*Name:*_____
ID : _____

ECS 129: Structural Bioinformatics
Midterm: solutions
February 15, 2024

Notes:

1) The midterm is open book, open notes.
2) You have 45 minutes, no more: I will strictly enforce this.
3) The midterm is divided into 2 parts and graded over 90 points.
4) You can answer directly on these sheets (preferred), or on loose paper.
5) Please write your name at the top right of each page you turn in!
6) Please, check your work! **Show your work** when multiple steps are involved.

Part I (5 questions, each 10 points; total 50 points)

(These questions are multiple choices; in each case, find the **most plausible** answer)

1) In the dynamic programming matrix below, what is the score in the cell identified with an interrogation mark (?). Assume that the score for a perfect match is set to 10, the score of a mismatch is set to -2, and gap penalties are set to -2, independent of length. Gaps at the beginning count.

	G	Y	W	W	C	A
W	-2	-4	8	8	-4	-4
W	-4	-4	6	18	6	4
C	-4	-6	-6	**4**	28	14

A) -6
B) 18
C) 6
D) 4
E) 0

2) We want to find the best alignment(s) between the protein sequences WWYCTY and WCFTY. The scoring scheme S is defined as follows: S(i,i) = 10, S(i,j) = 5 if i and j are both aromatic amino acids (i.e. W, F, or Y), and S(i,j) = 0 otherwise. There is a constant gap penalty of 5 (gaps at the beginning are considered, see below). The score Sbest and the number N of optimal alignments are (show your final dynamic programming matrix for full credit):

	W	W	Y	C	T	Y
W	10	5	0	-5	-5	0
C	-5	10	5	15	5	5
F	0	10	15	5	15	15
T	-5	5	10	15	20	15
Y	0	10	15	10	15	30

A) Sbest = 40, N = 1
B) Sbest = 35, N = 2
C) Sbest = 35, N = 1
D) Sbest = 40, N = 2
E) Sbest = 30, N = 1

*Name:*_____
ID : _____

This was in fact a "trick" question. The matrix shows that the best score is 30. However, there are two best alignments:

```
WWYCTY          WWYC-TY
WCF-TY          W--CFTY
```

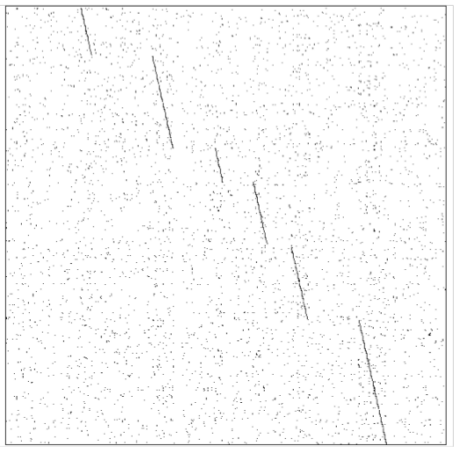
both with scores of 30. … I did not penalize anyone who answered E with a correct matrix, or anyone who changed the 30 to 35 to give the answer B. I rewarded those who found the "trick"…

3) How many DNA coding sequences (where a coding sequence includes the START and STOP codon, but no introns) could lead to the following protein sequence: **Met- Lys-Leu-Trp-Ser-Phe-Trp-Thr** assuming the standard genetic code?

A) 1
B) 576
C) 1152
D) 1728
E) 4096

N = 1 (Met) x 2 (Lys) x 6 (Leu) x 1 (Trp) x 6 (Ser) x 2 (Phe) x 1 (Trp) x 4 (Thr) x 3 (STOP)

4) The dotplot shown below compares the DNA sequence of the actin muscle gene from *Pisaster ochraceus* (horizontal) with the mRNA corresponding to the same gene (vertical). The six regions of high similarity that shows as black lines correspond to:

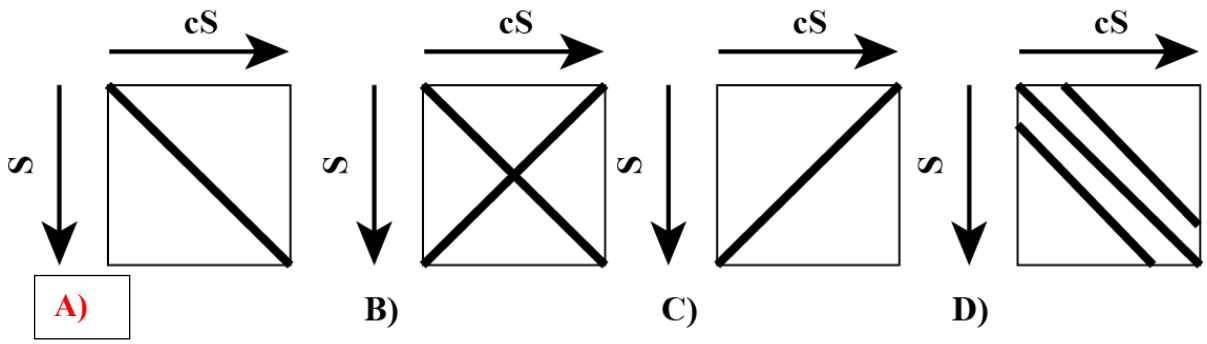


A) Introns
B) Repeats
C) Inverted repeats
D) Exons
E) All of the above

Conserved regions between RNA and DNA corresponds to coding regions, hence exons.

