International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume: 3 | Issue: 2 | Jan-Feb 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 - 6470

Computational Analysis of RNA Nucleotide Sequences

Shall Juneja[1], Deepayan Mukherjee[2], Sachi Garg[2]

[1]Assistant Professor, [2]Student
[1]Computer Science and Engineering,
[2]Electronics and Communication Engineering
Maharaja Agrasen Institute Of Technolgoy, Delhi, India

ABSTRACT
One of the most fascinating and deeply researched fields of all of the biochemical components are the nucleic acids and proteins, especially because of their dynamic structure and function and their fundamental role in the course of evolution. Sequence studies have been carried out and is still being researched on to decode completely the world of genetics and molecular biology. And so, here we present our concept in the form of a simple program that has been developed for analyzing the Amino acid sequence from a given sequence of RNA nucleotide bases, and using the result of which, further, interpretation of the corresponding protein can be done.

KEYWORDS: *Amino acid, RNA, Nucleotide Sequence, Protein, Gene, Codon*

I. INTRODUCTION

Through the lens of cell biology, the research in the field of gene expression is closely linked with the understanding of the proteins, their functions ad structures.[1] Way back, from the very early work of Christian Anfinsen around 1950s, it is a stated fact that the amino acid sequence in a protein determines the final three-dimensional protein structure. Not only his, scientists have also continuously observed and somewhat concluded that the protein structure has the fundamental role in dictating where the protein will act and what will be its course of action, a prominent example of which comes with the study of functioning of enzymes, in which the shape - structure of proteins fundamentally determines its expression and functioning capabilities. Besides, proteins have a crucial role in determining the cellular sub-regions where the modulation of gene expression will occur - nucleus, cytoplasm, cell membrane etc...[2]

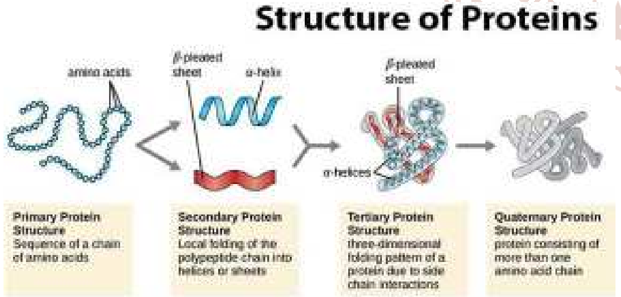

Structure of Proteins
amino acids
β-pleated sheet
α-helix
β-pleated sheet
α-helices
Primary Protein Structure
Sequence of a chain of amino acids
Secondary Protein Structure
Local folding of the polypeptide chain into helices or sheets
Tertiary Protein Structure
three-dimensional folding pattern of a protein due to side chain interactions
Quaternary Protein Structure
protein consisting of more than one amino acid chain

Fig. 1 - STRUCTURE OF PROTEIN [3]

But as stated above, the fundamental step in studying a protein, its structure and function is the analysis and determination of the amino acid sequence present in it, corresponding to the nucleotides present in it. Also, amino acid sequencing helps in predicting the charge of a molecule, its size, and thereby a probable 3D structure of it.

Biologically, the synthesis of protein is quite complex. It involves many steps. In human body DNA is present and this DNA has segments called as genes which in turn fundamentally guide the process of protein formation in human body. However, the human body cannot directly decide and synthesis which protein is to be formed, when and how, from the DNA itself. And so, firstly, part-wise, the double stranded helical DNA is opened with the help of various enzymes present in the body and each strand is replicated. Each strand of the DNA has a 5' and 3' end in which nucleotides are attached. In DNA, there are only 4 Nucleotides, namely A= Adenosine, T=Thymine, G= Guanine and C= Cytosine and between the 2 strands, the rule of pairing states that A will be paired with T and G with C. And so, during replication if a strand has a sequence ATTGGC, the replicated strand will always have TAACCG in it.

