*Name:*_____
ID : _____

ECS 129: Structural Bioinformatics

March 15, 2022

Notes:

1) The final exam is open book, open notes.
2) The final is divided into 2 parts, and graded over 100 points, with an additional 5 extra credit 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 (16 questions, each 5 points; total 80 points)

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

1) Two homologous genes:
   a) Would be expected to have very similar sequences in related organisms
   b) Would be expected to be more similar in distantly related organisms than in organisms that are closely related
   c) May have become similar to each other by random mutations
   d) Cannot be found on the same genome
   e) All of these

2) 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 0, and gap penalties are ignored

	A	T	W	C	Y	T
A	10	0	0	0	0	0
T	0	20	10	10	10	20
Y	0	10	20	20	**30**	

A) 20
B) 10
C) 30
D) 40
E) 0

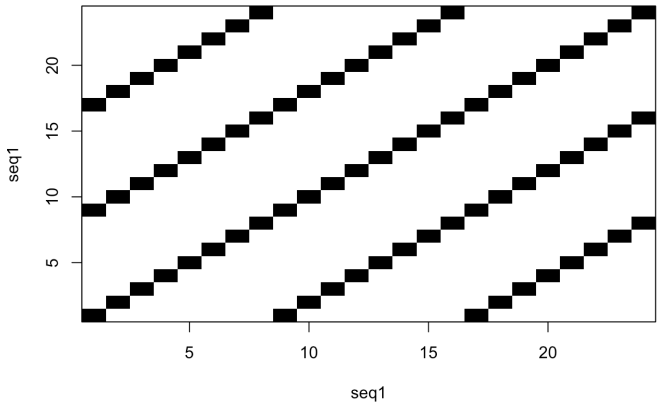
3) Which of the following statements on the Needleman and Wunsch algorithm for pair-wise sequence alignment is most likely true?
   a) The Needleman-Wunsch algorithm is always better in aligning homologous sequences than a multiple sequence alignment
   b) The Needleman-Wunsch algorithm generates one alignment only, namely the single best alignment.
   c) The alignment is optimal (as measured by the alignment score), but there might be many equally optimal pathways/traces through the scoring table.
   d) The Needleman and Wunsch algorithm computes the best local alignment
   e) None of the above

*Name:*_____
ID : _____

4) The current estimate for the number of human genes is 20,000. If a gene contains on average 1,000 nucleotides, genes occupy 20,000,000 bases out of the estimated 3.2 billion nucleotides of the genome. The remaining nucleotides are:

a) Alien DNA that currently remains dormant,
b) Junk, random sequences of DNA,
c) The real genetic information, as genes are only "backup" storage, which is only used in case of emergency
d) A combination of control regions for genes, RNA coding regions, and regions whose purpose is still not known
e) Unusual nucleotides (i.e., different from A, T, G, C) whose origin is still unknown

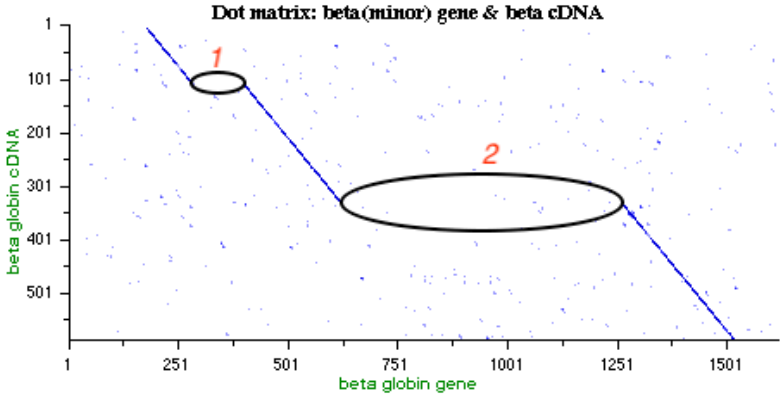
5) The figure below shows the dotplot obtained when comparing a protein sequence Seq1 with itself.



Which of these sequences is most likely to be Seq1:

a) ACDEFGHIACDEFGHIACDEFGHI
b) AAAAAAAAAGGGGGGGGGGG
c) ACDEFGHIACDEEDCAIHGFEDCA
d) ACDACDACDEFGEFGEFGHIHIHI

6) The figure below shows the dot plot between the DNA sequence for the gene beta globin (horizontal axis), and its cDNA, or equivalently the corresponding mRNA.