5) Given the DNA sequence S= 5'-GAATTC-3', how does the dotplot between S and its complementary, cS, look like?

*Name:*_____
ID : _____



Note that cS = 5'-GAATTC-3', i.e. S and cS are equal (S forms a double stranded molecule with itself). As such, we will mostly see the first diagonal on the dotplot. You note that I indicated direction on the axes… this means that we need to know how to read the sequences, and sequences are always written 5' to 3'. Note also that D is not completely wrong as there is some repeats, but A was definitely the most plausible answer.

Part II (one question, 10 points)
In one of the strands of a double stranded DNA molecule there is 30 % of Adenine ([A]=30%) and 24 % of guanine ([G] = 24%). Calculate the following, if possible (if impossible, write "I"):

a) [A]+[G] : 54%
b) [T] : I
c) [C] : I
d) [T]+[C] : 46 %
e) [A] on the other strand : I
f) [T] on the other strand : 30%
g) [A]+[T] on the other strand : I
h) [G] on the other strand : I
i) [C] on the other strand : 24%
j) [G]+[A] on the other strand : 46%

Part III (two questions, 10 points each: total 20)
Below is the double-stranded DNA sequence of part of a hypothetical bacterial genome, which happens to contain a very small gene.

```
5′ — TATAAATTATGTCTGCTATAAAATAACCCGGT— 3′
     |||||||||||||||||||||||||||||||||
3′ — ATATTTAATACAGACGATATTTTATTGGGCCA— 5′
```

a) What is the sequence of gene and of the longest protein that can be produced by this DNA sequence? Label the N and C termini.

```
5′ — TATAAATTATGTCTGCTATAAAATAACCCGGT— 3′
     |||||||||||||||||||||||||||||||||
3′ — ATATTTAATACAGACGATATTTTATTGGGCCA— 5′
```

The top strand sequence S contains one ATG (start codon) with one TAA (stop codon), in phase with the ATG. Consequently, the longest ORF is:

5' ATG TCT GCT ATA AAA TAA -3'

The corresponding RNA sequence is:

5' AUG UCU GCU AUA AAA UAA -3'

The protein sequence is obtained directly using the genetic code:

Nter – Met Ser Ala Ile Lys – Cter

b) Propose a single base pair **deletion** that will lead to the mutated sequence still coding for a protein, albeit smaller, with the same START codon. ***Note that you still need a STOP codon in phase with the START codon***. Give the sequence of the shorter protein. Label the N and C termini.

One option (there are others) is to remove the A before the TAAAA in the top sequence:

```
5′ — TATAAATTATGTCTGCTATAAAATAACCCGGT— 3′
     ||||||||||||||||||||||||||||||||
3′ — ATATTTAATACAGACGATATTTTATTGGGCCA— 5′
```

This leads to the new sequence:

```
5′ — TATAAATTATGTCTGCTTAAATAACCCGGT— 3′
     ||||||||||||||||||||||||||||||
3′ — ATATTTAATACAGACGAATTTTATTGGGCCA— 5′
```

This sequence has a new TAA, in phase with the original START codon ATG. This leads to a new longest ORF is:

5' ATG TCT GCT TAA -3'

The corresponding RNA sequence is:

5' AUG UCU GCU UAA -3'

The protein sequence is obtained directly using the genetic code:

Nter – Met Ser Ala– Cter

In fact, coming back to the original sequence:

5' ATG TCT GCT ATA AAA TAA -3'

Notice that deletion of any of the letters colored in green would lead to the same effect, possibly with different protein sequences

*Name:*_____
ID : _____

Part III: (one question, 10 points)

We want to find the best alignment(s) between the protein sequences FAFWC and FWFC. The scoring scheme S is defined as follows: S(i,i) = P, and S(i,j) = M otherwise. There is a constant gap penalty of G (gaps at the beginning are considered). The dynamic programming matrix is shown below. What were the values of P, M, and G? Write the best alignment found using those values.

	F	A	F	W	C
F	5	-4	3	-4	-4
W	-4	3	1	8	1
F	3	1	8	-1	6
C	-4	1	-1	6	11

P = 5, G = -2, M = -2

There was a typo in the text… it should have been the "best alignments", as there are three alignments with scores of 11:

```
FAFWC        FAFW-C        FAFW-C
FWF-C        --FWFC        F--WFC
```

No one was penalized for only writing one alignment.

*Name:*_____
ID : _____

Appendix:

Appendix A: Genetic Code

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met/Start	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G