

Name: _____
 ID : _____

ECS 129: Structural Bioinformatics

March 16, 2026

Notes:

- 1) The final exam is open book, open notes.
- 2) The final is divided into 3 parts and graded over 100 points.
- 3) You can answer directly on these sheets (preferred), or on loose paper.
- 4) Please write your name at least on the front page!
- 5) Please, check your work! If possible, show your work when multiple steps are involved.

Part I (15 questions, each 4 points; total 60 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 -3, and gap penalties are ignored.

	A	T	C	Y	A	G
A	10	-3	-3	-3	10	-3
Y	-3	7	7	20	7	7
G	-3	7	4	4	17	30

- a) 10
- b) 20
- c) 30**
- d) 40
- e) 0

2) 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:

	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

- a) P= 5, M=0, G=0
- b) P= 5, M=-1, G=-1
- c) P= 5, M=-2, G=-2**
- d) P= 5, M=-3, G=-1
- e) P=5, M=-1, G=-3

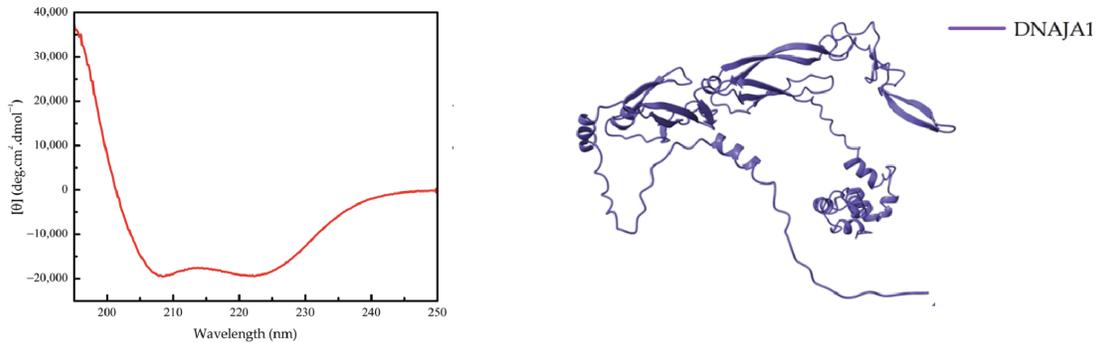
On first row, first position, from F-F=5, we get P = 5; From second F-F=3, we get P+G=3, hence G = -2. Finally, from F-W = -4, we get M+G=- 4, hence M=-2

3) Which of the following statements on the Smith and Waterman algorithm for pair-wise sequence alignment is most likely true?

- a) The dynamic programming matrix for the Smith and Waterman algorithm only includes non-negative values,**
- b) The Smith and Waterman algorithm is based on dynamic programming and as such requires multiplicative scores,
- c) Contrary to the Needleman and Wunsch algorithm, the Smith and Waterman algorithm generates a single optimal alignment,

Name: _____
ID : _____

- d) The Smith and Waterman algorithm has a complexity of $O(N^5)$ and as such can only be used on very short sequences,
e) None of the above.
- 4) You are given the task to find the structure of human DNAJA1, a heat shock co-chaperone involved with proteostasis. You have a sample of the protein in solution, and you were able to generate its CD spectrum (left, below). As you are in a hurry, instead of trying to find its 3D structure experimentally, you use AlphaFold and find the structure show on the right, below.



Based on your understanding of CD spectra,

- a) The AlphaFold prediction is spot on as the CD spectrum is typical of a protein with a lot of coil regions and many β -sheets
b) The AlphaFold prediction is mostly correct as the CD spectrum shows that the protein has many coil regions
c) **The AlphaFold prediction should be questioned: the CD spectrum reveals a mostly helical protein.**
d) I cannot say as the CD spectrum does not provide information on the structure of a protein.

5) You are given a single strand S of DNA. You are told that this sequence contains 20 % of Adenine and that the corresponding double stranded DNA contains 46% of GC base pairs. How much thymine (in percent) does S contain?

