Homework 1: Sequence Analysis

Due: Friday October 16th by end of day. Submit by email.

Please solve the following problems and hand in your answers as a zip file. Make sure to read the entire problem set before attempting the individual problems. For the programming question (1) you may use any programming language you choose.

1. Implement a dynamic programming algorithm for global alignment of two nucleotide sequences. It should take as input two strings q and d of arbitrary length m and n, the provided symmetric scoring matrix M_{4X4} in file 'score_matrix.txt' provided, and gap penalty $g = 3$. (*i.e.*, the Needleman-Wunsch algorithm from class). In this version of the problem, you are looking for the alignment with the **smallest** score. For example aligning A with T adds 5 to the score, aligning A with a gap adds 3 to the score, aligning A with A adds 0 to the score (see score_matrix.txt).
   Write a function that takes two sequences and the scoring matrix as input, and produces the alignment score and actual alignment as output. The identified alignment should be the alignment with minimum score. If there is more than one alignment with the minimum score (*i.e.,* ties) you only need to output one minimal alignment. Write a program that reads in the provided sequences (in file sequence_q1.fasta – 3 sequences of length 20 and 2 of length 100), the scoring matrix and the gap penalty and aligns the three 20 nucleotide sequences to each other (3 alignments), the two 100 nucleotide sequences to each other (1 alignment), and one 20 nucleotide sequence to one 100 nucleotide sequence (1 alignment)). Report your outputs (both score and actual alignment) for the given data set.
   Assume that you are aligning nucleotide sequences, so your symbols will be A, T/U, C, and G (be sure to handle both T and U). Your program must be able to handle sequences of length up to 100. If your program places limits on m and n please state so. Your output should show the alignment including a '-' to represent a gap. You should email the instructor a zip file containing your source and the output requested.

2. You are collaborating with an experimental cancer research lab. The lab recently identified a gene mRNA transcript X associated with a specific type of cancer and they need your help. The sequencing of the transcript revealed that it had the following sequence:

1	CTCTGCTCTA	CAAGCCTGTG	GACCGTGTGA	CGAGGAGCAC	GCTGGTCCTC	CATGACTTGC
61	TGAAGCACAC	TCCTGCCAGC	CACCCTGACC	ACCCCTTGCT	GCAGGACGCC	CTCCGCATCT
121	CACAGAACTT	CCTGTCCAGC	ATCAATGAGG	AGATCACACC	CCGACGGCAG	TCCATGACGG
181	TGAAGAAGGG	AGAGGGAGAA	GACAGGATGA	AAGCTTCATC	AACGAGGAAG	AGATTACTCC
241	TTATGGAAGA	AGCCCTTCAG	CGGCCAGTAG	CATCTGACTT	TGAGCCTCAG	GGTCTGAGTG
301	AAGCCGCTCG	TTGGAACTCC	AAGGAAAACC	TTCTCGCTGG	ACCCAGTGAA	AATGACCCCA
361	ACCTTTTCGT	TGCACTGTAT	GATTTTGTGG	CCAGTGGAGA	TAACACTCTA	AGCATAACTA
421	AAG					

   Do the following:

   (a) BLAST this sequence against human genomes using the NCBI Blast server at `http://blast.ncbi.nlm.nih.gov/Blast.cgi`. In the BLAST page, select "blastx" (to match the nucleotide sequence against human protein sequences). Enter the above query sequence (also contained in the file sequence_q2.fasta) and select "Reference proteins (refseq_protein)" for the search database. Note that RefSeq genes and proteins are generally well studied and manually verified genes and proteins. Click the "Algorithm parameters" and notice the parameters we discussed in class – change the "Max target sequences" to 50 but do not change any other parameters. Click the "BLAST" button to start the search.

   (b) Copy and paste the Graphic Summary and the Descriptions for the search result into your answer for this question.

   (c) From the Graphic Summary and the matches of the sequences to different parts of the genome assembly, what interesting thing do you notice about the continuity of the matches against the query sequence? Is the entire query sequence matched to any sequence in the database?

   (d) Notice that the descriptions for some of the matches appear similar to each other. For the first two matches, with quite different descriptions, follow the web links to the Genes corresponding to the proteins (look

for link starting with ”GENEID” under the proteins). These should take you to the Entrez gene page which describes the genes in more detail. Entrez genes are usually tied to functions using GeneRIFs. These are usually in the middle of the Entrez gene web pages. For each of the two genes, copy the first function available from GeneRIFs (you may have to look at several genes to find ones with GeneRIFs). Alternatively, extract the top function from the Gene Ontology section in the Entrez gene page for each gene.

(e) (Bonus question) In the Summary sub-section of the Summary section of the Entrez Gene Webpages for either of the two genes you looked at, did you notice anything that would explain your observation in (c) above?