University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

Exam#2 (27 More Review Problems)

Question 1. Suppose you have two sequences S_1 and S_2 where $|S_1| < 2|S_2|$. Is it possible to have two optimal local alignments between S_1 and S_2 that do not overlap? In other words the two alignments involve completely different bases from both sequences.

ANSWER: True. The fact that one is more than twice as long as the other is not even relevant, it's always true.

Question 2. Assume a pairwise alignment scoring scheme that scores +1 for a match and −1 for a mismatch or indel. In the following multiple sequence alignment, what's the "sum of pairwise alignments" score? *Note: an indel/indel scores 0*

A T _ G C G
A _ C G T _
A T C A C _

ANSWER: The total is −3. If a column has three equal things then the contribution from that column is $+1+1+1 = 3$. If two are equal and one is different then it's $+1 - 1 - 1 = -1$. But we don't score an indel/indel, so the last colum is just $-1 - 1 = -2$. So it's $+3 - 1 - 1 - 1 - 1 - 2 = -3$.

Question 3. True or False. A greedy algorithm never finds an optimal solutions.

ANSWER: False.

Question 4. A position weight is square. What are its dimensions?

ANSWER: 4×4.

Question 5. True or False. A distance matrix is symmetric.

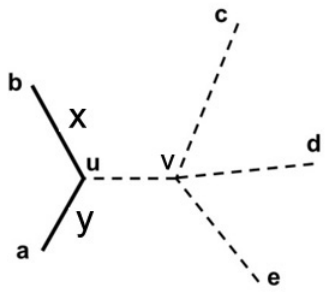
ANSWER: True. That's because the distance from a to b is the same as the distance from b to a.

Question 6. True or False. A distance matrix has ones on the diagonal.

ANSWER: False.

Question 7. Consider the following distance matrix and unrooted tree. What is $d(a,c)+d(b,c)-2d(u,v)-2d(v,c)$?

	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: The answer is 5. Draw a line along the path from a to c and do the same for b to c. That's the first two terms. Notice that the segment from u to v was added twice and then subtracted twice (that's the third term). And same for the segment from v to c. So all that's left is the segment from a to u and the segment from u to b, and that's $d(a,b)$ which from the distance matrix is 5.

Question 8. True or False. In a SAM file, the CIGAR string determines the exact location of the alignment.

ANSWER: False, you also need the chromosome and start position (two other entries in each SAM record).

Question 9. Suppose we are clustering rows of the following block to construct a BLOSUMN matrix. Is there a value of N such that $0 < N < 100$ that will result in only one cluster?

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

ANSWER: No because the second row is 100% different from all other rows, so it's always going to be in its own cluster.

Question 10. Suppose we are aligning to a DNA database. Is it possible that the background frequencies are different from 25% per nucleotide?

ANSWER: Yes. They'll probably be close to 25% but no reason to be exacly and depending on the species in the database might be quite different.

Question 11. True or False. Extension is the most time-consuming part of the BLAST algorithm.

ANSWER: True.

Question 12. True or False. In the BLAST random walk, it's possible that the max excursion is 0.

ANSWER: True. Could be 100% mismatches.

Question 13. True or False. An RNA-Seq read can be a multimapper to the genome but unambiguously increment the count of one gene.

ANSWER: True, it could align to two different places in the same gene.

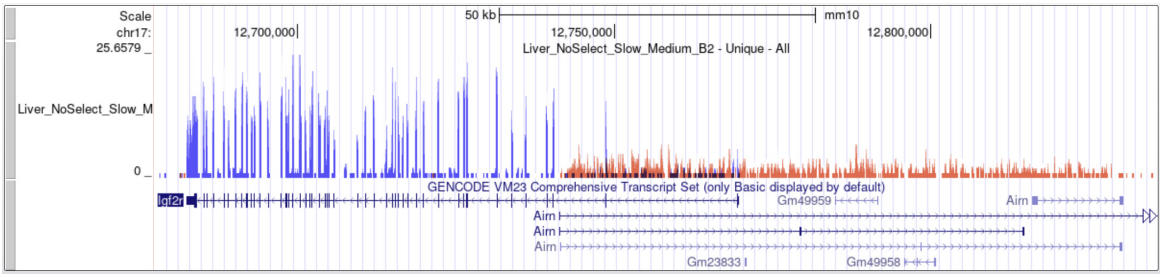
Question 14. True or False. It's possible for an RNA-Seq read to align uniquely to the genome but be ambiguous about whether it aligns to an intron or an exon.

