**Translation of mRNA to Protein**

A MINI PROJECT REPORT

Submitted by

BL.EN.U4AIE21006 Akhilesh P

BL.EN.U4AIE21008 Amal Krishna

BL.EN.U4AIE21012 Ahn Karthik

**In partial fulfilment for the award of the degree Of**

**BACHELOR IN TECHNOLOGY**

**IN**

COMPUTER SCIENCE AND ENGINEERING

(ARTIFICIAL INTELLIGENCE)



AMRITA SCHOOL OF ENGINEERING, BANGALORE

AMRITA VISHWA VIDYAPEETHAM

BANGALORE – 560035

**ABSTRACT**

This project aims on developing a system which bio mimic the functionality of Ribosomes and tRNAs by converting mRNA sequence to a chain of amino acids (Protein).

A program which runs in python is developed. The code makes full use of some built in functions in python like reading from files, dictionary data structures -etc.

**TABLE OF CONTENTS**

Pg No.

1.Abstract 2

2.Contents 3

3.Tranlation 4

3.1 Initiation 5

3.2 Elongation 6

3.3 Termination 7

4. Role of tRNA 8

5. Role of Ribosome 9

6. Genetic Code 10

7. Python Code 11

8. Algorithm 13

9. Result and analysis 14

10. Conclusion 15

**TRANSLATION**

In molecular biology and genetics, translation is the process in which ribosomes in the cytoplasm or endoplasmic reticulum synthesize proteins after the process of transcription of DNA to RNA in the cell’s nucleus. The entire process is called gene expression.

In translation, messenger RNA (Mrna) is decoded in a ribosome, outside the nucleus, to produce a specific amino acid chain, or polypeptide. The polypeptide later folds into an active protein and performs its functions in the cell. The ribosome facilitates decoding by inducing the binding of complementary Trna anticodon sequences to Mrna codons. The tRNAs carry specific amino acids that are chained together into a polypeptide as the Mrna passes through and is “read” by the ribosome.

In prokaryotes (bacteria and archaea), translation occurs in the cytosol, where the large and small subunits of the ribosome bind to the Mrna. In eukaryotes, translation occurs in the cytoplasm or across the membrane of the endoplasmic reticulum in a process called co-translational translocation. In co-translational translocation, the entire ribosome/Mrna complex binds to the outer membrane of the rough endoplasmic reticulum (ER) and the new protein are synthesized and released into the ER; the newly created polypeptide can be stored inside the ER for future vesicle transport and secretion outside the cell, or immediately secreted.

Many types of transcribed RNA, such as transfer RNA, ribosomal RNA, and small nuclear RNA, do not undergo translation into proteins.

A number of antibiotics act by inhibiting translation. These include anisomycin, cycloheximide, chloramphenicol, tetracycline, streptomycin, erythromycin, and puromycin. Prokaryotic ribosomes have a different structure from that of eukaryotic ribosomes, and thus antibiotics can specifically target bacterial infections without any harm to a eukaryotic host’s cells.

**INITIATION :**

The ribosome assembles around the target Mrna. The first Trna is attached at the start codon (mostly AUG which codes for methionine).

Initiation factors are proteins that bind to the small subunit of the ribosome during the initiation of translation, a part of protein biosynthesis.

Initiation factors can interact with repressors to slow down or prevent translation. They have the ability to interact with activators to help them start or increase the rate of translation. In bacteria, they are simply called Ifs (i.e.., IF1, IF2, & IF3) and in eukaryotes they are known as eIFs (i.e.., Eif1, Eif2, Eif3). Translation initiation is sometimes described as three step process by which initiation factors help to carry out. First, the Trna carrying a methionine amino acid binds to the small ribosome, then binds to the Mrna, and finally joining together with the large ribosome. The initiation factors that help with this process each have different roles and structures.

Initiation on most eukaryotic mRNAs involves scanning by ribosomal 43S preinitiation complexes on the 5′ untranslated region (UTR) from the 5′ cap-proximal point of initial attachment to the initiation codon. The mechanism of scanning is incompletely understood, but requires an ‘open’ conformation of the 40S subunit, which is induced by Eif1 and Eif1A, is coupled to the activities of the DEAD-box RNA helicase Eif4A, its cofactor Eif4B and Eif4G, and may involve additional DEAD box- or DexH box-containing proteins such as DexH box protein 29 (DHX29; higher eukaryotes) and DEAD box helicase 1 (Ded1; yeast).

