2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579606

Codon Optimization and Converting DNA Sequence into Protein Sequence using Deep Neural Networks

Y. Tarakaram
Department of Computer Science and Engineering
Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham, India
yadallatarakaram@gmail.com

Y.Mounika
Department of Computer Science and Engineering
Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham, India
ymouni626@gmail.com

Y.Lakshmi Prasanna
Department of Computer Science and Engineering
Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham, India
prasanna3002@gmail.com

Tripty Singh
Department of Computer Science and Engineering
Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham, India
tripty_singh@blr.amrita.edu

***Abstract* - In bioinformatics, finding the protein sequences plays a crucial role, as they help in developing anti-biotics, which help in curing many diseases. Protein sequence analysis is used to identify amino acid sequence. Identification of structure and function of proteins plays a crucial role in understanding cellular processes. These protein sequences are obtained from DNA sequences by dividing them into codons. DNA sequences are first divided into codons and these codons can be optimized before converting them into protein sequence. Codon optimization is done to achieve transcription efficiency, translation efficiency, gene synthesis, protein folding, etc as this optimization helps in reducing the GC contents. It can also solve some of the problems of protein expression like cannot be expressed in heterologous systems, expression level is very low, unable to fold properly, lose functional activity, etc. After codon optimization, the resultant sequence is converted into protein sequence using deep neural networks.**

***Keywords* - BioInformatics, Codon Optimization, Protein sequences, Neural networks.**

I. INTRODUCTION

DNA (Dioxyribonucleic Acid) contains all the necessary information required to maintain an organism. Every living creature has DNA in their cells. It contains genetic information which plays an important role in heridity of organisms. It contains two strands - one parent strand and one complementary strand which are made of four nucleotides A (Adenine), G (Guanine), C (Cytosine), T (Thymine). Three consecutive nucleotides present in DNA or RNA are considered as codons.

As we have 4 nucleotides, 64 triplets can be formed using them. So there are 64 codons, out of which 61 codons are coded into 20 standard amino acids and the other 3 codons are considered as stop codons namely - TAA, TGA, TAG[2]. Some amino acids are coded by multiple codons. For example, TTT and TTC code for same amino acid i.e., Phenylalanine similarly, TAT and TAC code for amino acid Tyrosine, etc.

Bias towards certain codons usage to code for same amino acid is exhibited differently by different organisms. The effect of these differences in different organisms is still a subject to debate in the field of biotechnology. These differences show a great impact on protein expression. Therefore codon optimization is important in performing study of protein expression.

A. Codon Optimization

Codon optimization is a technique used to optimize codons which helps in increasing the transcriptional and translational efficiency of a particular gene between the foreign cells[1]. It uses codon bias to increase translational and transcriptional efficiency. There are different parameters or constraints for codon optimization. Some of them are-

1. Codon Usage Bias:

Transcriptional inefficiency occurs when there is less amount of amino acid in the tRNA so in those cases due to degeneracy we can replace the codon with another codon representing the same amino acid. For example, we have UUU codon and one tRNA carrying phenylalanine if there is any transcriptional inefficiency we can replace the last nucleotide of the codon with C then we get UUC codon which also codes for phenylalanine. This is done by a codon optimizer. This process is called codon usage bias.

2. GC-content:

Transcriptional and translational efficiency can also be increased based on GC - content in the DNA. As the GC bond provides more stability than AU bond if the GC content is more then it will be difficult to unwind the DNA which makes it difficult to produce mRNA. So here also we can replace a codon with another codon that codes for the same amino acid to reduce GC - content[6].

3. mRNA structure:

Translational inefficiency is caused when there is any hairpin loop in mRNA. When an mRNA strand folds and forms base pairs with another section of the same strand which leads to an unpaired loop known as hairpin loop[3]. This happens when there are complementary sequences in a single mRNA strand. This can also be avoided by optimizing codon.

4. Repeated sequences:

Repeated sequences also decrease the

transcriptional and translational efficiency so this can also be avoided using codon optimization.

5. Avoid restriction site of enzymes:

If there is any restriction site in the gene that might lead to cleavage of the gene when it is inserted into foreign cell. Therefore, in order to avoid that we can replace the restriction sites of the enzyme. This can also be achieved by codon optimization[10]

6. Non-coding regions:

Non-coding regions are called introns. After mRNA synthesis the introns are replaced and exons are joined together to get matured RNA. If this splicing process is absent then it leads to poor transcription and translation so this can be done using codon optimization.

