*Name:*_____
ID : _____

ECS 129: Structural Bioinformatics
Midterm
February 18, 2026

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 3 parts and graded over 80 points (i.e. there is 5 extra credit 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 (6 questions, each 10 points; total 60 points)

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

1) Cytochrome P450 enzymes form a super-family of heme-containing oxygenases that are found in all kingdoms of life. Let us consider these three examples: (A) the human CYP46A1, (B) a human prostacyclin synthase, and (C) Xpla, a cytochrome P450 from rhodococcus (aerobic, gram-positive bacterium). (A), (B) and (C) are homologous proteins; what else can you say?
   a) (A), (B) and (C) are orthologous
   b) (A), (B) and (C) are paralogous
   c) (A), (B) and (C) are analogous
   d) (A) and (B) are paralogous, while (A) and (C) are orthologous
   e) (A) and (B) are orthologous, while (A) and (C) are paralogous

A and B are both human proteins and as such paralogous, while A and C are from different species, hence orthologous.

2) The protein sequence alignment shown below has a total score of 20. Knowing that the score for an exact match is 5 and the score for a mismatch is -4, what is the score used for the (constant, i.e. independent of length) gap penalty:

```
GCTGGAAG-GCA-T
GC-C-TAGAGCACT
```

   a) -1
   b) -2
   c) -3
   d) -4
   e) Undefined (any value would give the same total score)

Alignment score: S = 5 + 5 + G -4 +G -4 + 5 + 5 +G+5+5+5+G+5 = 20, hence 4G+32 = 20, G=-3

*Name:*_____
ID : _____

3) You are told that the DNA coding strand that encodes the peptide: Met-Trp-X-Trp-Met contains **60% guanine (G)**. What is amino acid X?
   a) Gly
   b) Ala
   c) Lys
   d) Phe
   e) Not enough information

The DNA includes 5 times 3 = 15 bases. As it contains 60 % guanine, it actually contains 9 guanines. The codons for Met and Trp are ATG and TGG, respectively. Therefore, we have 6 guanines with the codons from Met and Trp; amino acid X includes three G in its codon. Hence the codon for X is GGG, and X=Gly.

4) We want to find the best alignment(s) between the DNA sequences TCATAGA and TCTAGC. The scoring scheme S is defined as follows: S(i,j) = 3 if i = j (perfect match), S(i,j) = 2 if i and j are two different purines, S(i,j)=1 if i and j are two different pyrimidines, and S(i,j)=0 otherwise. There is a constant gap penalty of -1 (penalty for the first position counts; see table below). The score Sbest and the number N of optimal alignments are (***show your final dynamic programming matrix and the best possible alignment (s) for full credit***):

	T	C	A	T	A	G	A
T	**3**	0	-1	2	-1	-1	-1
C	0	**6**	2	3	2	2	2
T	2	3	6	**8**	5	5	5
A	-1	2	8	6	**11**	9	10
G	-1	2	7	8	9	**14**	12
C	0	5	5	8	8	10	**14**

   a) Sbest = 13, N = 2
   b) Sbest = 13, N = 1
   c) Sbest = 14, N = 1
   d) Sbest = 14, N = 2
   e) Sbest = 14; N = 3

The best alignment is:

```
TCATAGA
TC-TAGC
```

5) Dynamic programming, popular for sequence alignment, can also be used for spell checking. Assuming that a match is worth 10, a mismatch is worth 5, and a gap "costs" -5, which of these two words is closest to the word "graffe" typed by a user? ***Write the score of the optimal alignment next to each word (gaps at the start or at the end do not count).***
   a) gaff best score: = 35
   b) graft best score: = 45
   c) grail best score: = 40

*Name:*_____
ID : _____

d) giraffe ***Best score: 55***
e) raft best score: 35

	G	R	A	F	F	E
G	10	5	5	5	5	5
A	5	10	15	10	10	10
F	5	10	10	25	20	15
F	5	10	15	20	**35**	25

	G	R	A	F	F	E
G	10	5	5	5	5	5
R	5	20	10	10	10	10
A	5	10	30	20	20	20
F	5	10	20	40	35	30
T	5	10	20	30	**45**	40

	G	R	A	F	F	E
G	10	5	5	5	5	5
R	5	20	10	10	10	10
A	5	10	30	20	20	20
I	5	10	20	35	30	30
L	5	10	20	30	**40**	35

	G	R	A	F	F	E
G	10	5	5	5	5	5
I	5	15	10	10	10	10
R	5	15	20	15	15	15
A	5	10	25	25	20	20
F	5	10	15	35	35	25
F	5	10	15	30	45	40
E	5	10	15	25	35	**55**