**ELONGATION :**

The last Trna validated by the small ribosomal subunit (accommodation) transfers the amino acid it carries to the large ribosomal subunit which binds it to the one of the precedingly admitted Trna (transpeptidation). The ribosome then moves to the next Mrna codon to continue the process (translocation), creating an amino acid chain.

Elongation factors are a set of proteins that function at the ribosome, during protein synthesis, to facilitate translational elongation from the formation of the first to the last peptide bond of a growing polypeptide. Most common elongation factors in prokaryotes are EF-Tu, EF-Ts, EF-G. Bacteria and eukaryotes use elongation factors that are largely homologous to each other, but with distinct structures and different research nomenclatures.

Elongation is the most rapid step in translation. In bacteria, it proceeds at a rate of 15 to 20 amino acids added per second (about 45-60 nucleotides per second). In eukaryotes the rate is about two amino acids per second (about 6 nucleotides read per second). Elongation factors play a role in orchestrating the events of this process, and in ensuring the high accuracy translation at these speeds.

The ribosome then translocates along the mRNA molecule to the next codon, again using energy yielded from the hydrolysis of GTP. Now, the growing peptide lies at the P site and the A site is open for the binding of the next aminoacyl-Trna, and the cycle continues. The polypeptide chain is built up in the direction from the N terminal (methionine) to the C terminal (the final amino acid).

**TERMINATION :**

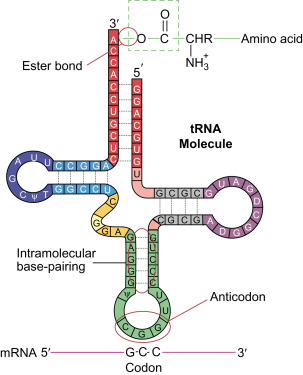
When a stop codon is reached, the ribosome releases the polypeptide. The ribosomal complex remains intact and moves on to the next Mrna to be translated.

Translation termination in eukaryotes occurs in response to a stop codon in the ribosomal A-site and requires two release factors (RFs), Erf1 and Erf3, which bind to the A-site as an Erf1/Erf3/GTP complex with Erf1 responsible for codon recognition. After GTP hydrolysis by Erf3, Erf1 triggers hydrolysis of the polypeptidyl-Trna, releasing the completed protein product. This leaves an 80S ribosome still bound to the Mrna, with deacylated Trna in its P-site and at least Erf1 in its A-site, which needs to be disassembled and released from the Mrna to allow further rounds of translation. The first step in recycling is dissociation of the 60S ribosomal subunit, leaving a 40S/deacylated Trna complex bound to the Mrna. This is mediated by ABCE1, which is a somewhat unusual member of the ATP-binding cassette family of proteins with no membrane-spanning domain but two essential iron–sulphur clusters. Two distinct pathways have been identified for subsequent ejection of the deacylated Trna followed by dissociation of the 40S subunit from the Mrna, one executed by a subset of the canonical initiation factors (which therefore starts the process of preparing the 40S subunit for the next round of translation) and the other by Legating or homologous proteins. However, although this is the normal sequence of events, there are exceptions where the termination reaction is followed by reinitiation on the same Mrna (usually) at a site downstream of the stop codon. The overwhelming majority of such reinitiation events occur when the 5′-proximal open reading frame (ORF) is short and can result in significant regulation of translation of the protein-coding ORF, but there are also rare examples, mainly bicistronic viral RNAs, of reinitiation after a long ORF.

**ROLE OF Trna**

Transfer RNA is an adaptor molecule composed of RNA, typically 76 to 90 nucleotides in length (in eukaryotes), that serves as the physical link between the Mrna and the amino acid sequence of proteins. Transfer RNA (Trna) does this by carrying an amino acid to the protein synthesizing machinery of a cell called the ribosome. Complementation of a 3-nucleotide codon in a messenger RNA (Mrna) by a 3-nucleotide anticodon of the Trna results in protein synthesis based on the Mrna code. As such, tRNAs are a necessary component of translation, the biological synthesis of new proteins in accordance with the genetic code.