ANSWER: True. For example, one gene can be located in the intron of another.

Question 15. True or False. It's possible for an RNA-Seq read to align uniquely to the genome but be ambiguous about whether it aligns to one of two different genes.

ANSWER: True. Different genes can overlap.

Question 16. Anti-sense transcription is when a gene is transcribed in the wrong direction. Does the following graphic show any anti-sense transcription?



ANSWER: Since the blue (reverse read) reads all lie within a gene on the reverse strand and the red (forward strand) reads all lie within a gene on the foward strand, there's not conclusive evidence for anti-sense transcription. It's possible, but not definite.

Question 17. True or False. We use different aligners for RNA-Seq as for ChIP-Seq.

ANSWER: True. They're very different alignment problems so need different algorithms. We used Bowtie for DNA and STAR for RNA.

Question 18. Which of the following are relevant for ChIP-Seq
(A) Epigenetics ⟵ **THIS ONE**
(B) DNA ⟵ **THIS ONE**
(C) RNA
(D) SNP Calling
(E) Peak Calling ⟵ **THIS ONE**
(F) Gene Regulation ⟵ **THIS ONE**

Question 19. Suppose there are 30,000 genes. Suppose we do an RNA-Seq DE analysis and that the largest q-value over all genes is 0.4. How many genes do we expect are DE in total?
(A) 12,000
(B) 18,000 ⟵ **THIS ONE**
(C) It cannot be determined

Question 20. You are performing 100 tests. You want to use the Bonferroni correction to control the FWER at the level 0.01. Suppose it turns out all genes are significant. What can you say about the maximum of the 100 p-values?

ANSWER: Bonferroni says the cutoff needs to be $0.01/100 = 0.0001$. If all genes are significant then all p-values must be less or equal to 0.0001. So, the max is less or equal to 0.0001.

Question 21. True or False. If all p-values and all q-values are one, then all null hypotheses are necessarily true.

ANSWER: False. There could still be false-negatives.

Question 22. True of False. The "F" in FWER and FDR stand for the same thing.

ANSWER: False. In FWER it stands for "Family" and in FDR it stands for "False".

Question 23. Consider five consecutive SNPs. Suppose a population has exactly two haplotypes with no exceptions. True or False. If an individual is heterozygous at one of the SNPs then they are heterozygous at all five.

ANSWER: False. Suppose for example the two haplotypes are $AAAAA$ and $ACCCC$. Then people can be homozygous at the first SNP and hetero at the other four.

Question 24. Consider the following Markov transition matrix. True or False. This Markov Chain perfectly models repeatedly flipping a fair coin.

$$\begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \end{array}$$

ANSWER: True. It doesn't matter if the current state is A or B, the probability that the next is A is $50/50$ and same for B, just like repeated coin flips.

Question 25. True or False. The BLAST algorithm one-hit method first extends hits as ungapped alignments.

ANSWER: True.

Question 26. We perform two independent tests $T1$ and $T2$ and consider them significant if their p-value is $\leq C$. True or False. The probability that at both of them are false-positives is exactly double the probability that one (and only one) of them is a false-positive.

ANSWER: False. That would be like saying it's 50% to flip a head in one flip, and therefore 100% to flip it if given two tries. That's obviously false.

Question 27. Consider the following CIGAR string: 3S97M. True or False. The entire read aligned from end-to-end with no indels.

ANSWER: False. The "S" part means the first 3 bases did not align.