- a) 10%,
b) 20%,
c) **34%,**
d) 44%,
e) Not enough information

We know $A(S) = 20\%$. We know GC total is 46%, therefore AT total is 54%. Hence $A(S)+A(cS)=54$, i.e. $A(cS)=34\%$ (where cS is the complementary strand). As $T(S)=A(cS)$, $T(S)=34\%$.

6) The best global alignment found between sequences TATC and TTC

is: TATC

T-TC. It has a total score of 12. The scoring scheme uses a match score of 5 and a mismatch score of -2. What is the constant gap penalty d used for this alignment?

- a) 0

Name: _____
ID : _____

- b) -1
- c) -2
- d) -3
- e) None of the above

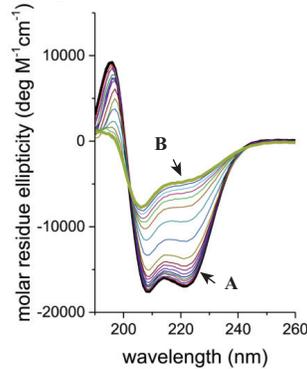
Score=5+G+5+5=12, hence G=-3.

7) A protein sequence contains one Histidine (HIS) residue. You want to create a new protein sequence, with this HIS being replaced with a Methionine (MET). To do this, you first generate the DNA corresponding to the original protein, then mutate this DNA using only substitution to get the sequence corresponding to the new protein. What is the minimum number of mutations needed?

- a) 1
- b) 2
- c) 3
- d) 0

The codon for MET is AUG, while the codon for His is either CAU or CAC; we need to change all three nucleotides.

8) The figure below illustrates the thermal unfolding of recombinant human Apo A-1 as seen by CD spectroscopy. Apo A-1 is a mostly helical protein. Note that increase in temperature leads to unfolding, i.e., loss of structure, including secondary structures).



The CD spectra were recorded between 5°C and 90°C recorded in 5°C steps. Which temperatures do you think the spectra A and B correspond to:

- a) A: 5°C and B: 90°C
- b) A: 90°C and B: 5°C

Name: _____
 ID : _____

9) We want to find the best alignment(s) between the DNA sequences TAGCTTG and AGTTTG. The scoring scheme S is defined as follows: $S(i,j) = 2$ if $i = j$, 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 S_{best} and the number N of optimal alignments are (show your final dynamic programming matrix and the best possible alignment (s) for full credit):

	T	A	G	C	T	T	G
A	0	1	-1	-1	-1	-1	-1
G	-1	0	3	0	0	0	2
T	1	-1	0	3	4	4	2
T	1	1	0	2	5	6	4
T	1	1	0	2	4	7	6
G	-1	1	3	2	2	4	9

- a) $S_{best} = 7, N = 2$
- b) $S_{best} = 9, N = 1$**
- c) $S_{best} = 9, N = 2$
- d) $S_{best} = 7, N = 1$
- e) None of the above

Best alignment:

TAGCTTG

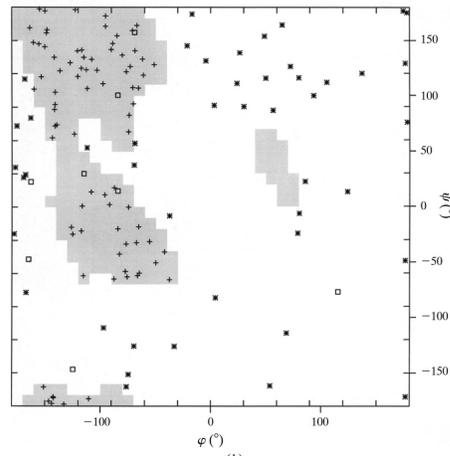
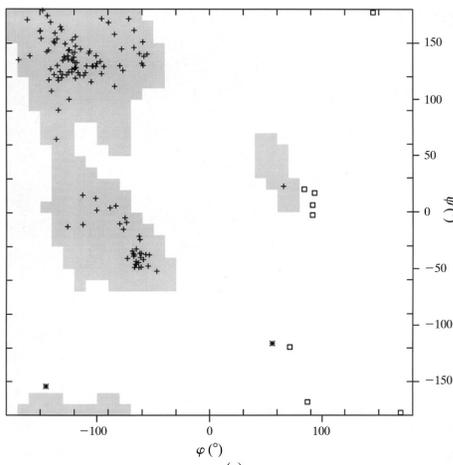
-AGTTTG

10) You are trying to align two protein sequences that share only 18% sequence identity. Which combination of program and substitution matrix is most appropriate for finding a biologically meaningful alignment between domains of those two sequences?

