Programming Assignment 5

Introduction

The aim of this assignment is to introduce you to C basics and get you familiar with Multi-D arrays. You are expected to implement a protein sequence analyser for bioinformatics.

Your DNA is made up of a series of four different nucleotides. These are adenine (A), thymine (T), guanine (G), and cytosine (C). Thymine (T) becomes urasil (U) when it comes to RNA as shown in Figure 2. Hidden within your genome lies the "triplet code," a series of three nucleotides that determine a single amino acid. RNA (or, more specifically messenger RNA - mRNA) serves as a copy of your chromosomal DNA and specifies the sequence of amino acids in proteins.

Proteins are large, complex molecules that play many critical roles in the body. They do most of the work in cells and are required for the structure, function, and regulation of the bodys tissues and organs. Proteins are made up of hundreds or thousands of amino acids, which are attached to one another in long chains. It had long been known that only 20 amino acids occur in naturally derived proteins.

Protein synthesis is one of the most fundamental biological processes by which individual cells build their specific proteins. It is performed in two steps. The first step is called transcription. During transcription, the information encoded in the DNA is copied to a RNA molecule as one strand of the DNA double helix is used as a template. The RNA molecule is sent to the cytoplasm, which helps to bring all components required for the actual protein synthesis together amino acids, transport RNAs, ribosomes, etc. In the cytoplasm the protein polymers are actually synthesized through chemical reactions that is why the process is known as protein synthesis or even more precisely protein biosynthesis. The second step of protein synthesis is mRNA Translation (or just Translation). The mRNA Translation step follows right after DNA Transcription (or just Transcription). The production of proteins happens during the second step of protein synthesis process the Translation. Sometimes protein synthesis process is referred only to Translation step, because no actual protein synthesis happens during the Transcription. However transcription is responsible for moving the genetic instructions from the nucleus to the cytoplasm, where the DNA/RNA code is translated by the ribosomes to a polypeptide sequence, which will later be folded into a protein.

The nucleotide triplet that encodes an amino acid is called a codon. Each group of three nucleotides encodes one amino acid as shown in Figure 1.

Examination of the full table of codons in Figure 1 enables one to immediately determine whether the "extra" codons are associated with redundancy or dead-end codes. Note that

both possibilities occur in the code. There are only a few instances in which one codon codes for one amino acid, such as the codon for tryptophan. Note also that the codon for the amino acid methionine (AUG) acts as the start signal for protein synthesis in a mRNA. Moreover, the genetic code also includes stop codons, which do not code for any amino acid. The stop codons serve as termination of the protein synthesis.

	U		C		A		G	
U	UUU	Fenilalanin(F)	UCU	Serin (S)	UAU	Triozin (Y)	UGU	Sistein (C)
	UUC	Fenilalanin(F)	UCC	Serin (S)	UAC	Triozin (Y)	UGC	Sistein (C)
	UUA	Lösin (L)	UCA	Serin (S)	UAA	Stop codon	UGA	Stop codon
	UUG	Lösin (L)	UCG	Serin (S)	UAG	Stop codon	UGG	Triptofan (W)
C	CUU	Lösin (L)	CCU	Prolin (P)	CAU	Histidin (H)	CGU	Arjinin (R)
	CUC	Lösin (L)	CCC	Prolin (P)	CAC	Histidin (H)	CGC	Arjinin (R)
	CUA	Lösin (L)	CCA	Prolin (P)	CAA	Glutamin (Q)	CGA	Arjinin (R)
	CUG	Lösin (L)	CCG	Prolin (P)	CAG	Glutamin (Q)	CGG	Arjinin (R)
A	AUU	İzolosin (I)	ACU	Treonin (T)	AAU	Asparajin (N)	AGU	Serin (S)
	AUC	İzolosin (I)	ACC	Treonin (T)	AAC	Asparajin (N)	AGC	Serin (S)
	AUA	İzolosin (I)	ACA	Treonin (T)	AAA	Lizin (K)	AGA	Arjinin (R)
	AUG	Metionin (M)	ACG	Treonin (T)	AAG	Lizin (K)	AGG	Arjinin (R)
G	GUU	Valin (V)	GCU	Alanin (A)	GAU	Aspartik acid (D)	GGU	Glisin (G)
	GUC	Valin (V)	GCC	Alanin (A)	GAC	Aspartik acid (D)	GGC	Glisin (G)
	GUA	Valin (V)	GCA	Alanin (A)	GAA	Glutamic acid (E)	GGA	Glisin (G)
	GUG	Valin (V)	GCG	Alanin (A)	GAG	Glutamic acid (E)	GGG	Glisin (G)

Figure 1: The amino acids specified by each mRNA codon. Multiple codons can code for the same amino acid.
AUG is an initiation (start) codon; UAA, UAG, and UGA are termination (stop) codons.

Protein synthesis is summarized in Figure 2. Part of the DNA is transcribed and mRNA is produced. Notice that the mRNA sequence starts with Metionin and ends with a stop codon. Then, the transcript is translated in to an amino acids sequence, which encodes a protein (a hypothetical one for this example).