B. Central Dogma

Central Dogma in biology is said to be the flow of genetic information from DNA to RNA, to make functional products called proteins. The instructions to make protein are present in DNA. DNA is converted to mRNA (messenger RNA) through transcription process. mRNA is converted to protein sequence by ribosomes through translation process[8].

1. Transcription:

The process in which the enzyme, RNA polymerase transfer information from one strand of the DNA to RNA is known as transcription. The template DNA strand which is converted to RNA contains three regions promoter, structural gene and terminator. Initiation, Elongation and Termination are the three stages in Transcription. The promoter region of the DNA tells the RNA polymerase where to bind in the stage one i.e., initiation stage.

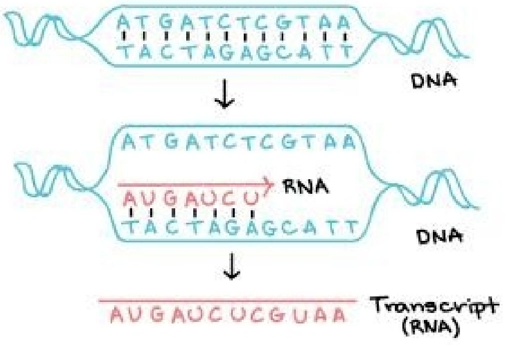


Fig 1: Creation of mRNA through transcription

Gene expression in mainly controlled in this region by restricting or allowing access to the promoter region by RNA polymerase. DNA unwinds upon binding[9]. During second stage i.e., elongation, the RNA polymerase moves across the template of the DNA strand. Nucleotides are joined to the 3' end of the RNA molecule by the RNA polymerase as the complementary nucleotides join together. After reaching the termination point of the DNA, DNA strand, RNA polymerase and the resultant mRNA dissociates from each other.

During transcription, the mRNA strand contains two regions - one is exons that code for proteins which are also known as coding regions and other is introns which are also known as non-coding regions. Introns must be removed to translate mRNA. This process of removing introns is called intron splicing as shown in fig 2. This process helps in getting mRNA strand which contains only the exons, the resultant mRNA is called mature mRNA[11]. This mature mRNA now enters the cytoplasm by leaving the nuclear through a nuclear pore.

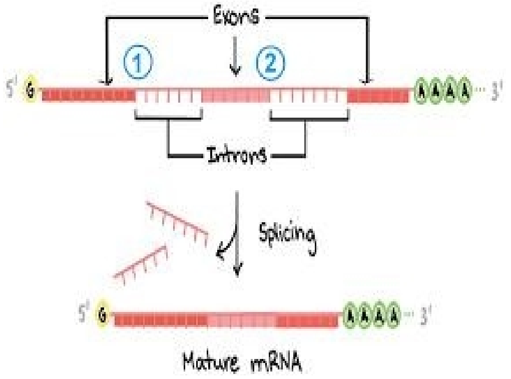


Fig 2: Process of Intron splicing

2. Translation

Initially, mRNA strand binds to ribosomes in translation. First, mRNA binds to the start codon at its ribosomal sub-unit. Specific tRNA molecule brings each amino acid to the ribosome. Anti-codon sequence present in the tRNA molecule specifies the amino acid. The complementary bases pair up between the codon of the mRNA strand and the anti-codon present on the tRNA.

After the initiator tRNA molecule and the start binds, the translation complex is formed from the binding of the large ribosomal sub-unit and the initiation is completed. Three distinct regions are present in the ribosomal sub-unit - E, P and A sites as shown in Fig 3.

During elongation, tRNA molecule brings individual amino acids to the mRNA through base pairing of the complementary bases of the codons and anti-codons. Each anticodon present in the tRNA molecule specifies a particular amino acid[12]. A peptide bond is formed between the tRNA molecule and its amino acid at the P site when the charged tRNA molecule is attached to the A site. When binding is done, the protein sequence moves down one codon to its right and the uncharged tRNA exists from E-site and now A-site accepts the next tRNA molecule. Until the stop codon is reached, this process of elongation is continued. The complex peptide is detached from the tRNA at the P site as the release factor gets bounded at the A site to the stop codon. Then the whole peptide is dislodged. To start the process at initiation, it can reassemble. To produce polypeptides rapidly and correctly the process of translation is used. After dissociation, the polypeptide is needed to be modified to make it ready to perform.