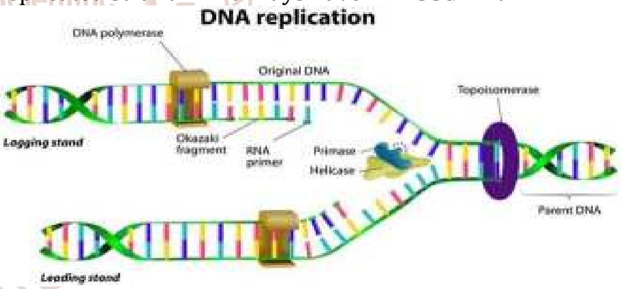


Fig. 2 - THE PROCESS OF DNA REPLICATION [4]

After replication, there comes the process of transcription, wherein the part of DNA is converted to RNA with the help of various enzymes. The change that RNA has with respect to DNA, in terms of nucleotides, is that instead of ATGC, RNA has AUGC i.e., T= Thymine is replaced by a nucleotide called Uracil,, denoted by U. Thence, now, the pairing is between A - U and G- C.

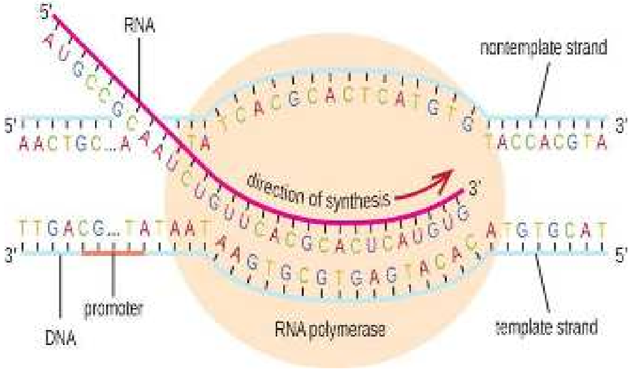


Fig. 3 - THE PROCESS OF TRANSCRIPTION [5]

After this, there comes the final process, in which out of the 2 strands of RNA – the template strand and the coding strand, the nucleotides on the template strand are coded for their corresponding amino acids. While coding, a triplet of nucleotide is considered. To mark the beginning and the end of coding, special factors automatically gets attached to it. Moreover, AUG which codes for Methionine is said to be the initiator codon. Similarly, there are 3 codons (triplet of nucleotide) which when encountered lead to termination of the process of translation – UUA, UAG and UGA. After amino acids are coded for/obtained, they are joined by special bonds called as peptide bonds which finally link them together to form a particular protein (chain of amino acid).

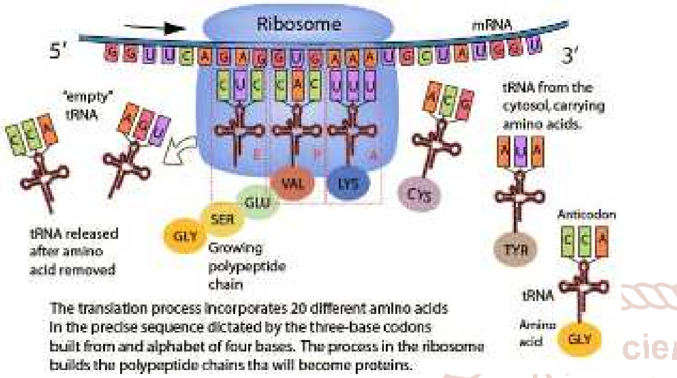

Ribosome
mRNA
5′
3′
GGUUCAGAGGUGAAAUGCUAUGGU
CUC
CAC
UUU
"empty" tRNA
CCA
AGU
ACG
tRNA from the cytosol, carrying amino acids.
AAA
VAL
LYS
CYS
GLU
SER
GLY
TYR
Anticodon
CCA
tRNA released after amino acid removed
Growing polypeptide chain
tRNA
Amino acid
GLY
The translation process incorporates 20 different amino acids In the precise sequence dictated by the three-base codons built from and alphabet of four bases. The process in the ribosome builds the polypeptide chains tha will become proteins.

Fig. 4 - THE PROCESS OF PROTEIN TRANSLATION [6]

It is already a known fact, that for proper functioning of biological components this process of protein translation and expression should be carried out with highest efficiency, precision and fidelity. Given a sequence of amino-acid from which protein is to be coded, there exists many multiple degrees of freedom which may introduce minor errors in the process, which in turn can drastically lead to severe genetic errors, defects and evolutionary flaws /faults. Particularly, the redundancy of the genetic code which provides the choice between alternative codons, coding for the same amino acid, which, although may sound 'synonymous,' but might exert catastrophic effects on the translation process. And so, it becomes an absolute necessity that the given sequence of nucleotide and thence the corresponding sequence of amino acid obtained, be thoroughly examined with the highest degree of precision. [7]

In general, the task of sequence detection is often accomplished by the means of a technique called as Mass Spectrometry, after which once the sequence is known, a comparison with various databases allow researchers and scientists to find whether or not there are any related proteins corresponding to the detected sequence. And thus, in this way, protein databases are formed, which form an essential part of modern biology.