*Name:*_____
ID : _____

I have labeled two regions on this dotplot, 1 and 2, with ovals. Those regions correspond to:

a) Regions in the gene that do not appear in the cDNA: introns
b) Regions in the gene that do not appear in the cDNA: exons
c) Regions in the cDNA that are not present in the gene: introns
d) Regions in the cDNA that are not present in the gene: exons

7) You are given a single strand S1 of DNA. You are told that: (i) if you mutate four Adenine of S1 to Thymine, the mutated DNA contains twice as many pyrimidines as purines (i.e., the number of pyrimidines is then twice the number of purines), and (ii), S1 contains as many purines as its complementary strand. What is the length of S1?

a) 14
b) 16
c) 18
d) 24
e) Not enough information

8) Which combination of program / substitution matrix will most likely give you the best alignment between two sequences that differ significantly?

a) BLAST / Blosum45
b) Dynamic programming / Blosum45
c) BLAST / Blosum90
d) Dynamic programming / Blosum90
e) BLAST / Blosum10

9) How many possible alignments, with no internal gaps, can you form when you compare a sequence of length 5 with a sequence of length 8? (Note that an alignment must have at least one letter match between the 2 sequences)

a) 6
b) 9
c) 10
d) 12
e) 13

10) We want to find the best alignment(s) between the 2 DNA sequences TATATGCA and ATATC. The scoring scheme S is defined as follows: S(i,i) = 10, S(i,j) = 5 if i and j are both purines, or both pyrimidines, and S(i,j) = 0 otherwise. There is a constant gap penalty of -1. The score Sbest and the number N of optimal alignments are (**show your final dynamic programming matrix for full credit**). Note that a gap at the beginning does not count.

a) Sbest = 50, N = 1
b) Sbest = 50, N = 2
c) Sbest = 49, N = 1
d) Sbest = 49, N = 2
e) None of the above

*Name:*_____
ID : _____

	T	A	T	A	T	G	C
A	0	10	0	10	0	5	0
T	10	0	20	9	20	9	14
A	0	20	9	30	19	25	19
T	10	9	30	19	40	29	34
C	5	10	24	30	34	40	49

11) We want to find the best alignment(s) between the DNA sequences AGTATCT and AGAACT. The scoring scheme S is defined as follows: S(i,j) = 1 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 Sbest and the number N of optimal alignments are (**show your final dynamic programming matrix and the best possible alignment (s) for full credit**):

a) Sbest = 3, N = 2
b) Sbest = 3, N = 1
c) Sbest = 4, N = 2
d) Sbest = 3, N = 3
e) None of the above

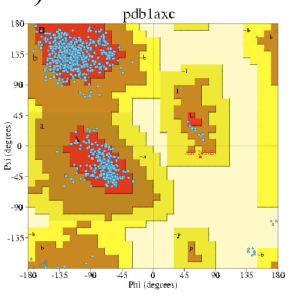
	A	G	T	A	T	C	T
A	1	-1	-1	0	-1	-1	-1
G	-1	2	0	0	0	0	0
A	0	0	2	2	1	1	1
A	-1	0	1	3	2	1	1
C	-1	0	1	1	3	3	2
T	-1	0	2	1	3	3	4

12) You have designed E-Coli such that it can react to light. In the presence of light it generates a white dot, while in the absence of light it generates a black dot. You want to use a bio-film covered with E-Coli as a synthetic camera. Assuming that the bacteria cover uniformly (with no overlap) your bio-film, and that each bacterium is approximately a square with a width of 1 μm, and assuming you want to generate an image with 400 Megapixels (1 Megapixel = 10^6 pixels) what would be a possible size for your bio-film?

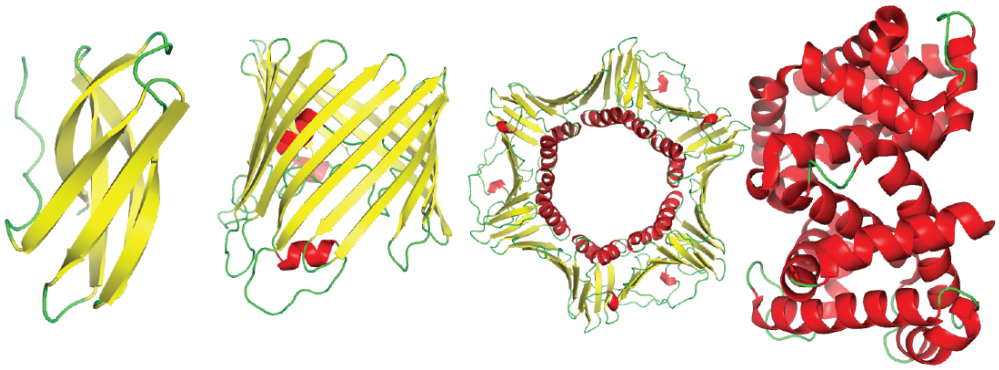
a) 1 cm x 1 cm
b) 1 cm x 2 cm
c) 1 cm x 3 cm
d) 1 cm x 4 cm

e) None of the above

13) The Ramachandran plot of the protein structure 1axc in the PDB databank is given below.