Ribosome 'sites' based on rRNA structure

A = amino acid entry (binding) site

P = 'peptidyl-tRNA' binding site

E = exit site

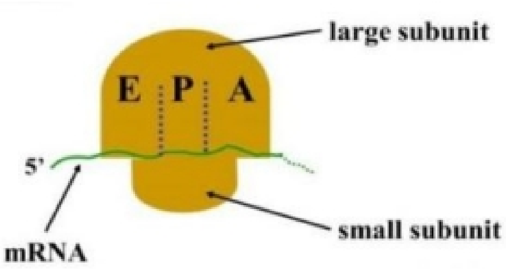


Fig 3: Different binding sites of ribosomes

II.RELATED WORK

In [1], they have mentioned the details of implementing and working of the codon optimization using a BiLSTM model. The codons that are active in the initiation of the process of translation is explained in [2].The protein expression of a particular gene is compared with the gene of E.coli to show the importance of the process of codon optimization in [3]. The work of implementing a special tool for codon optimization that is codon optimizer is done by [7]. The use of codon usage bias to diminish the efficiency of the protein expression in the human gene is explained in [10]. The efficient protein expression in mammalian cells is done using the selected specified human codons whose GC content is over 60 percent (performed using the process of codon optimization) in [6].

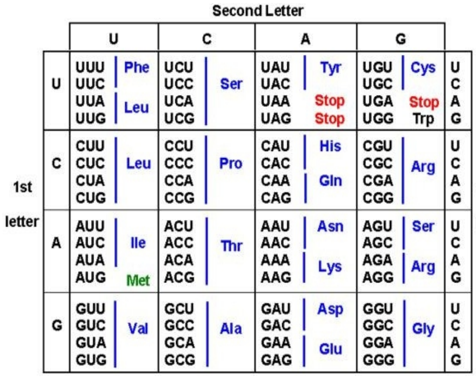
Instead of implementing a model for codon optimization in [1], we directly use the codon optimizer tool to perform the codon optimization to reduce the system work and later we convert the sequence into protein.

III.METHODS

A. Training set:

For training the model, all the 64 codons formed using 4 nucleotides and their corresponding amino acids are taken.

Training input is codons and output is corresponding amino acids.



Second Letter

1st letter	U		C		A		G		3rd letter
U	UUU UUC UUA UUG	Phe Leu	UCU UCC UCA UCG	Ser	UAU UAC UAA UAG	Tyr Stop Stop	UGU UGC UGA UGG	Cys Stop Trp	U C A G
C	CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC CAA CAG	His Gln	CGU CGC CGA CGG	Arg	U C A G
A	AUU AUC AUA AUG	Ile Met	ACU ACC ACA ACG	Thr	AAU AAC AAA AAG	Asn Lys	AGU AGC AGA AGG	Ser Arg	U C A G
G	GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC GAA GAG	Asp Glu	GGU GGC GGA GGG	Gly	U C A G

Fig 4: Codon with corresponding amino acids which is used as training data

B. Codon Optimization model

Many parameters or constraints are considered to optimize a codon. Some of them are -

1. Checking that the sequence has no matches longer than N in a given index.
2. Avoiding Hairpin patterns in DNA or mRNA structure. mRNA has more Hairpin structures than DNA.
3. Checking whether the given pattern is absent in the sequence.
4. Specifying the proportion of GC content in a particular window size in the DNA.
5. Making sure that no new stop codon is introduced into the frame.
6. Checking that the optimized sequence also produces same amino acid sequence or protein after translation.

C. Network Architecture

In our project we used ANN (Artificial Neural Networks) model which is a feed forward neural network. Codon is taken as input for the model. Output obtained from the model is amino acid. In this model we have four neural network layers - 1 input layer, 2 hidden layers and 1 output layer. 1st hidden layer consists of 512 neurons and next hidden layer consists of 256 neurons. Most of the processing of the data is done in hidden layers. Activation function for both the hidden layers is taken as reLU (Rectified Linear Unit). Softmax is taken as activation function for output layer. This softmax function helps in multi-class classification. It gives us output in such a way that sum of all the outputs is 1. So, they can be considered as probabilities. Loss function used for the model is sparse categorical cross entropy. It is a variant of categorical cross entropy. This function is used when classes are mutually exclusive i.e., each sample input can be categorized into only one class. Since, only one amino acid is coded by each amino acid, all the amino acid classes are said to be mutually exclusive.

In the sample codon, each nucleotide is one hot encoded i.e., A, T, G and C are encoded as [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0] and [0, 0, 0, 1] respectively. This encoded codon is converted into tensor and given as input to the neural network. Input layer of the neural network is taken as flat layer which converts the 3x4 tensor into 12x1 and pass to hidden layers for the processing. Amino acids are label encoded i.e., a numeric value is given to each amino acid. So, the neuron which gives the highest probability is then decoded to give the predicted amino acid.

