Middle East Technical University
Department of Computer Engineering

CENG 140

Spring 2024

Take Home Exam 1

REGULATIONS

Due date: 15 April 2024, 23:59, Monday *(Not subject to postpone)*

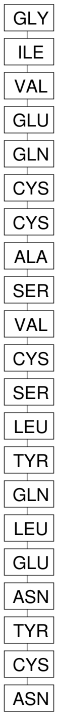
Submission: Electronically. You will be submitting your program source code written in a file which you will name as `the1.c` through the ODTUCLASS system. Resubmission is allowed (till the last moment of the due date), The last will replace the previous.

Team: There is **no** teaming up. The take home exam has to be done/turned in individually.

Cheating: **This is an exam:** all parts involved (source(s) and receiver(s)) get zero+both parts will be subject to disciplinary action.

BACKGROUND

The life form on our planet is primarily based on a molecule that we call DNA. It carries within its structure the heredity information that determines the structure of proteins and the instructions that direct cells to grow and divide. So are the messages that bring about the differentiation of fertilized eggs into the multitude of specialized cells that are necessary for the successful functioning of higher plants and animals.



In DNA molecules, *nucleotides* are linked together to form long chains by bonds. The order and sequence of this chain is the information content of the DNA. Each DNA molecule contains many subparts that we call *gene*. A *gene* is the smallest functional unit and serves as a 'program' that synthesizes a protein (a chain of amino-acids).
On the left you see a part of such a synthesed protein: The A-chain of an *insulin*.
Just for your information the names of the building blocks of proteins namely the Amino acids are:

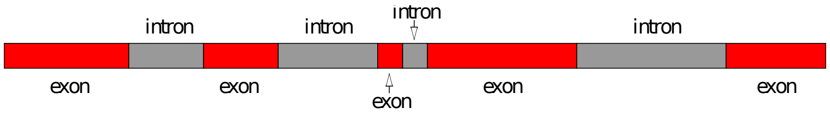
Amino Acid	**Three letter symbol**	**One letter symbol**	**Amino Acid**	**Three letter symbol**	**One letter symbol**
Glycine	GLY	g	Lysine	LYS	k
Alanine	ALA	a	Arginine	ARG	r
Valine	VAL	v	Asparagine	ASN	n
Isoleucine	ILE	i	Glutamine	GLN	q
Leucine	LEU	l	Cysteine	CYS	c
Serine	SER	s	Methionine	MET	m
Threonine	THR	t	Tryptophan	TRP	w
Proline	PRO	p	Phenilalanine	PHE	f
Asparic acid	ASP	d	Tyrosine	TYR	y
Glutamic acid	GLU	e	Histidine	HIS	h

A *nucleotide* is one of the the four molecules:

A	:	Adenine
C	:	Cytosine
G	:	Guanine
T	:	Thymine

In a gene, each group of three successive nucleotides is called a *codon*. Each codon is is responsible of synthesizing an amino acid (one of the table on the previous page).

Although a gene is responsible of synthesis of a protein, it has parts which are ‘garbage’ and parts which are ‘functional’. These parts are called *introns* and *exons*, respectively. So, it is actually the exon zones which produces the protein and introns can practically be omitted from the sequence.



The mechanism of where an exon stops/starts and intron starts/stops is not known to us as far as this THE is concerned. When it is time to produce a protein, a kind of a mask is generated from the gene. This mask is called the messenger RNA (mRNA)which also consists of only four kind of nucleotides. These are

A	:	Adenine
C	:	Cytosine
G	:	Guanine
U	:	Urasil

It is only the useful parts that are masked, namely the exons. The introns are simply ignored (skipped over).

