Student name _____

Student number_____

CSc 428/589B
Algorithms in bioinformatics

Midterm exam 1

	Max points	Results
Question 1	**10**	
Question 2	**10**	
Question 3	**10**	
Question 4	**10**	
Question 5	**10**	
Question 6	**10**	
Question 7	**10**	
Question 8	**10**	
Question 9	**10**	
Question 10	**10**	
Total:	**100**	

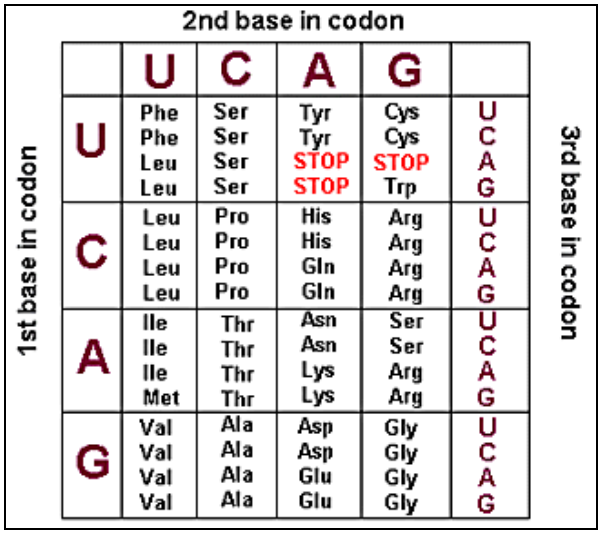
Question 1.

Part I. If we know that 40% of a given *double-stranded DNA* sequence is A, what are the proportions for the rest of the nucleotides?

A. 25% G, 50% T, 25% C
B. 40% T, 5% G, 5% C
C. 50% G, 25% C, 25% T
D. Insufficient information

Part II. If we know that 10% of a given *protein* sequence is amino acid *Leu*, what is the proportion of *Phe* residues in the same sequence

A. 10%
B. 20%
C. Insufficient information
D. 5%



	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	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

Question 2.

Part I. What pair of m-RNA sequences encodes for the most similar pair of proteins

A. UUU CCU UUA and UUA CUU UUC
B. UUU UUG CUA and UUC CUU UUA
C. GUC CCC ACU and CAU AAA GGG

Part II. The protein synthesis starts by translating the following m-RNA sequence from the first codon. What is the length (in the number of aminoacids) of the synthesised protein?_____

CUGGUUAAUUAAUUG

Hint: pay attention to the sequence

Question 3. (Attention: the solution of this question is used for questions 4 and 5)
Draw the suffix tree for string *S=MISSISSIPPI*. Enumerate internal nodes (for example, by *, **, or Roman numbers).
Make sure that there is a separate leaf for each suffix.

1	2	3	4	5	6	7	8	9	10	11	
M	I	S	S	I	S	S	I	P	P	I	

Question 4.

Using the suffix tree for *S=MISSISSIPPI* (see question 3), list all the repetitive substrings of *S* and their corresponding internal nodes.

Question 5.

Using the suffix tree for *S=MISSISSIPPI* (see question 3), find all maximal repeats of *S*. Present the result in form of a truncated suffix tree (the subset of the original tree, which contains only the internal nodes which correspond to the maximal repeats, and the leaves with the positions where these repeats occur)

Question 6.

Consider the following snapshot of a pattern vs text alignment during the KMP search. Let text *T=aaaabaabaa*, and pattern *P=aaaaa*.
The 4 first characters of *P* and *T* were compared (*k*=1), and the character at position *i*=5 of the text does not match the character at position *j*=5 of the pattern.

i	1	2	3	4	**5**	6	7	8	9	10
T	*a*	*a*	*a*	*a*	***b***	*a*	*a*	*b*	*a*	*a*
P	*a*	*a*	*a*	*a*	***a***					

Part I. What is the start position (in *T*) for the next alignment of a pattern vs text?

k=_____

Part II. In what position (*i*) of *T* and in what position *j* of *P* are the characters to be compared next?

i=_____

j=_____

Question 7.

Part I. Compute the edit distance between S_1=*actg* and S_2=*catg* by filling out the DP table for the edit distance.

		c	a	t	g
a					
c					
t					
g					

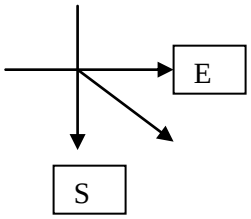
Part II. Find a path which cost corresponds to the value computed in Part I, and show an optimal alignment of S_1 and S_2 according to this path.

Alignment

Question 8.

Solve the following puzzle:
The objective is to go from point A1 to point D4 in the table below, moving through the cells of the table: only East, South and South-East directions are allowed.
While passing through a cell of the table, you collect (positive values) or pay (negative values) the specified number of gold coins.

Find a path that will end up with the maximum number of coins collected.
In your answer, specify the number of collected coins and the path you used to collect them.



	1	2	3	4
A	1	--1	6	5
B	1	3	1	--1
C	--3	5	1	4
D	1	--1	4	5

Question 9. The result after the first iteration of Hirschberg's algorithm is shown in the figure below. Mark all the cells of the dynamic programming table which will be computed in the next iteration.

	a	*c*	*g*	*c*	*g*	*a*
a						
g						
c	2	1	2	1	2	3
g	3	2	1	2	1	2
c						
a						

Question 10.

When searching for the similarity between a query sequence Q of length 50 and the set of 300 sequences of approximately the same length (50) each, we want to find all the sequences in the set which approximately match sequence Q with a maximum of 4 differences (i.e. insertions, deletions, substitutions) . Describe a method to speed up the search by filtration.