D.Implementation

Codon optimization is done using DNA Chisel tool. It is a python library which basically used in bioinformatics. It optimizes the codon with respect to the given constraints and objectives. It has almost 15 classes of sequence applications to codon optimize genes. It helps in meeting the contraints of DNA provider, to tune GC content and many more at once.

We built neural network model using Keras API in tensorflow. Model is trained using adam optimizer.

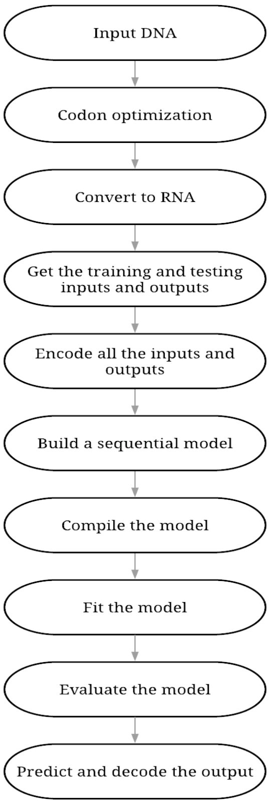


Fig 5: Flow diagram of the implementation

III. RESULTS AND DISCUSSIONS

Initially accuracy is less and loss is more in the model. Gradually as the epochs increases, accuracy increased and reached 1.0 and loss also decreased for the model. As we are converting a codon to amino acid accuracy must be more to get appropriate amino acid. So, this is achieved using our model.

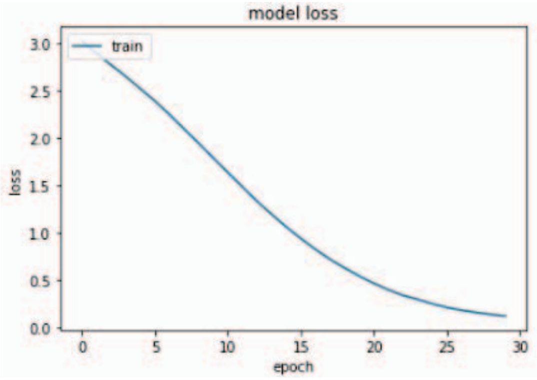


Fig 6: Accuracy vs Epoch of the model

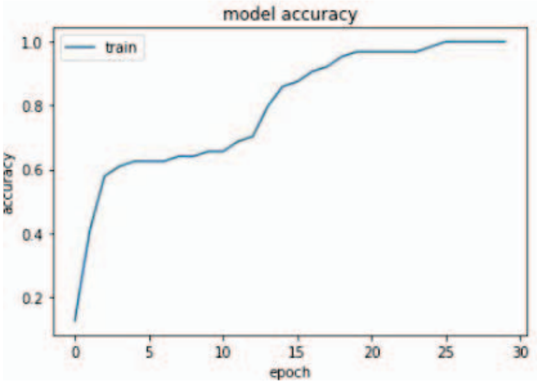


Fig 7: Loss vs Epoch of the model

For testing our model we have used TP53 tumour protein [Homo sapiens] DNA sequence. This sequence is codon optimized and given as testing input to the model. After evaluating the model accuracy 1.0 is achieved. Protein sequence after and before codon optimization is also same.

IV. CONCLUSION

Through this project the transcriptional and translational efficiency of the DNA sequence is increased using codon otptimization. This helps in improving gene expression, protein expression, increase elongation rate and can also stabilize DNA sequence. The optimized DNA sequence can also be used in gene therapy, mRNA therapy, producing DNA or RNA vaccines, recombinant protein drugs, etc.

Although DNA contains the important information of an organism, most of the biological activities are carried out by proteins. DNA contains the information which is required to produce a protein. Protein synthesis is very much important because it is responsible for proper cell and organism functionality. Therefore, the solution to the production of functional proteins is the assembly of the amino acids in the order encoded in the DNA. Protein have varying functions in biological systems like some proteins are structural components (like keratin) and the other proteins acts like enzymes. Some other functions of are immunity (like antibodies), transport (like haemoglobin) and regulation. These protein sequences are also used to make antibiotics. Antibiotics are used to cure some bacterial diseases, etc.

In our work, we can add a codon optimizer tool which changes its constraints and parameters based on the input DNA sequence. We can also add a model which compares the output protein sequence with some known protein sequences to get a comparative study of the new sequence which is obtained.

REFERENCES

[1] Hongguang Fu, Yanbing Liang, Xiuqin Zhong, ZhiLing Pan, Lei Huang, HaiLin Zhang, Yang Xu, Wei Zhou & Zhong Liu, “Codon optimization with deep learning to enhance protein expression”, Sci Rep 10, 17617 (2020).

