7.36/7.91/20.390/20.490/6.802/6.874
PROBLEM SET 1. Sequence search, global alignment, BLAST statistics (19 Points)

Due: Thursday, February 20th at noon.

Problem 1. Sequence search (6 points)
To better understand inborn disorders of metabolism, you isolate a strain of mice that becomes ill unless fed a diet lacking phenylalanine. You sequence the genome of this mouse and find several differences from wildtype including a change to a region that encodes a highly expressed 68 nucleotide RNA which has sequence
5'-UGUACAUGAUGAAGUCAUAGCGAACGGAGAAGGGCCGGCUGAGGAA
ACUGCACGUCACCCUCCUGAAA-3'
in your strain and
5'-UGUACAUGAUGAAAACAGUCUCCCUCUUCUGAAUCUCGCUGAGGAA
ACUGCACGUCACCCUCCUGAAA-3'
in wildtype mice.
Search the sequence in your strain against the mouse genome and transcriptome using NCBI's BLASTn: from the BLAST homepage, click on "nucleotide blast" (not "Mouse") and use the "Mouse genomic + transcript" (G+T) Database, optimized for "Somewhat similar sequences". By expanding the "Algorithm Parameters" box at the bottom, set the Match/Mismatch scores to +1/-3.

(A) (1 pt.) How many statistically significant hits are there at an E-value of 0.05? In one sentence, what does an E-value of 0.05 mean? For transcript hits, what are the maximum reported scores, and are they raw scores or bit scores? (Click on the hyperlink to view individual hits.) To what parts of your RNA do these hits correspond, and what is the % match?

(B) (1 pt.) Using the E-value and reported score from the result with the highest % identity match from part (A), calculate the approximate length of the Mouse (G+T) Database.

(C) (1 pt.) Consider a query sequence Q of length L that matches perfectly to a sequence in the database, yielding a BLAST E-value E_1. How would the E-value change if only the first half of Q were searched against the database? In particular, would it stay the same, go up, go down, and how (linearly, exponentially, etc.)?

(D) (1 pt.) Returning to the BLAST results from part (A), to what genes and RNA classes do the transcript hits with E-values below 0.05 belong? Does your RNA match the sense or antisense direction of these hits? (Click on the hyperlink of the hit and look at the "Strand" section, which tells you the DNA strand of the Hit/Query.)

(E) (2 pts.) After performing an RNA-protein affinity purification (pull-down) from mouse cell lysates followed by mass spectrometry, you determine that your RNA interacts with the product of the *ADAR1* gene. What does this enzyme do, and what type of RNA does this enzyme act on? Looking back at the function and strand of the gene hit to the second part of your RNA, state a hypothesis as to how your RNA might function to cause your mouse's metabolic disorder. (Hint: on the BLAST hit entry corresponding to the mRNA, click on the "Graphics" link to see the hit in red and how your query at the bottom overlaps with it. If ADAR1 acts at the UAU codon, what is the resulting change during translation?)

Problem 2. Gapped sequence alignment (6 points)
In this problem, you will use the algorithms discussed in class to find the optimal alignment for a pair of short peptides.

(A) (1 pt.) In order to perform this alignment, you must first choose a scoring matrix. For example, you could use a constant match and mismatch penalty of 1 and -1, respectively, so that $S_{ij} = 1$ if $i = j$ and $S_{ij} = -1$ otherwise. Is this a good idea? Why or why not? In one sentence, briefly describe how you might obtain a better scoring matrix for protein comparison.

(B) (1 pt.) You decide to explore more commonly used protein alignment scoring matrices instead. Compare the score for aligning two tryptophans (W) to the score for aligning two alanines (A) in the PAM250 scoring matrix. Both of these alignments are "matches", so why are these scores so different?

(C) (2 pts.) Perform a **global** alignment of the two peptides ATWES and TCAET, using the Needleman-Wunsch algorithm to fill out the alignment matrix below. Use the **BLOSUM62** scoring matrix and a linear gap penalty of 2.
After filling out the matrix, circle the traceback path and write the final alignment. If there are multiple traceback paths, write out all top-scoring alignments.

	Gap	A	T	W	E	S
Gap	0					
T						
C						
A						
E						
T						

Final Alignment:

(D) (**2 pts.)** Different scoring matrices and gap penalties can give very different alignment results. Below is the alignment of the peptides from part (C) using the **PAM250** scoring matrix (same gap penalty). The traceback path is shaded.

	Gap	A	T	W	E	S
Gap	0	-2	-4	-6	-8	-10
T	-2	**1**	1	-1	-3	-5
C	-4	**-1**	-1	-3	-5	-3
A	-6	-2	**0**	**-2**	-3	-4
E	-8	-4	-2	-4	**2**	0
T	-10	-6	-1	-3	0	3

What is the resulting alignment?

Compare the optimal alignments obtained using the BLOSUM62 and PAM250 scoring matrices. Why are they different?

Problem 3. Sequence similarity search statistics (7 points)
You are conducting local nucleotide sequence alignments with your favorite local alignment tool (e.g. BLAST) with match and mismatch scores of +1 and -1 respectively. You align a 100bp query sequence to a 1Mbp genome and find that a 20-nt subsequence from your query is a perfect match.

For each of the following cases, calculate the significance of a 20-nt perfect match (assume $K = 1$ in each case):

(A) (2 pts.) Query sequence and genome both have approximately balanced base composition A=C=G=T=25%).

(B) (1 pt.) Query sequence and genome are both highly A-T rich (A=T=40%, C=G=10%).

(C) (1 pt.) Query is moderately **A**+**T**-rich (A = T = 30%, C = G = 20%) but genome is moderately **C**+**G**-rich (A = T = 20%, C = G = 30%).

(D) (1 pt.) Briefly explain why the ordering of the P-values from (A) - (C) makes sense.

(E) (2 pts.) Design a new scoring system for application to searching a 20 nt query of unbiased composition against a highly A+T-rich genome (as in (B) above) that will increase the sensitivity for detection of matches to that genome by drawing lines from each box on the left to its new score in the right box (+1, 0, or -1 for different types of matches/mismatches). What would the P-value of a perfect match to this query (with 5 A's, 5 C's, 5 G's, 5 T's) be using your new scoring system?

A/A or T/T match	+1
C/C or G/G match	0
mismatch	-1

MIT OpenCourseWare
http://ocw.mit.edu

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology

Spring 2014

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.