Problem Set 1

General Instructions

• The answer sheet for this problem set should be submitted as a PDF file. You may use any word processing software to create the answer sheet. The name of the PDF file to be submitted should follow the following format: [CMSC 197] < Student Number > − < Last Name, First Name > − Problem Set 1.pdf. For example: [CMSC 197] 201536149 - De la Cruz, Juan - Problem Set 1.pdf. Include the answer sheet, programs of items 4-6 in the .zip file detailed next.

• This problem set will require you to submit a .zip file containing several files (including the answer sheet) as detailed below. The name of the .zip file to be submitted should follow the following format: [CMSC 197] < Student Number > − < Last Name, First Name > − Problem Set 1.zip. For example: [CMSC 197] 201536149 - De la Cruz, Juan - Problem Set 1.zip

• Submission of the problem set answers should be done via google form. Send your answers to https://docs.google.com/forms/d/e/1FAIpQLScZDXDIRL1OvbrHWaOKEklMaqvCsNrsRfD7zo-Tfr0C7UYEdA/viewform

• If you have any questions regarding an item (EXCEPT the answer and solution) in the problem set, do not hesitate to e-mail me to ask them. However, questions regarding this problem set forwarded/received on or after 12:01am of 24 January 2018 will NOT be entertained.

1. Given a double-stranded DNA with an ATGGCA sequence in one of its strands. How many hydrogen bonds are present in this six-base pair sequence

2. A given segment of double-stranded DNA consists of 100 base pairs. If strand 1 contains 23 A's and 36 C's while strand 2 possesses 21 A's.

    a. How many G's are there in strand 1?

    b. How many T's are there in strand 2?

    c. How many U's are there in strand 1?

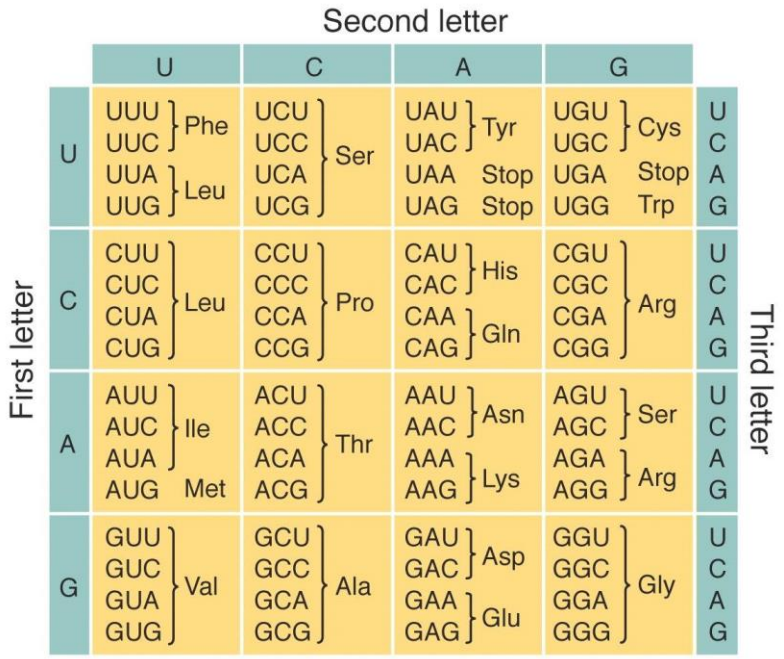
3. The table below gives the DNA sequence for the beta chain of the hemoglobin gene oriented in the 5' to 3' direction. Hemoglobin is an iron-containing protein that transports oxygen in the blood of vertebrates.

    a. Given the nucleotide sequence of the gene, provide the DNA complement by filling out row 3.

    b. Using the complementary DNA as template, determine the mRNA that will be transcribed by filling out row 4.

    c. Determine the amino acid sequence translated based on the mRNA template by filling out the last row. Use the Genetic code on Table 1 as your guide.