In today's world, such huge amounts of data regarding different protein structures, their varied functions, and particularly their sequences are being continuously analyzed and researched about. But such dynamic and extremely large data is quite complex to be handled manually and there may be chances of error too. And so, that's where the computers step in. [8] Today, in the age of digitization, there have been many ongoing researches in this field to computerize the process and probably that is one of the main reason why the field of computational biology is one of the most demanding and emerging field in the recent times.

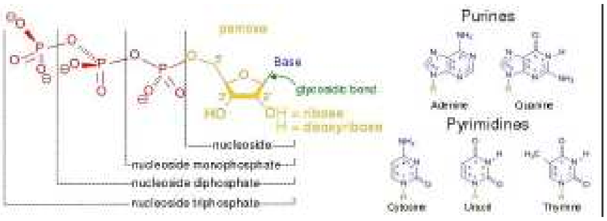
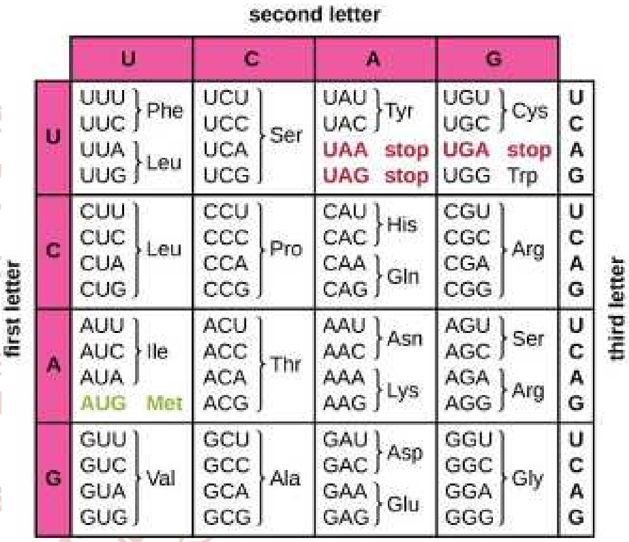


Fig.5. - NULCEOTIDES AN NULEOSIDE STRUCTURES [9]

And so, integrating our knowledge of cell biology and computers, we have tried developing a simple algorithm and thus a program, which helps in determining the nucleotide sequence of a protein which when used in simulation with other pre and post programs can be beneficial in studying proteins and determining its function and structure. Moreover it also shows how the impact and evolution of computers have led to a wonderful collaboration of it with the field of biology and how this has helped in handling and analysis of the sequence data.



	second letter				
first letter	U	C	A	G	third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	UAA stop	UGA stop	A
	UUG	UCG	UAG stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lys	AGA Arg	A
	AUG Met	ACG	AAG	AGG	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glu	GGA	A
	GUG	GCG	GAG	GGG	G

Fig. 6 - RNA NUCLEOTIDE – AMINO ACID CHART [10]

II. METHODOLOGY OF APPROACH

Under the supervision of Ms. Shallu Juneja, I along with Ms.Sachi has been researching across the internet and reading various previously published papers with regard to different aspects of this paper, on which we have been working for many months now.

We, firstly, studied from books of varied authors regarding cell biology and the ongoing research in this field. Secondly, we studied the gene expression patters and how from amino acids proteins are expressed. We, also, went on to study about nucleotides, nucleosides, transcription, translation, replication and other associated process, which we thought might be useful for us in the arena of writing the research paper and conducting our research.

After this we tried and plotted an algorithm using which, corresponding amino acid sequences could be determined from the inputted sequence of nucleotide. Thereafter, we were in search of a programming language which could sustain the dynamicity of the algorithm but also at the same time be sufficiently simple to write in, to understand in and to be used so that even a layman running the code could get the cent percent result out of it. And so, we chose C++ as our foundational base language and developed the program in it.

Another reason for choosing C++ was that, since it is well taught in almost all schools and/or other educational institutions, the algorithm and thereby the program could thence be used by many and could be beneficiary to a greater public in general.