- a) Smith and Waterman / Blosum90
- a) Needleman and Wunsch / Blosum45
- b) BLAST / Blosum90
- c) Needleman and Wunsch / Blosum90
- d) Smith and Waterman / Blosum45**

You need a low BLOSUM matrix; as you want to detect domains, it is best to use Smith and Waterman.

11) The two Ramachandran plots relate to the same protein, cellular retinoic acid binding protein type II. One is based on the experimental X-ray structure of the protein (at 1.8 Angstroms resolution), while the other is from an intentionally bad, backwards-traced model of the same protein. Can you guess which one is which?



Name: _____
ID : _____

- a) (a, left) is the experimental structure, while (b, right) is the bad model
- b) (a, left) is the bad model, while (b, right) is the experimental structure
- c) It is impossible to say from just the Ramachandran plots

Clearly, b has many residues in “forbidden regions”

12) The mRNA sequence 5’-AUG AAA UGC GUU AAC UAA-3’ codes for the small peptide Met-Lys-Cys-Val-Asn. The first G from the 5’ end is mutated to a U. The new mRNA sequence codes for

- a) Ile-Lys-Cys-Val-Asn
- b) Lys-Cys-Val-Asn
- c) Met-Arg
- d) Met-Arg-Asn
- e) None of the above as the new sequence does not contain a START or a STOP codon

If G is mutated to U, we lose the START codon. The new sequence is;

5’-AUUAAAUGCGUUAACUAA-3’

Which contains one short gene, AUG CGU UAA coding for Met – Arg.

13) Which of the following best describes the function and processing of exons and introns?

- a) Both introns and exons are translated into proteins in the cytoplasm.
- b) Introns are spliced out of the mature mRNA, while exons are joined together to be translated.
- c) Exons are removed from the pre-mRNA, while introns are joined together.
- d) Introns are the coding sequences, and exons are the non-coding sequences.

14) The best alignment found between the 2 DNA sequences TAGATTC and TGTCTC is:

TAGATGC

T-GTCAC

The scoring scheme S is defined as follows: $S(i,i) = 3X$, $S(i,j) = 2X$ if i and j are different purines, $S(i,j)=X$ if I and j are different pyrimidine, and $S(i,j)=0$ otherwise. There is a constant gap penalty of -1. The score of this alignment is 35. What is the value of X?

- a) $X = 2$
- b) $X = 3$
- c) $X = 4$
- d) $X = 5$
- e) None of the above

Score = $3X + 3X - 1 + X + 2X + 3X = 36$, hence $12X = 36$, $X = 3$.

15) Which of the following is a known limitation of AlphaFold?

- a) It cannot predict structures for proteins with more than 10 amino acids.
- b) It struggles with modeling intrinsically disordered regions and ligand-bound states.
- c) It is slower than traditional experimental methods.
- d) It only works for human proteins.

Name: _____
 ID : _____

Part II (3 problems, total 40 points)

Problem 1 (20 points)

We want to find the best **local** alignment(s) between the protein sequences S1=WCAPT and S2=PTWCA. 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 count). Note that M and G are negative. The partial dynamic programming matrix (based on **local alignment**) is shown below:

	W	C	A	P	T
P	0	0	0	9	0
T					19
W					5
C					5
A					15

a) (10 points) Can you find P, M, and G. Explain your work

We fill the first row; since M and G are negative, and we perform a local sequence alignment, the first row is filled with 0, except for P with P, where the score is P+G:

	W	C	A	P	T
P	0	0	0	P+G	0

Therefore $P+G = 9$;

Now we fill the second row:

	W	C	A	P	T
P	0	0	0	P+G	0
T	0	0	0	0	2P+G

Therefore $2P+G = 19$. As $P+G=9$, we get **$P=10$ and $G = -1$** .