This mask generation is called *transcription*. Each nucleotide in the gene is represented by a one-to-one correspondence. Here is the table of transcription:

Original DNA nucleotide	Transcripted into mRNA nucleotide
A	U
C	G
G	C
T	A

After the transcription is completed, the codons of the mRNA are used to manufacture the protein. Since there are 4 nucleotides, as you might have realized already, there are $4 \times 4 \times 4$ possibilities for them to group in four. That means there are 64 codons. But as we know they synthesize only 20 amino acids and some codons stop the process. So, evidently, some codons must synthesize the same amino-acid. That is true, and is called *degeneracy*.

Here is the table of which mRNA codon is synthesizing which amino acid.

First position	Second position				Third position
	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	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

PROBLEM

You will be given a list of nucleotides which represent a gene. You know also that there is exactly 3 exons and 2 introns. The gene starts with an exon and ends with an exon. You don't know where the intermediate exon is located. You also know that none of the introns are empty. Since exons are sequences of codons (three nucleotides together) they are a multiple of 3. This is not so for introns, since they are garbage their length has not to be a multiple of 3.

The input will be read from the terminal. Whitespaces are possible. So the input can be scattered over lines (as it is in the following example) The input terminates with a period. Here is an example of such a gene input:

```
TCTGCAGCAGAGGGGCCGTC
GGCAGAAGGAGGGCTCGGGC
AGGCTCTGCGACTCGTAGGC
ACCAGGCGTGAGACCTGTAG
CCCCCGATCACCATGTACAG
CTTCATGGGTGGTGGCCTGT
TCTGTGCCTGGGTGGGGACC
ATCCTCCTGGTGGTGGCCAT
GGCAACAGACGGGGCCAAGG
ACACCTGTATTCCAGATGGA
GAACTCTGCGGCTCAAAGAG
GGAAAGGGAGCAACCCAAGG
TCACTCAGCGGAGGCTGACT
CCTGGTCCTAGGCTGGAAGG
AGGAAGAATAGGGCCCATGG
GAGGGAGCTGAGAAGACT.
```

The next input is a protein (as a sequence of amino acid one-letter symbols). This input input will also be read from the terminal. Whitespaces are possible. Similar to the gene input the input can be scattered over lines. The input terminates with a period (this period serves only to let you know that the DNA sequence has terminated).

Here is the protein input example:

```
rrrlpgsrlppepvrdaelvvhvevpttgqdtdpplvggpppvplspptedqdptflllipgtlprlf.
```

In this input the period represents also the ”stop” (so it has a corresponding coding in the exon part). You are expected to find the introns’ start and end positions (counting starts at zero). You will output four integers:

$intron_{1_{start}} \quad intron_{1_{end}} \quad intrn_{2_{start}} \quad intron_{2_{end}}$

If you discover that this protein is not the product of the gene given, then you print ”`NONE`” (without the quotes).
You are assured that each of the intron fields are not empty.
For the example above the expected input is:

```
54 84 169 248
```

SPECIFICATIONS

- **There is no need for string** reading/manipulation. You are expected to read the input character by character and store the ascii codes in `char` arrays. All the processing is expected to be done over these arrays.
- There will be no erroneous input.
- The maximal DNA length is 3000 nucleotides.
- The minimal and maximal length of the protein input is 10 and 2900 amino-acids, respectively.
- If there are more then one solution you can provide any one of them.

Mini Summary for the Perplexed

- Gene is a part of a DNA molecule that codes a single protein. A gene is made of a sequence of nucleotides.
- A nucleotide is one of A,C,G,T in the gene (In the mRNA it is one of A,C,G,U).
- In a gene there are garbage areas that are called introns. The useful areas (used in protein synthesis) are called exons.
- GENE → $\boxed{\text{intron removal \& transcription}}$ → mRNA → $\boxed{\text{protein synthesis}}$ → PROTEIN.
- A Protein is a sequence of amino acids.
- There are 20 different amino acids.
- In the exon (useful) parts of a gene three successive nucleotides are called a codon. A codon is coding a single amino acid or a stop (the protein synthesis) information.
- There are 64 possible codons but only 20 amino acids. So some codons code the same type of amino acid.