The program which we developed uses the nucleotide chart as a reference to give as output the corresponding amino acid from the detected/inputted sequence.

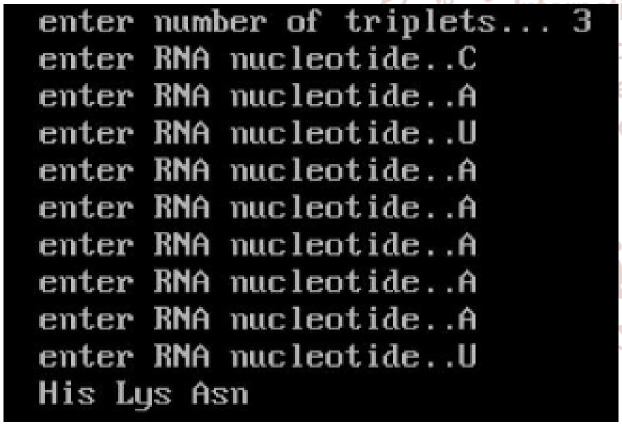
III. ALGORITHM INSIGHTS

The basic algorithm that runs as the foundation of the program is:-

1. Inputting of sequence of nucleotides from the user or from any external source.
2. The Nucleotides that are accepted from the user follows two criterions :-
   A. They should be in the upper case (A,U,G,C)
   B. They should be in triplets- as per the Nucleotide table
3. The program runs through the codes and as per the Nucleotide table, gives as output the corresponding amino acid in the order of inputted sequence
4. If :-
   A. No triplet is formed at some point in the end Or
   B. A triplet of Stop codon is encountered the program gives as output, in order, the corresponding amino acids till before it.

IV. PROGRAM OUTPUT AND DISCUSSION

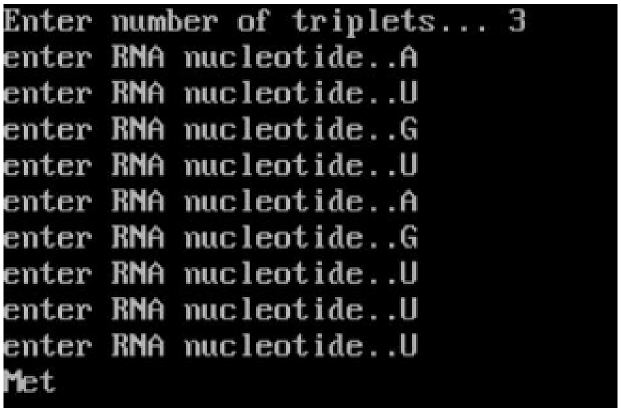
1.



```
enter number of triplets... 3
enter RNA nucleotide..C
enter RNA nucleotide..A
enter RNA nucleotide..U
enter RNA nucleotide..A
enter RNA nucleotide..A
enter RNA nucleotide..A
enter RNA nucleotide..A
enter RNA nucleotide..A
enter RNA nucleotide..U
His Lys Asn
```

In this particular screenshot of the output, we can see that how a given sequence of nucleotide is inputted by the user and the program gives the corresponding amino acid in the order of the triplet of nucleotide (As per the Nucleotide table).

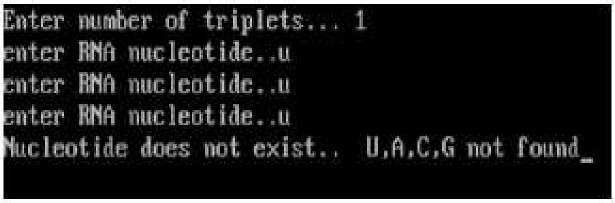
2.



```
Enter number of triplets... 3
enter RNA nucleotide..A
enter RNA nucleotide..U
enter RNA nucleotide..G
enter RNA nucleotide..U
enter RNA nucleotide..A
enter RNA nucleotide..G
enter RNA nucleotide..U
enter RNA nucleotide..U
enter RNA nucleotide..U
Met
```

As stated earlier, that there are 3 stop codon which signal the termination of the process of translation– UAA, UGA, UAG. In this particular screenshot of the output, it can be clearly observed that how the upon detecting UAG, which is a stop codon, the program stops i.e., codes only till before it.