Table 1. The Genetic Code



Codon	1	2	3	4	5	6	7	8	9	10	11	12
DNA	**ATG**	GTG	CAC	CTG	ACT	CCT	GAG	GAG	AAG	TCT	GCG	GTT
Complementary DNA	TAC	CAC	GTG	GAC	TGA	GGA	CTC	CTC	TTC	AGA	CGC	CAA
mRNA	AUG	GUG	CAC	CUG	ACU	CCU	GAG	GAG	AAG	UCU	GCG	GUU
Protein	Met	Val	His	Leu	Thr	Pro	Glu	Glu	Lys	Ser	Ala	Val

Codon	13	14	15	16	17	18	19	20	21	22	23	24
DNA	ACT	GCC	CTG	TGG	GGC	AAG	GTG	AAC	GTG	GAT	GAA	GTT
Complementary DNA	TGA	CGG	GAC	ACC	CCG	TTC	CAC	TTG	CAC	CTA	CTT	CAA
mRNA	ACU	GCC	CUG	UGG	GGC	AAG	GUG	AAC	GUG	GAU	GAA	GUU
Protein	Thr	Ala	Leu	Trp	Gly	Lys	Val	Asn	Val	Asp	Glu	Val

Codon	25	26	27	28	29	30	31	32	33	34	35	36
DNA	GGT	GGT	GAG	GCC	CTG	GGC	AGG	CTG	CTG	GTG	GTC	TAC
Complementary DNA	CCA	CCA	CTC	CGG	GAC	CCG	TCC	GAC	GAC	CAC	CAG	ATG
mRNA	GGU	GGU	GAG	GCC	CUG	GGC	AGG	CUG	CUG	GUG	GUC	UAC
Protein	Gly	Gly	Glu	Ala	Leu	Gly	Arg	Leu	Leu	Val	Val	Leu

Codon	37	38	39	40	41	42	43	44	45	46	47	48
DNA	CCT	TGG	ACC	CAG	AGG	TTC	TTT	GAG	TCC	TTT	GGG	GAT
Complementary DNA	GGA	ACC	TGG	GUC	TCC	AAG	AAA	CTC	AGG	AAA	CCC	CTA
mRNA	CCU	UGG	ACC	CAG	AGG	UUC	UUU	GAG	UCC	UUU	GGG	GAU
Protein	Pro	Trp	Thr	Gln	Arg	Phe	Phe	Glu	Ser	Phe	Gly	Asp

Codon	49	50	51	52	53	54	55	56	57	58	59	60
DNA	CTG	TCC	ACT	CCT	GAT	GCA	GTT	ATG	GGC	AAC	CCT	AAG
Complementary DNA	GAC	AGG	TGA	GGA	CTA	CGT	CAA	TAC	CCG	TTG	GGA	TTC
mRNA	CUG	UCC	ACU	CCU	GAU	GCA	CUU	AUG	GGC	AAC	CCU	AAG
Protein	Leu	Ser	Thr	Pro	Asp	Ala	Leu	Met	Gly	Thr	Pro	Lys

Codon	61	62	63	64	65	66	67	68	69	70	71	72
DNA	GTG	AAG	GCT	CAT	GGC	AAG	AAA	GTG	CTC	GGT	GCC	TTT
Complementary DNA	CAC	TTC	CGA	GTA	CCG	TTC	TTT	CAC	GAG	CCA	CGG	AAA
mRNA	GUG	AAG	GCU	CAU	GGC	AAG	AAA	GUG	CUC	GGU	GCC	UUU
Protein	Val	Lys	Ala	His	Gly	Lys	Lys	Val	Leu	Gly	Ala	Phe

Codon	73	74	75	76	77	78	79	80	81	82	83	84
DNA	AGT	GAT	GGC	CTG	GCT	CAC	CTG	GAC	AAC	CTC	AAG	GGC
Complementary DNA	TCA	TCA	CCG	GAC	CGA	GTG	GAC	CTG	TTG	GAG	TTC	CCG
mRNA	AGU	GAU	GGC	CUG	GCU	CAC	CUG	GAC	AAC	CUC	AAG	GGC
Protein	Ser	Asp	Gly	Leu	Ala	His	Pro	Asp	Asn	Leu	Lys	Gly

Codon	85	86	87	88	89	90	91	92	93	94	95	96
DNA	ACC	TTT	GCC	ACA	CTG	AGT	GAG	CTG	CAC	TGT	GAC	AAG
Complementary DNA	TGG	AAA	CGG	TGT	GAC	TCA	CTC	GAC	GTG	ACA	CTG	TTC
mRNA	ACC	UUU	GCC	ACA	CUG	AGU	GAG	CUG	CAC	UGU	GAC	UUC
Protein	Thr	Phe	Ala	Thr	Leu	Ser	Glu	Leu	His	Cys	Asp	Phe

