Name: Tianru Zhang PID: A15432834

Q1. Why does the DNA sequence have more dots than the protein sequence plot? HINT: what do you know about DNA composition vs protein composition?

Because DNA only has 4 nucleotides(characters), but protein has 20 amino acids(20 characters), there is a higher chance for 2 random DNA sequences to match compared to protein sequences. There are more signal in protein compared to DNA.

Q2. How can we increase the signal to noise ratio?

(more noise in DNA sequences compared to protein) Increase window size(the number of monomers included in the sequence) and decrease the stringency moderately to have more signals.

Q3. What does a 'Match stringency' larger than 'Window size' yield and why?

It will give an error, because the match stringency specifies number of match characters required per window, and cannot be larger than window size(number of total characters included). For example, when the window size is 4, the number of maximum match is 4, so the stringency cannot be more than 4.

Q4. What are the major weaknesses of this approach? HINT: is your inner nerd happy with this approach? How would you use it to determine if a second set of sequences was more similar to each other than a first set of sequences?

There are only visual/graphical and qualitative representations, no numerical/quantatative representations telling whether two sequences are more related. Inefficient when comparing long sequences.

Section 2: Needleman-Wunsch Alignment

Sequence alignment methods often use something called a 'dynamic programming' algorithm that can be usefully considered as an extension of the dot plot approach. Here we have two sample sequences, and we'd like to use the **Needleman-Wunsch algorithm** discussed in class to align them.

Sequence 1: **ATTGC**

Sequence 2: **AGTTC**

Alignment 1: A _ T T G C
| | | | |
A G T T _ C

Score: 4×2 − 2×2 = 4

Alignment 2: A G T T C
| | | | |
A T T G C

Score: 3×2 − 2 = 4

		A	G	T	T	C
	0	-2	-4	-6	-8	-10
A	-2	2	0	-2	-4	-6
T	-4	0	1	2	0	-2
T	-6	-2	-1	3	4	2
G	-8	-4	0	1	2	3
C	-10	-6	-2	-1	0	4

Q5. Using a **match score of +2**, a **mismatch score of -1**, and a **gap score of -2**. Fill in the table below and translate it into a alignment. What is the optimal score for this alignment? Is there one unique alignment with this score?

Optimal Alignment Score: 4

There are 2 alignments.

Practice makes perfect. Again use the **Needleman-Wunsch algorithm** discussed in class to align the following sequences:

Sequence 1: **TATAG**

Sequence 2: **GTTAC**

G T T A C
T A T A G.

2×2 − 3 = 1

Optimal Alignment Score: 1

		G	T	T	A	C
	0	-2	-4	-6	-8	-10
T	-2	-1	0	-2	-4	-6
A	-4	-3	-2	-1	0	-2
T	-6	-5	-1	0	-2	-1
A	-8	-7	-3	-2	2	0
G	-10	-6	-5	-4	0	1

Q6. Using a **match score of 2**, a **mismatch score of -1**, and a **gap score of -2**. Write out your alignment matrix (table), fill in the values and translate your results into all optimal alignments. What is the optimal alignment score for these sequences? Write out all alignments consistent with this score?

Q7. Using the default settings for NCBI BLAST, can you find any homologs for this protein in Humans? HINT: try using the *LIMITS* and *FILTERING* options we covered in the last lab.

No homologs of human protein are founded.

Q8. Try changing the database to **refseq_protein**. From the results, select a few proteins and find the common name for the species. What trend do you notice as you move down the results list? HINT: search google for the species name.

The protein is hedgehog, and the fly name is Drosophila, the common name is fruit fly. As I move down the results, other fly species show up, with lower percentage identical.

Q9. Finally, try also limiting the search to only *H. Sapiens*. HINT: you can simply type the Taxon ID **9606** in the "**Organism**" box. What function do these proteins have?

Hedgehog protein preproproteins, including desert hedgehog protein, indian hedgehog protein, sonic hedgehog protein. Percent indentical are around 50%. The low e-value indicates that it's unlikely to have false positive hits, and the result is trustable. The proteins aresignaling protein is instrumental in patterning the early embryo. They act as inductive signals in patterning of the ventral neural tube, the anterior-posterior limb axis. Mutations disrupt limb patterning.

Q10. What function do you think this protein performs for your collaborators' organism?

The functions are similar to those in human embryos, they contribute to embryogenesis and limb paterning.