Which of the model of protein structures given below is most likely the corresponding structure:



a) b) c) d)

14) You are given a single strand S1 of DNA of 80 nucleotides. You are told that: (i) S1 contains 3 times as many Thymine as its complementary strand, and (ii), The double stranded DNA formed by S1 and its complementary contains as many A-T base pairs as G-C base pairs. How many Adenine does S1 contains?

a) 20
b) 16
c) 10
d) 40
e) Not enough information available

15) The protein sequence alignment shown below has a total score of 28. 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 gap penalty (assuming that this penalty is constant, i.e., independent of length):

```
GCTGGAAG-GCA-T
GC----AGAGCACT
```

a) -1
b) -2
c) -3
d) -4

e) Undefined (any value would give the same total score)

16) 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 four words is closest to the word "graffe" typed by a user? ***For full credit, write the alignment matrices for all four cases, and write the optimal score next to each word (gaps at the start or at the end do not count).***

a) gaff
b) graft
c) grail
d) giraffe

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

*Name:*_____
ID : _____

Part II (2 problems; total 20 points)
Problem 1 (10 points)

The **template** strand of a sample of double-helical DNA contains the sequence:
(5') TTACGAGATCAT (3')

a) What is the sequence of the corresponding coding sequence (label its 5' and 3' end)?

(5') ATG ATC TCG TAA (3')

b) What is the corresponding coding mRNA sequence (label its 5' and 3' end)?

(5') AUG AUC UCG UAA (3')

c) What is the resulting amino acid sequence (label its Nter and Cter)?

Nter-Met-Ile-Ser-Cter

Problem 2 (10 points)

A probe that was sent to the moon Titan of Saturn brings back a strange bacterium. Analysis of this bacterium shows that its DNA only contains 3 bases, D, N, and K, and that its proteins are made of up to 8 amino acids, which we will label as 1, 2, 3, 4, 5, 6, 7, and 8, for simplicity. You assume that the codon size associated with the unusual genetic code of this bacterium is 2 and it proves to be correct. You are assigned the task to find this genetic code. You test some DNA sequences and check the resulting proteins that you obtained. Your results can be summarized as follows:

*Name:*_____
ID : _____

DNA contains	Protein contains
….KKKK…..	Amino acid 1 only
….NNNN…..	Amino acid 2 only
….DDDD…..	Amino acid 3 only
…KNKNKN….	Amino acid 4 or amino acid 5
…KDKDKD….	Amino acid 6 or amino acid 7
….NDNDND….	Amino acid 8
….KNDKNDKND…	Amino acid 4, 6, or 8

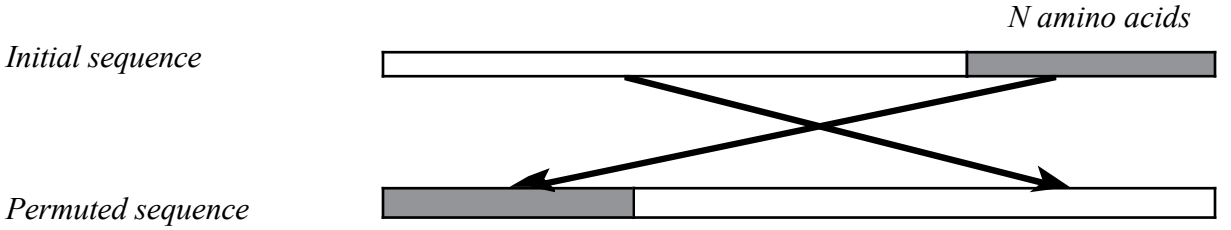
The DNA sequences given are fragments. The reading frame could correspond to the first or the second nucleotide on each of these fragments. Can you fill up the genetic code table below?

		Second base		
		K	N	D
First base	K	1	4	7
	N	5	2	8
	D	6	8	3

Part III extra credit (1 problem; total 5 points)

You have isolated an important gene that regulates the size of a newly found frog from the island of Borneo. You have also been able to find the sequence of the protein encoded by this gene. You suspect that sequences similar to this sequence can be found in other organism, but with circular permutation:

*Name:*_____
ID : _____



In a circular permutation, N amino acids (N can take any value between 1 and M-1, where M is the total length of the protein) at the end of the original sequence will appear at the beginning of the permuted sequence (i.e. before the remaining M-N amino acids).

Propose an efficient strategy for detecting all possible permuted sequences of your frog sequence in a large database of protein sequences.

Use a chimeric sequence that include two copies of the original frog sequence.

*Name:*_____
ID : _____

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