	G	R	A	F	F	E
R	5	10	5	5	5	5
A	5	10	20	10	10	10
F	5	10	15	30	25	20
T	5	10	15	20	**35**	30

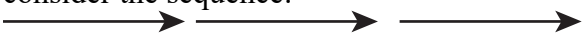
*Name:*_____
ID : _____

6.A **single-stranded DNA** molecule contains: 30% adenine, and 50% pyrimidine. This strand is used as a template to synthesize its perfectly complementary strand. This complementary strand also contains 30% Adenine. What percentage of bases in the final double stranded DNA are guanines?

a) 15%
b) 20%
c) 25%
d) 30%
e) Not enough information

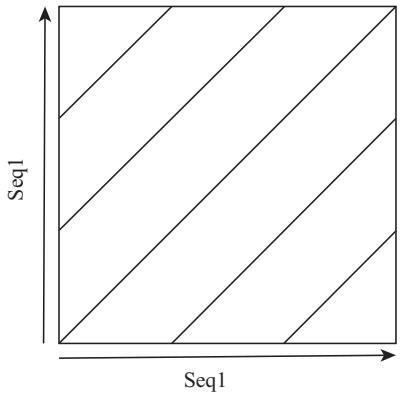
The original strand contains 30% A, 20% G, and 50% (T+C). Let T1 and C1 be the percentage of T and C in this strand. The complementary strand contains 30% T, 20% C, and 50% purines. As it contains 30% A, it contains 20% G. Both strands contain 20% G, hence the double stranded DNA contains 20% G.

Part II (one question, 15 points)
Dotplots are useful, graphical tools for detecting information within a sequence. For example, consider the sequence:



ACDEFGHIACDEFGHIACDEFGHI

The dotplot below detects the presence of internal repeats as lines parallel to the main diagonal:

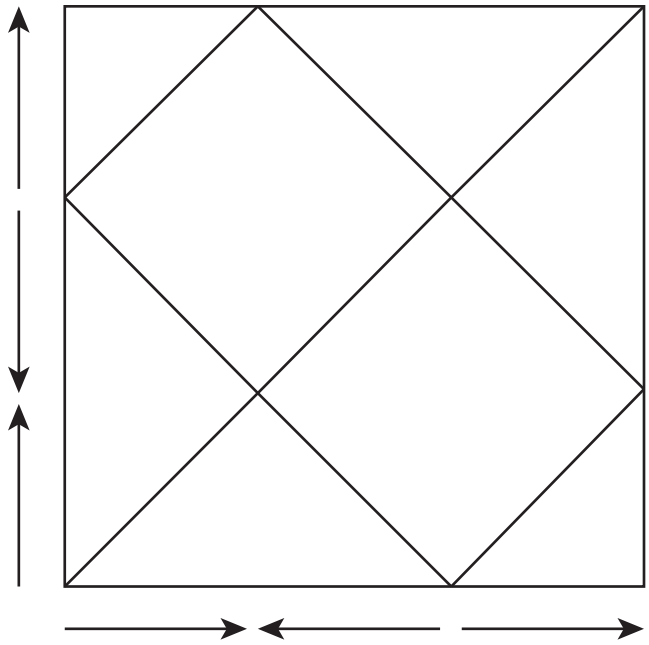


Draw schematically the dotplot for the following sequence:

ACDEFGHIIHGFEDCAACDEFGHI

*Name:*_____

ID : _____



*Name:*_____
ID : _____

Part III (total 10)

We want to find the best alignment(s) between the protein sequences WWYCTY and WCYTY. The scoring scheme S is defined as follows: S(i,i) = P (perfect match), and S(i,j) = M otherwise (mismatch). There is a constant gap penalty of G (gaps at the beginning are considered). The partial dynamic programming matrix is shown below:

	W	W	Y	C	T	Y
W						3
C						13
Y						26
T						28
Y	3	13				43

a) Can you find P, M, and G (note that we assume $P > 8$, $4 < M < 8$, and $-3 < G < 0$). Explain your work

Note first that:

	W	W	Y	C	T	Y
W	P					
C	M+G					
Y	M+G					
T	M+G					
Y	M+G					

Hence M+G = 3.
Now (as P > 3):

	W	W	Y	C	T	Y
W	P	P+G				
C	3	P+M				
Y	3	P+M+G=P+3				
T	3	P+M+G=P+3				
Y	3	P+M+G=P+3				

Therefore P+3 = 13, P = 10.

	W	W	Y	C	T	Y
W	10	10+G	3	3	3	3
C	3	10+M	13	20+G	13	13
Y	3	13	20+M	13+M	23	30+2G
T						
Y						

Hence 30+2G= 26, i.e. G = -2 and since M+G = 3, M = 5.

b) Complete the dynamic programming matrix and find the best alignment (s)

	W	W	Y	C	T	Y
W	10	8	3	3	3	3
C	3	15	13	18	13	13
Y	3	13	25	18	23	26
T	3	13	18	30	33	28
Y	3	13	23	28	35	43

The best alignment is:

```
WWYCTY
WCY-TY
```

*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