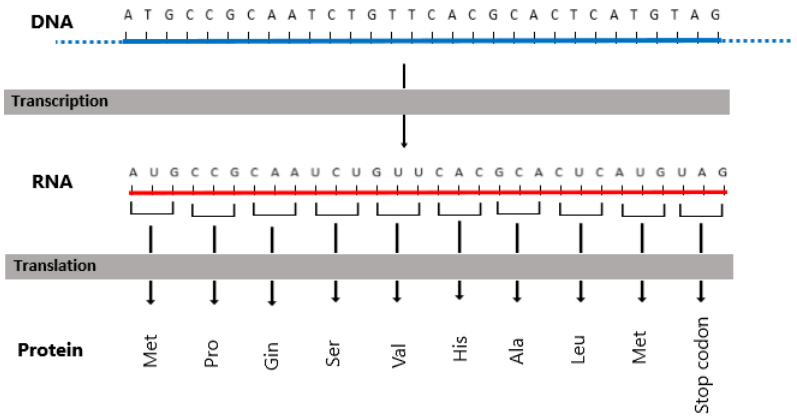

Figure 2: Steps of the protein synthesis

The size of the proteins found in real life is very long. Therefore, hypothetical example proteins called myProteinA, myProteinB, myProteinC, myProteinD and myProteinE are created to be used in this assignment, which are seen in Table 1. Assume that these proteins are the ones known by the scientific community. The number of nucleotides in the example proteins is fixed and set to 30. Assume that they are the only known proteins by the scientific community as of today.

Table 1: Nucleotide sequences of example proteins

Protein	**Nucleotide Sequence**
myProteinA	AUGGUGGCGGAGGGGACGAAGAGGAUCUAA
myProteinB	AUGGGAGAAGCAGUAAGAAAAACAAUAUAG
myProteinC	AUGUUUUCCUAUUGCCUGCCACAACGCUGA
myProteinD	AUGUUCUUGGUCCCUACUUACGAUCAUUAA
myProteinE	AUGUUUUCCUAUUGCCUGCCAAAACGCUGA

Problem

In this assignment, you will be given a nucleotide sequence and asked *i)* if that sequence is one of your known proteins (myProteinA-E) or *ii)* to determine if that sequence might be a protein -if not a known one-.

To test your solution, you will be given a protein sequence in an input text file (Input.txt). Firstly, you are expected to find whether the given sequence in the input file is an alternative of is an alternative of one of the proteins listed in Table 1. Remember, due to redundancy in aminoacid formation, a protein might be represented with several different nucleotide sequences. If this is the case, print "myProteinX is identified in sequence."(X should be either A, B, C, D or E) and list the amino acids that make up this protein as shown in Figure 3. If not, a warning message should be given as "It is not a known protein." and check if it might be a possible unknown protein. If yes, print "It is probably a new protein". If the given sequence does not conform with the conditions to form a protein out of a nucleotide sequence print "It is not a protein!". The input file (input.txt) should be given as an argument to your program.

Sample Inputs / Outputs:

The outputs of this assignment according to various input files are given in Figure 3, Figure 4 and Figure 5.

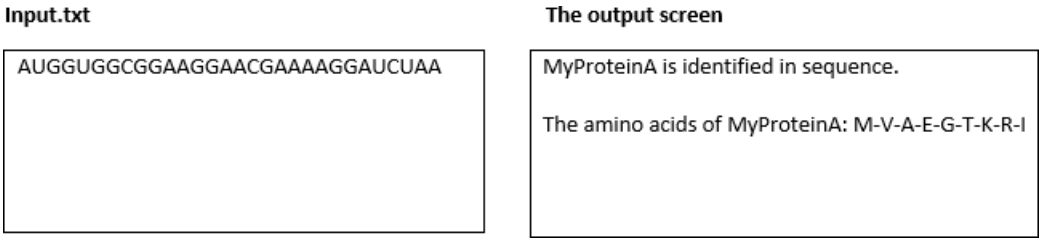

Figure 3: If the sequence given in the text file is an alternative of myProteinA-E

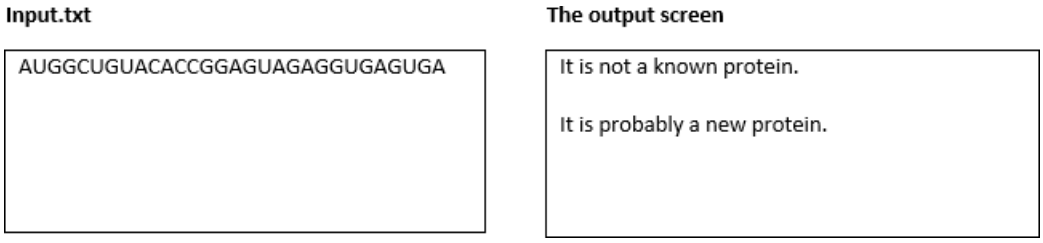

Figure 4: If the sequence given in the text file except myProteinA-E

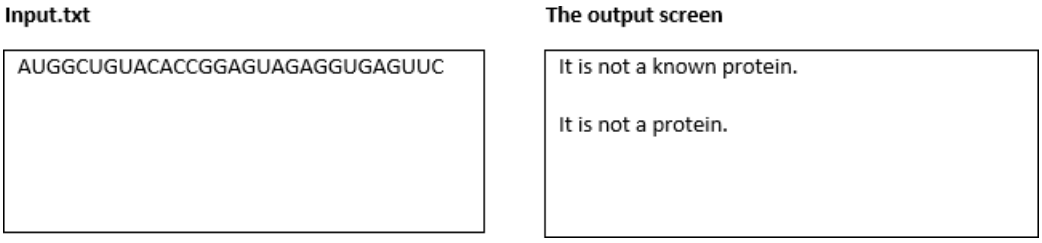

Figure 5: If the sequence given in the text file is not a protein