The structure of tRNA can be decomposed into its primary structure, its secondary structure (usually visualized as the cloverleaf structure), and its tertiary structure (all tRNAs have a similar L-shaped 3D structure that allows them to fit into the P and A sites of the ribosome). The cloverleaf structure becomes the 3D L-shaped structure through coaxial stacking of the helices, which is a common RNA tertiary structure motif. The lengths of each arm, as well as the loop ‘diameter’, in a Trna molecule vary from species to species.



**ROLE OF RIBOSOME**

Ribosomes also called Palade granules (after discoverer George Palade and due to their granular structure), are macromolecular machines, found within all cells, that perform biological protein synthesis (Mrna translation). Ribosomes link amino acids together in the order specified by the codons of messenger RNA (Mrna) molecules to form polypeptide chains. Ribosomes consist of two major components: the small and large ribosomal subunits. Each subunit consists of one or more ribosomal RNA (Rrna) molecules and many ribosomal proteins (RPs or r-proteins). The ribosomes and associated molecules are also known as the translational apparatus.

Ribosomes are the workplaces of protein biosynthesis, the process of translating Mrna into protein. The Mrna comprises a series of codons which are decoded by the ribosome so as to make the protein. Using the Mrna as a template, the ribosome traverses each codon (3 nucleotides) of the Mrna, pairing it with the appropriate amino acid provided by an aminoacyl-Trna. Aminoacyl-Trna contains a complementary anticodon on one end and the appropriate amino acid on the other. For fast and accurate recognition of the appropriate Trna, the ribosome utilizeS LArge conformational changes (conformational proofreading).

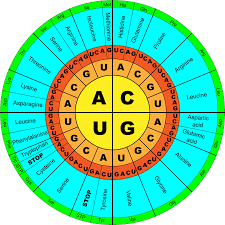
The small ribosomal subunit, typically bound to an aminoacyl-Trna containing the first amino acid methionine, binds to an AUG codon on the Mrna and recruits the large ribosomal subunit. The ribosome contains three RNA binding sites, designated A, P and E. The A-site binds an aminoacyl-Trna or termination release factors; the P-site binds a peptidyl-Trna (a Trna bound to the poly-peptide chain); and the E-site (exit) binds a free Trna. Protein synthesis begins at a start codon AUG near the 5’ end of the Mrna. Mrna binds to the P site of the ribosome first. tHE ribosome recognizes the start codon by using the Shine-Dalgarno sequence of the Mrna in prokaryotes and Kozak box in eukaryotes.

**GENETIC CODE**

The genetic code is the set of rules used by living cells to translate information encoded within genetic material (DNA or RNA sequences of nucleotide triplets, or codons) into proteins. Translation is accomplished by the ribosome, which links proteinogenic amino acids in an order specified by messenger RNA (Mrna), using transfer RNA (Trna) molecules to carry amino acids and to read the Mrna three nucleotides at a time. The genetic code is highly similar among all organisms and can be expressed in a simple table with 64 entries.

A series of codons in part of a messengeR rnA (mRNA) molecule. Each codon consists of three nucleotides, usually corresponding to a single amino acid. The nucleotides are abbreviated with the letters A, U, G and C. This is Mrna, which uses U (uracil). DNA uses T (thymine) instead. This Mrna molecule will instruct a ribosome to synthesize a protein according to this code.

The codons specify which amino acid will be added next during protein synthesis. With some exceptions, a three-nucleotide codon in a nucleic acid sequence specifies a single amino acid. The vast majority of genes are encoded with a single scheme (see the RNA codon table). That scheme is often referred to as the canonical or standard genetic code, or simply the genetic code, though variant codes (such as in mitochondria) exist.