[2] A. Hecht, J. Glasgow, P. R. Jaschke, L. Bawazer, M. S. Munson, J. Cochran, D. Endy, and M. Salit, "Measurements of translation initiation from all 64 codons in e. coli," bioRxiv, p. 063800, 2016.

[3] Liu, B., Kong, Q., Zhang, D. & Yan, L. "Codon optimization significantly enhanced the expression of human 37-kDa iLRP in Escherichia coli." 3 Biotech 8(4), 210 (2018).

[4] Al-Hawash, A. B., Zhang, X. & Ma, F. "Strategies of codon optimization for high-level heterologous protein expression in microbial expression systems." Gene Rep. 9, 46–53 (2017).

[5] Natalie J. Ward, Suzanne M. K. Buckley, Simon N. Waddington, Thierry VandenDriessche, Marinee K. L. Chuah, Amit C. Nathwani, Jenny McIntosh, Edward G. D. Tuddenham, Christine Kinnon, Adrian J. Thrasher, John H. McVey; Codon optimization of human factor VIII cDNAs leads to high-level expression. Blood 2011; 117 (3): 798–807.

[6] Inouye, S., Sahara-Miura, Y., Sato, J. I. & Suzuki, T. Codon optimization of genes for efficient protein expression in mammalian cells by selection of only preferred human codons. Protein Expr. Purif. 109, 47–54 (2015).

[7] Anders Fuglsang, Codon optimizer: a freeware tool for codon optimization, Protein Expression and Purification, 2003, ISSN 1046-5928

[8] CRICK, F. Central Dogma of Molecular Biology. Nature 227, 561–563 (1970).

[9] Shapiro JA. Revisiting the central dogma in the 21st century. Annals of the New York Academy of Sciences. 2009 Oct; 1178:6-28

[10] Nicola A. Burgess-Brown, Sujata Sharma, Frank Sobott, Christoph Loenarz, Udo Oppermann, Opher Gileadi, Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study, Protein Expression and Purification, Volume 59, Issue 1, 2008, Pages 94-102, ISSN 1046-5928

[11] Chan, R.T., Peters, J.K., Robart, A.R. et al. Structural basis for the second step of group II intron splicing. Nat Commun **9,** 4676 (2018)

[12] Cole, M., Cowling, V. Transcription-independent functions of MYC: regulation of translation and DNA replication. Nat Rev Mol Cell Biol **9,** 810–815 (2008)

[13] Koonin, E.V. Why the Central Dogma: on the nature of the great biological exclusion principle. Biol Direct 10, 52 (2015)

[14] Alexaki, A., Hettiarachchi, G.K., Athey, J.C. et al. Effects of codon optimization on coagulation factor IX translation and structure: Implications for protein and gene therapies. Sci Rep 9, 15449 (2019)

[15] Satoshi Inouye, Yuiko Sahara-Miura, Jun-ichi Sato, Takahiro Suzuki, Codon optimization of genes for efficient protein expression in mammalian cells by selection of only preferred human codons, Protein Expression and Purification, Volume 109, 2015, Pages 47-54, ISSN 1046-5928

[16] T. Babu, Dr. Tripty Singh, Dr. Deepa Gupta, and Hameed, S., "Prediction of Normal & Grades of Cancer on Colon Biopsy Images at Different Magnifications using Minimal Robust Texture & Morphological Features", International Journal of Bioinformatics Research and Applications, 2019

[17] N. Oruganti and Dr. Tripty Singh, "Best Fit Polygonal Approximation for Multiple ROI Estimation", 2020 11th International Conference on Computing, Communication and Networking Technologies

[18] S.Suresh Shastri, Priyanka Vivek, Dr. Deepa Gupta, Nayar, R. C., Rao, R., and Ram, A., "Breast Cancer Diagnosis and Prognosis using Machine Learning Techniques", in International Symposium on Intelligent Systems Technologies and Applications (ISTA'17), Manipal University, Karnataka, 2017

[19] Dr. Deepa Gupta, Khare, S., and Aggarwal, A., "A method to predict diagnostic codes for chronic diseases using machine learning techniques", in 2016 International Conference on Computing, Communication and Automation (ICCCA), 2016

[20] L. Kolasani and Dr. Supriya M., "Parallelized Heat Map Algorithm Using Multiple Cores", in International Conference on Data Science, Machine Learning & Applications (ICDSMLA 2019), CMR Institute of Technology, Hyderabad, India , 2020.

[21] R. R. Nair, Karumanchi, S. H., and Dr. Tripty Singh, "Neuro-Fuzzy based Multimodal Medical Image Fusion", 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). IEEE, 2020.