Codon	97	98	99	100	101	102	103	104	105	106	107	108
DNA	CTG	CAC	GTG	GAT	CCT	GAG	AAC	TTC	AGG	CTC	CTG	GGC
Complementary DNA	GAC	GTG	CAC	CTA	GGA	CTC	TTG	AAG	TCC	GAG	GAC	CCG
mRNA	CUG	CAC	GUG	GAU	CCU	GAG	AAC	UUC	AGG	CUC	CUG	GGC
Protein	Leu	His	Val	Asp	Pro	Glu	Asn	Phe	Arg	Leu	Leu	Gly

Codon	109	110	111	112	113	114	115	116	117	118	119	120
DNA	AAC	GTG	CTG	GTC	TGT	GTG	CTG	GCC	CAT	CAC	TTT	GGC
Complementary DNA	TTG	CAC	GAC	CAG	ACA	CAC	GAC	CGG	GTA	GTG	AAA	CCG
mRNA	AAC	GUG	CUG	GUC	UGU	GUG	CUG	GCC	CAU	CAC	UUU	GCC
Protein	Asn	Val	Leu	Val	Cys	Val	Leu	Ala	His	His	Phe	Ala

Codon	121	122	123	124	125	126	127	128	129	130	131	132
DNA	TTT	GGC	AAA	GAA	TTC	ACC	CCA	CCA	GTG	CAG	GCT	GCC
Complementary DNA	AAA	CCG	TTT	CTT	AAG	TGG	GGT	GGT	CAC	GTC	CGA	CGG
mRNA	UUU	GGC	AAA	GAA	UUC	ACC	CCA	CCA	GUG	CAG	GCU	GCC
Protein	Phe	Gly	Lys	Glu	Phe	Asn	Pro	Pro	Val	Gin	Ala	Ala

Codon	133	134	135	136	137	138	139	140	141	142	143	144
DNA	TAT	CAG	AAA	GTG	GTG	GCT	GGT	GTG	GCT	AAT	GCC	CTG
Complementary DNA	ATA	GTC	TTT	CAC	CAC	CGA	CCA	CAC	CGA	TTA	CGG	GAC
mRNA	UAU	CAG	AAA	GUG	GUG	GCU	CCU	GUG	GCU	AAU	GCC	CUG
Protein	Tyr	Gin	Lys	Val	Val	Ala	Pro	Val	Ala	Ile	Ala	Leu

Codon	145	146	147	148	149	150
DNA	GCC	CAC	AAG	TAT	CAC	TAA
Complementary DNA	AGG	GTG	TTC	ATA	GTG	ATT
mRNA	GCC	CAC	AGG	UAU	CAC	UAA
Protein	Ala	His	Lys	Ile	His	Stop

For problems 4-6, write a program for each problem. You may use any programming language to solve each problem.

4. Counting DNA Nucleotides

A **string** is simply an ordered collection of symbols selected from some **alphabet** and formed into a word; the **length** of a string is the number of symbols that it contains.

An example of a length 21 **DNA string** (whose alphabet contains the symbols 'A', 'C', 'G', and 'T') is "ATGCTTCAGAAAGGTCTTACG."

Given: A DNA string ss of length at most 1000 nt.
Return: Four integers (separated by spaces) counting the respective number of times that the symbols 'A', 'C', 'G', and 'T' occur in ss.

Sample Dataset

AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC

Sample Output

A C G T

20 12 17 21

5. Transcribing DNA into RNA

An **RNA string** is a string formed from the alphabet containing 'A', 'C', 'G', and 'U'. Given a DNA string tt corresponding to a coding strand, its transcribed RNA string uu is formed by replacing all occurrences of 'T' in tt with 'U' in uu.

Given: A DNA string tt having length at most 1000 nt.
Return: The transcribed RNA string of tt.

Sample Dataset

GATGGAACTTGACTACGTAAATT

Sample Output

GAUGGAACUUGACUACGUAAAUU

6. Translating RNA into Protein

The 20 commonly occurring amino acids are abbreviated by using 20 letters from the English alphabet (all letters except for B, J, O, U, X, and Z). **Protein strings** are constructed from these 20 symbols. Henceforth, the term **genetic string** will incorporate protein strings along with DNA strings and RNA strings.

The **RNA codon table** dictates the details regarding the encoding of specific codons into the amino acid alphabet.

Given: An RNA string ss corresponding to a strand of mRNA (of length at most 10 kbp).
Return: The protein string encoded by ss.

Sample Dataset

AUGGCCAUGGCGCCCAGAACUGAGAUCAAUAGUACCCGUAUUAACGGGUGA

Sample Output

MAMAPRTEINSTRING