**PYTHON CODE**

inputfile=”Mrna\_sequence.txt”

f=open(inputfile,”r”)

seq=f.read() #reads the genome sequence from input file

seq=seq.replace(“\n”,””) #removes empty lines

seq=seq.replace(“\r”,””)

print(“length of Mrna sequence : “,len(seq))

print()

if(len(seq)%3!=0):

print(“Nucleotide sequence does not have whole number of codons”)

print(“remove “,len(seq)%3,” nucleotides to get accurate results”)

def translate(seq):

table={‘AUA’:’Ile’, ‘AUC’:’Ile’, ‘AUU’:’Ile’, ‘AUG’:’Met’,

‘ACA’:’Thr’, ‘ACC’:’Thr’, ‘ACG’:’Thr’, ‘ACU’:’Thr’,

‘AAC’:’Asn’, ‘AAU’:’Asn’, ‘AAA’:’Lys’, ‘AAG’:’Lys’,

‘AGC’:’Ser’, ‘AGU’:’Ser’, ‘AGA’:’Arg’, ‘AGG’:’Arg’,

‘CUA’:’Leu’, ‘CUC’:’Leu’, ‘CUG’:’Leu’, ‘CUU’:’Leu’,

‘CCA’:’Pro’, ‘CCC’:’Pro’, ‘CCG’:’Pro’, ‘CCU’:’Pro’,

‘CAC’:’His’, ‘CAU’:’His’, ‘CAA’:’Gln’, ‘CAG’:’Gln’,

‘CGA’:’Arg’, ‘CGC’:’Arg’, ‘CGG’:’Arg’, ‘CGU’:’Arg’,

‘GUA’:’Val’, ‘GUC’:’Val’, ‘GUG’:’Val’, ‘GUU’:’Val’,

‘GCA’:’Ala’, ‘GCC’:’Ala’, ‘GCG’:’Ala’, ‘GCU’:’Ala’,

‘GAC’:’Asp’, ‘GAU’:’Asp’, ‘GAA’:’Glu’, ‘GAG’:’Glu’,

‘GGA’:’Gly’, ‘GGC’:’Gly’, ‘GGG’:’Gly’, ‘GGU’:’Gly’,

‘UCA’:’Ser’, ‘UCC’:’Ser’, ‘UCG’:’Ser’, ‘UCU’:’Ser’,

‘UUC’:’Phe’, ‘UUU’:’Phe’, ‘UUA’:’Leu’, ‘UUG’:’Leu’,

‘UAC’:’Tyr’, ‘UAU’:’Tyr’, ‘UAA’:’\_’, ‘UAG’:’\_’,

‘UGC’:’Cys’, ‘UGU’:’Cys’, ‘UGA’:’\_’, ‘UGG’:’Trp’,}

protein=””

if len(seq)%3 == 0:

for i in range(0,len(seq),3):

codon=seq[i:i+3]

protein+=table[codon]+”-“

return protein

p=translate(seq[3:len(seq)])

print(“Number of codons translated : “,int(len(seq)/3))

print()

print(“Protein sequence is”,p)

**ALGORITHM**

The code reads Mrna sequence from txt file

Then it trims any white spaces and empty lines from the sequence.

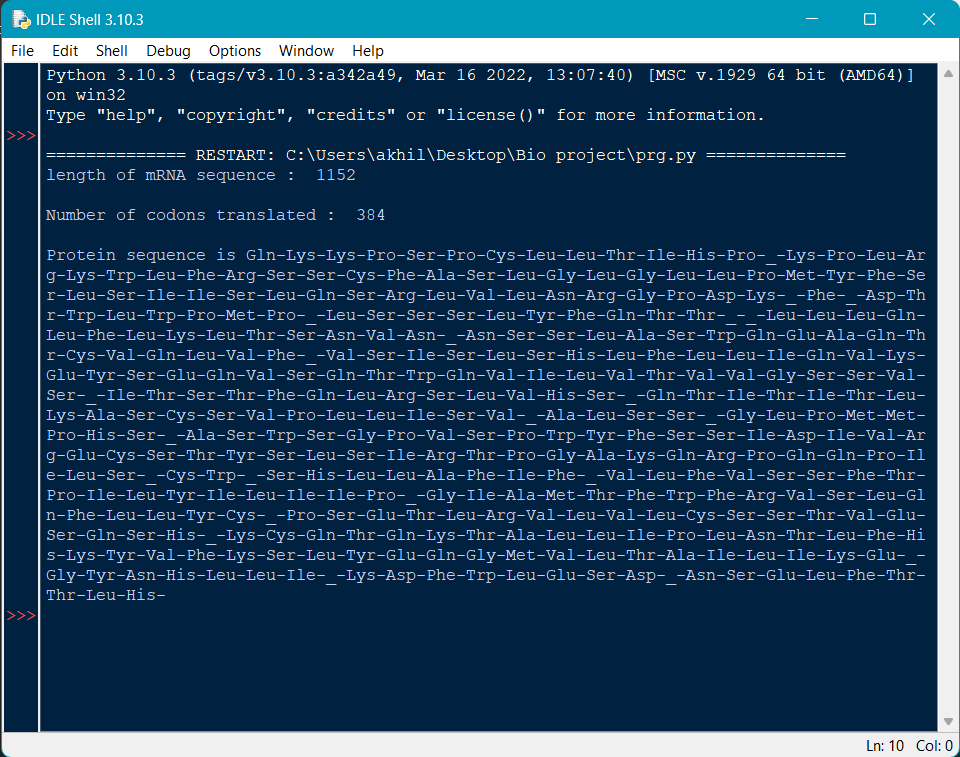
Then the code checks if the number of codons in the sequence is a whole number or not.