3.



```
Enter number of triplets... 1
enter RNA nucleotide..u
enter RNA nucleotide..u
enter RNA nucleotide..u
Nucleotide does not exist..  U,A,C,G not found_
```

Since, in biology, by convention the nucleotides are represented in upper case, hence it can bee seen in this screenshot of the output that the program is sensitive to the Upper Case Alphabets(A,U,G,C)

V. CONCLUSIONS

1. It was thence observed that when the program is run, it successfully inputs the triplets of nucleotides from the user and gives as output the corresponding amino acid to the detected sequence.
2. It can also been clearly seen that the program is sensitive to the letters in Upper Case as the names of the nucleotides are written, by convention, the upper case.
3. Also, it is evident that the program can recognize the stop sequences and respond to it/ take necessary actions i.e., code the amino acid only till before it encounters a stop codon.

VI. FUTURE SCOPE OF IMPLEMENTATION

The program that has been developed is basic in terms of the language used for execution and coding. But, it caters to and takes into account all what is necessary in analyzing and determining which nucleotide triplet gives which corresponding amino acid, as per the sequence given. If this program is further extended in terms of implementation and used in collaboration with programs developed in the sphere of biodata analysis or artificial intelligence, it can be very helpful in the field of research regarding protein sequences and help in various associated field. The robustness of the program can definitely be increased accordingly and so can be its dynamicity, but as stated earlier, in this attempt of ours we aim just to realize the concept of nucleotide sequence recognition and analysis in form of coding and that to on/in a language which is readily usable, available and can be understood even by layman.

A. Appendix

1. **Nucleotide** - Organic molecule which is the fundamental unit or basic building block of DNA and/or RNA.
2. **Codon** - Sequence of 3 DNA or RNA nucleotides which corresponds to a specific amino acid
3. **Stop Codon** - A nucleotide triplet which signals termination of translation into proteins. It may be UAA, UAG, and UGA
4. **Initiator codon** – A triplet which signals the process of translation of protein, which generally codes for methionine in eukaryotes and fMet in prokaryotes. The most common start codon is AUG.

ACKNOWLEDGMENT

First and foremost, I would like to thank my teachers, especially Mrs. Kanishka Kumar, Dr. Subhash Singh and Mrs.

Meeta Chopra for their valuable inputs and their painstaking efforts in teaching and guiding me theoretically and conceptually, and without whom this paper wouldn't have been a success.

Also, my gratitude goes to my co-author, my colleague and friend Ms. Sachi, without who's codding expertise I wouldn't have been able to realize the algorithm developed into a proper program. Not only that, it is because of her inputs that the program developed could run successfully on all given platforms, for which I would again like to thank her.

Alongside, I would also like to thank my grandparents and parents for their constant motivation and support, especially during the many times I met failures and above all, almighty God for giving me the energy to carry out, in theory and practical, whatever I could to make this research paper a success.

REFERENCES

[1] Dayhoff M.O. (1974)Computer analysis of protein sequences. In: Siler W., Lindberg D.A.B. (eds) Computers in Life Science Research. FASEB Monographs, vol 2. Springer, Boston, MA

[2] https://www.nature.com/scitable/topic/proteins-and-gene-expression-14122688

[3] Fig.1- https://www.biosciencetimes.com/study-notes/biochemistry/protein-structure/169/

[4] Fig.2- https://www.news-medical.net/life-sciences/DNA-Replication-and-Repair.aspx

[5] Fig.3. https://courses.lumenlearning.com/microbiology/chapter/rna-transcription/

[6] Fig.4- http://hyperphysics.phy-astr.gsu.edu/hbase/Organic/translation.html

[7] Gingold, H., & Pilpel, Y. (2011). Determinants of translation efficiency and accuracy. *Molecular systems biology*, *7*, 481.

[8] Xu D, Xu Y. Protein databases on the internet. *Curr Protoc Mol Biol*. 2004; Chapter 19: Unit 19.4.

[9] Fig.5- https://www.diffen.com/difference/Nucleoside_vs_Nucleotide

[10] Fig.6- http://biology-pictures.blogspot.com/2013/10/table-of-genetic-code.html

[11] Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965 from the Journal of the History of Biology 43(4):pp 623-60

[12] Molecular Basis of Inheritance : chapter 19 from National Council of Education Research and Training book in Biology, Class 12th