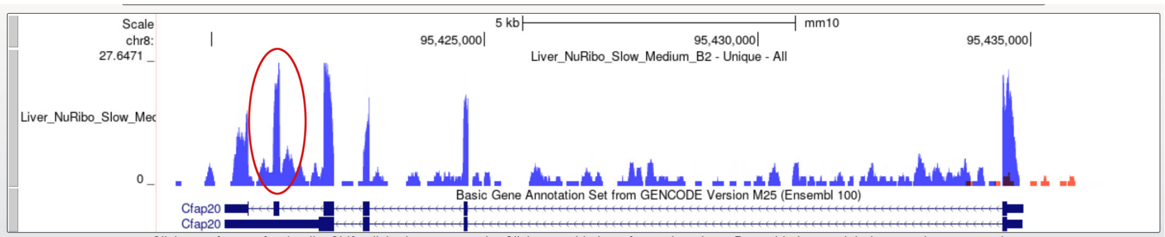
**ANSWER:** The RNA-Seq since signal is concentrated at the genes. Since it's the same number of reads, narrower footprint of signal has to mean higher peaks.

**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?



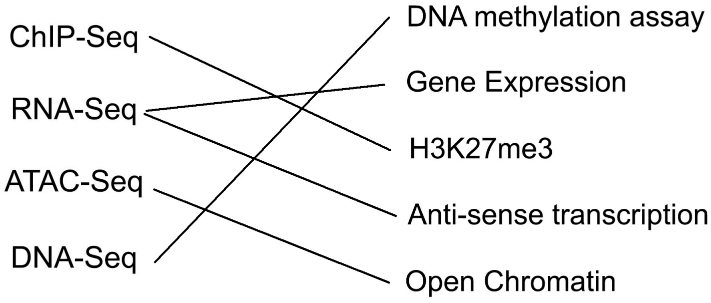
**ANSWER:** It is because Gm45854 is on the opposite strand, while we cannot easily tell the difference between signal from Gm45853 and signal from the intron of Csnk2as.

**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?


Scale
chr8:
27.6471 _
5 kb
mm10
95,425,000|
95,430,000|
95,435,000|
Liver_NuRibo_Slow_Medium_B2 - Unique - All
Liver_NuRibo_Slow_Mec
0 _
Basic Gene Annotation Set from GENCODE Version M25 (Ensembl 100)
Cfap20
Cfap20

**ANSWER:** The top one because we see a clear peak of signal (circled in red) that can only come from that small second exon (from the left).

**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.



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

**ANSWER:** True.

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

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

**ANSWER:** The FWER.

**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, the next generation must also have only one haplotype.

**ANSWER:** True. Even if there's a crossover, there's only one haplotype so it will basically cross over with itself and effect no change.

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

**ANSWER:** False, it's the opposite.

**Question 20.** A "progressive" alignment means: (circle the one correct answer)
(A) A multiple sequence alignment that adds one base at a time.
(B) A multiple sequence alignment that adds one sequence at a time. ⟵ **THIS ONE**
(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.

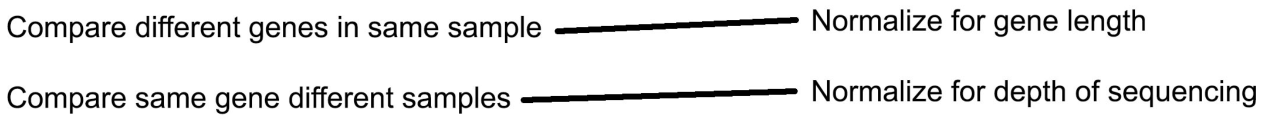
**ANSWER:** It is translated into amino acid sequence in all six possible frames and then each is aligned against the protein database.

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

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

**ANSWER:** Eventually it will end up in node B or C and then it will bounce back and forth between B and C forever.

**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?

**ANSWER:** It's either all A's or all T's, so $\frac{1}{2^5} + \frac{1}{2^5} = \frac{2}{32} = 0.0625$

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

**ANSWER:** 20 + 3 + 37 + 20 + 20 = 100. (I contributes to read length, but not D).