If no user is prompted to correct the number of nucleotides in the sequence.

If yes translation is done by taking one codon at a time and finding its key-value pair from the genetic code dictionary.

Number of codons translated and protein sequence generated is printed at the end of execution.

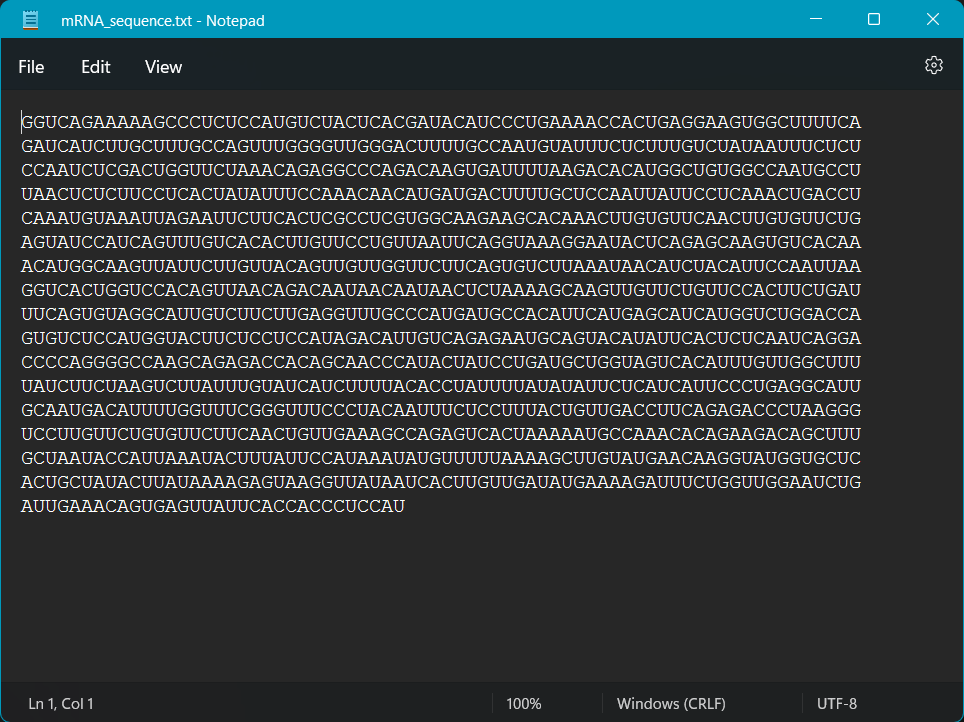
**RESULT AND ANALYSIS**



Output amino acid sequence generated.

Stop codons are translated as ‘\_’

Adjacent amino acids are separated using ‘-‘ which symbolizes the peptide bond which exists between adjacent amino acids in proteins.

mRNA sequence taken from ncbi

**CONCLUSION**

The python code perfectly bio mimics the functionality of ribosomes and tRNA molecules inside cell mechanisms.

The amino acid sequence generated can further be used to study the chemical and physiological effect of the protein translated from mRNA on the body.

The system can be further enhanced by stopping the translation process whenever a stop codon is encountered as is the case with the actual translation process inside cells.