Middle East Technical University
Department of Computer Engineering

CENG 111

Fall '97

Worksheet 2

REGULATIONS

Due date: 15 December, Monday *(Not subject to postpone)*

Submission: Electronically. You will be submitting your program source code through a file which you will name as `hw3.sh` by issuing the following single line at the Unix prompt.

```
submit111 hw3.sh
```

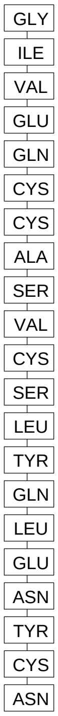
The last moment for submission is 24:00 of the due date.

Team: There is **no** teaming up. The homework has to be done and turned in individually.

Cheating: All parties involved (source(s) and receiver(s)) get zero.

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:

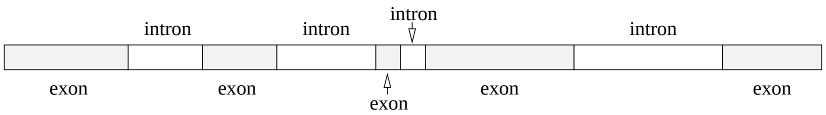
Glycine	GLY	Lysine	LYS
Alanine	ALA	Arginine	ARG
Valine	VAL	Asparagine	ASN
Isoleucine	ILE	Glutamine	GLN
Leucine	LEU	Cysteine	CYS
Serine	SER	Methionine	MET
Threonine	THR	Tryptophan	TRP
Proline	PRO	Phenilalanine	PHE
Asparic acid	ASP	Tyrosine	TYR
Glutamic acid	GLU	Histidine	HIS

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.



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
T	: Urasil

It is only the useful parts that are masked, namely the exons. The introns are simply ignored. 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 the both of the introns are non-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. Here is an example of such a gene:

```
(T C T G C A G C A G A G G G G C C G T C G G C A G A A G G A G
 G G C T C G G G C A G G C T C T G C G A C T C G T A G G C A C
 C A G G C G T G A G A C C T G T A G C C C C C G A T C A C C A
 T G T A C A G C T T C A T G G G T G G T G G C C T G T T C T G
 T G C C T G G G T G G G G A C C A T C C T C C T G G T G G T G
 G C C A T G G C A A C A G A C G G G G C C A A G G A C A C C T
 G T A T T C C A G A T G G A G A A C T C T G C G G C T C A A A
 G A G G G A A A G G G A G C A A C C C A A G G T C A C T C A G
 C G G A G G C T G A C T C C T G G T C C T A G G C T G G A A G
 G A G G A A G A A T A G G G C C C A T G G G A G G G A G C T G
 A G A A G A C T)
```

You will also be given a protein (as a list of amino acids). Which is something like:

```
(ARG ARG ARG LEU PRO GLY SER ARG LEU PRO PRO GLU PRO VAL ARG ASP ALA GLU
 LEU VAL VAL HIS VAL GLU VAL PRO THR THR GLY GLN ASP THR ASP PRO PRO LEU
 VAL GLY GLY PRO PRO PRO VAL PRO LEU SER PRO PRO THR GLU ASP GLN ASP PRO
 THR PHE LEU LEU LEU ILE PRO GLY THR LEU PRO ARG LEU PHE stop)
```

You are expected to find the introns' start and end positions. The answer will be given a list of two lists. Where each inner list is a pair of two positions.

$$((intron_{1_{start}} \quad intron_{1_{end}}) \quad (intron_{2_{start}} \quad intron_{2_{end}}))$$

If you discover that this protein is not the product of the gene given, then you shall return an empty list.

Name your function that takes the gene list as its first argument and the protein list as the second, as `locate_intron`

The answer for the above example shall be:

```
((55 85) (170 249))
```