Now we fill the third row:

	W	C	A	P	T
P	0	0	0	9	0
T	0	0	0	0	19
W	9	0	0	0	P+M+2G

Therefore $10+M-2=5$, hence **$M = -3$** .

Name: _____
 ID : _____

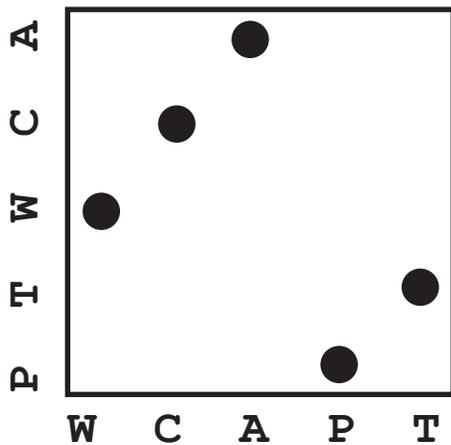
b) (5 points) Complete the local dynamic programming matrix and find the best **local** alignment(s)

	W	C	A	P	T
P	0	0	0	9	0
T	0	0	0	0	19
W	9	0	0	0	5
C	0	19	5	5	5
A	0	5	29	15	15

The best local alignment is:

WCA
 WCA

c) (5 points) Build a simple dotplot between S1 and S2 (assuming a dot in only present for a perfect match). What does it reveal that is not seen from the dynamic programming matrix?



The dotplot reveals a circular permutation in the sequence.

Name: _____
ID : _____

Problem 2 (20 points)

1) (10 points) The following eukaryotic DNA sequence was given to you:

5'-CCCTTAATGCGTATCGCTACCATGCGATACGCATGTAA-3'

You are told that this sequence, or its complementary, codes for one gene.

Find the **longest** "gene" corresponding to this DNA sequence; remember that there are 6 possibilities, i.e. 3 possible reading frames for one strand and 3 possible reading frames for its complementary.

Transcribe this gene into an RNA sequence and then translate it into a protein sequence

The sequence contains 2 ATGs (start codon) and one TAA (stop codon), all in phase. The second ATG is a Met inside the gene:

5'-CCCTTA **ATG** CGT ATC GCT CAC CAT GCG ATA CGC ATG **TAA**-3'

The complementary strand contains one ATG, and one STOP codon in phase:

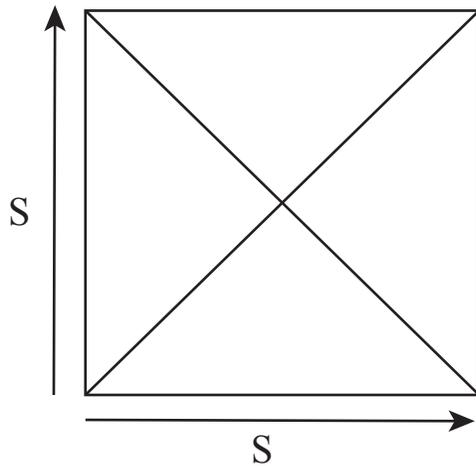
5'-TTAC **ATG** CGT ATC GCA TGG **TGA** GCG ATA CGC ATT AAG GG-3'

The longest gene is 5'-**ATG** CGT ATC GCT CAC CAT GCG ATA CGC ATG **TAA**-3'

The mRNA is: 5'-**AUG** CGU AUC GCU CAC CAU GCG AUA CGC AUG **UAA**-3'

The protein sequence is: **M-R-I-A-H-H-A-I-R-M**

2) (10 points) Draw schematically the dotplot of this protein sequence against itself. What does this dotplot reveal?



The sequence contains an inverted repeat.

Name: _____

ID : _____

Appendix A: Genetic Code

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