# University of Pennsylvania

## BIOL4536 Fall 2023

**Professor: Gregory R. Grant**

# Exam#2 (SOLUTIONS)

**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.

**ANSWER:** False.

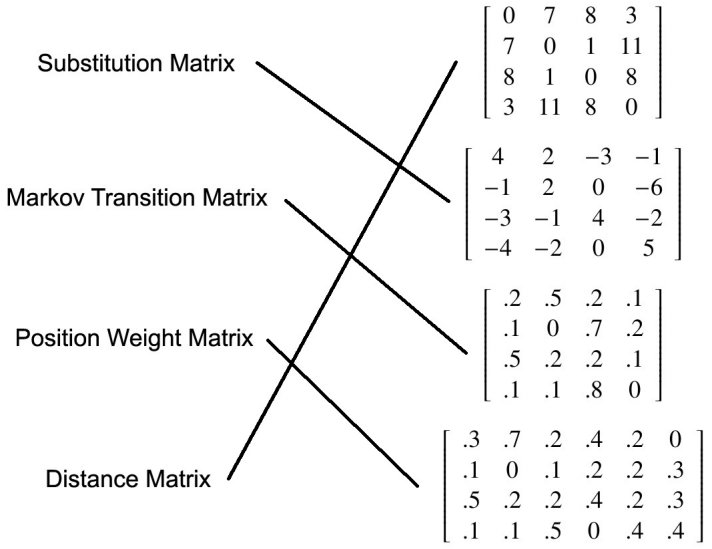
**Question 2.** True or False. Suppose you have two sequences $S_1$ and $S_2$ and a local alignment that involves the same number of bases of $S_1$ as it does $S_2$. Then the alignment does not have any indels.

**ANSWER:** False.

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

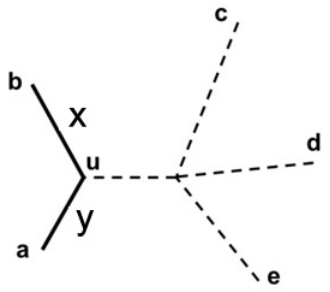
**ANSWER:** True.

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




**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 |



**ANSWER:** From the distance matrix we know $d(a, b) = 5$. Thus $x + y = 5$ so if $x = 3$ then $y = 2$.

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

**ANSWER:** True.

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

**ANSWER:** One.

Answer the same question for BLOSUM80:

**ANSWER:** Four.

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

**Question 8.** Consider protein BLAST. The 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.

**ANSWER:** They are simply the frequencies of each amino acid in the database.

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

**ANSWER:** False. It's applied to fewer sequences, not shorter ones.

**Question 10.** True or False, the total accumulated score of the random walk must be positive at the position where a max-excursion happens.

**ANSWER:** False.

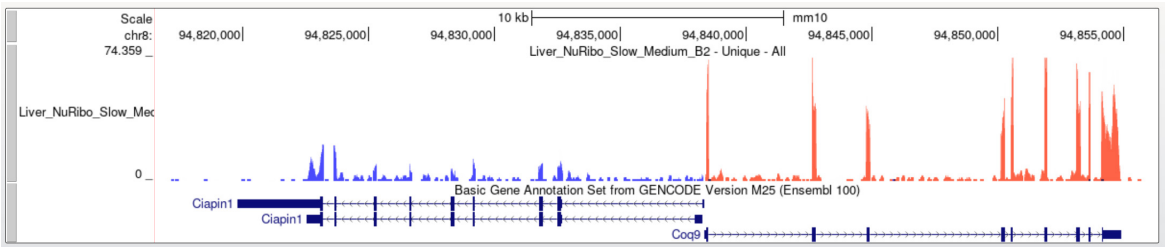
**Question 11.** Give one reason why RNA-Seq aligment is a more difficult problem than DNA-Seq.

**ANSWER:** Because of introns which constitute very large gaps that we don't see in DNA alignment.

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

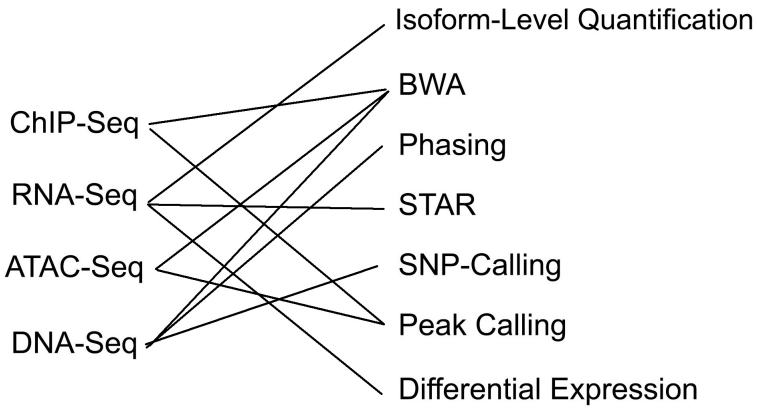
**ANSWER:** Because different isoforms of the same gene tend to share exons, so a read that aligned uniquely to the genome might still come from multiple isoforms.

**Question 13.** What do the colors represent? Which of the two isoforms of Ciapin1 appear to be the primary one that is expressed? Which of the two genes shown has a higher intron signal to exon signal ratio?




**ANSWER:** Blue means the read came from the minus DNA strand and red means it came from the plus strand. The second (lower) isoform seems to be the one expressed, because there is mainly only signal above the smaller 3'UTR.

**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.




**Question 15.** What are the two tasks for ChIP-Seq that we've discussed?
(A) Transcription Factor Binding ⟵ **THIS ONE**
(B) Gene Differential Expression Analysis
(C) SNP Calling
(D) Histone Modification ⟵ **THIS ONE**
(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 ⟵ **THIS ONE**
(C) 18
(D) 32
(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:** $0.01/20 = 0.0005$.

**Question 18.** True or False. You should never use a $q$-value cutoff great er than 0.05.

**ANSWER:** False.

**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 ⟵ **THIS ONE**

**Question 20.** Suppose a population has two haplotypes at three conscutive SNP locations: 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.

**ANSWER:** False.

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

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

**ANSWER:** Eventually it must visit node C and once it does it can never leave.

**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.

**ANSWER:** False. The $E$-value is not about the heuristic, it's about the multiple-testing problem of searching a large database.

**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.

**ANSWER:** True.

**Question 24.** Suppose an individual is heterozygous at a somatic 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 two reads do identify both of the SNP variants?

**ANSWER:** All possibilities are equally likely for (read1,read2). They could be: AA, AB, BA or BB. Two of these see both variants. So the probability is 50%.

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

**ANSWER:** 3 + 99 + 47 + 1 = 150.