BIOINFORMATION
Discovery at the interface of physical and biological sciences

OPEN ACCESS GOLD

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 16(5)

Research Article

A report on DNA sequence determinants in gene expression

Ravail Singh[1,#], & Yengkhom Sophiarani[2,#]

[1]Indian Institute of Integrative Medicine, CSIR, Canal Road, Jammu-180001; [2]Department of Biotechnology, Assam University, Silchar-788011, Assam, India; Ravail Singh - Email: ravailsin91@gmail.com; rubail2010@googlemail.com; Yengkhom Sophiarani- Email: Sophia.yengkhom90@gmail.com; #Equal contribution; *Corresponding author: Ravail Singh

Received April 23, 2020; Accepted April 24, 2020; Published May 31, 2020

DOI: 10.6026/97320630016422

Declaration on Publication Ethics:
The authors state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Declaration on official E-mail:
The corresponding author declares that official e-mail from their institution is not available for all authors

Abstract:
The biased usage of nucleotides in coding sequence and its correlation with gene expression has been observed in several studies. A complex set of interactions between genes and other components of the expression system determine the amount of proteins produced from coding sequences. It is known that the elongation rate of polypeptide chain is affected by both codon usage bias and specific amino acid compositional constraints. Therefore, it is of interest to review local DNA-sequence elements and other positional as well as combinatorial constraints that play significant role in gene expression.

Keywords: Gene expression; codon; amino acid; genome.

Background:

DNA is the building block of life. The DNA sequence of genes and genomes is the blueprint of the gene function. All the information related to hereditary and species evolution is contained in these macromolecules. DNA sequences form the basis for all bioinformatic analysis tools and resources. In bioinformatics, these sequences are analyzed using a wide range of analytical methods to discover genes, its features, function, structure, or evolution **[1].** The discovery of DNA sequencing technology has made available many sequence datasets from different genomes over the last three decades. Subsequently, it has become apparent that the nucleotide composition varies from genome to genome and is one of the major perplexing characteristics of each genome. DNA sequence analysis is widely used as a highly accurate and flexible tool for classifying and identifying organisms **[2].** All the nucleotide sequences present in a genome do not code for proteins. Protein coding sequences usually start with an initiator codon (usually AUG) and end with a terminator codon (either UAA or UGA or UAG) in standard genetic code **[2].** Any continuous stretch of DNA that has the tendency to encode a specific protein is firstly transcribed as a messenger RNA then translated into a protein. The protein translation machinery present within any organism reads the DNA sequence in groups of three nucleotides (codons), which eventually determine the outline of amino acids that will appear in the ultimate protein. Therefore, a single-stranded DNA and a double-stranded DNA have three and six distinct reading frames, respectively. An open reading frame (ORF) is a DNA sequence that has the potential to code for a protein. Out of the six ORFs, only one is used in translating a gene (in eukaryotes), and this is often the longest ORF. The nucleotide sequence present within each ORF provides the instruction for encoding different proteins. A specific region of the DNA sequence is destined to code a protein for a specific function. For example, variation in the flower color among different plants belonging to the very subfamily is due to the variation in a specific region of a gene, which determines the flower color. These variations of a particular gene are called alleles, which produce the alternative or *different forms* of one trait. Gene expression is a fundamental cellular process through which proteins are synthesized within all living cells using the nature gifted genetic codes. Gene expression is a complex biological process separated into several phages, including synthesis and processing of mRNA, export (in eukaryotes), translation, and decay regulated at a variety of steps, from DNA to RNA to protein. Four major steps are involved in regulating the rate of gene expression, (a) timing and rate of transcription initiation and elongation; (b) processing of transcripts; (c) rate of transcript degradation, and (d) post-transcriptional modification of transcripts.

The physiological state of a cell is determined by the absolute concentration of proteins. Maintaining the amount of proteins at a steady state is a crucial feature of controlling gene expression. The process of gene expression is far more complex in eukaryotes than prokaryotes and the concentration at which a particular protein is produced depends on various factors **[3]**. The unique DNA sequence assets present in each gene can have conspicuous effects on its expression. Various sequence determinants present within a gene and their role in expression (tissue specific), operate through complex pathways, and determine the differential regulation of gene expression during development and differentiation. Advances in computational biology and its genetic engineering applications have opened the doors for a systematic study of the compositional constraints that affect gene expression **(Figure. 1).** We here discuss the sequence determinants with some astonishing results emerged from extensive research done in the field of gene expression and genetic engineering.

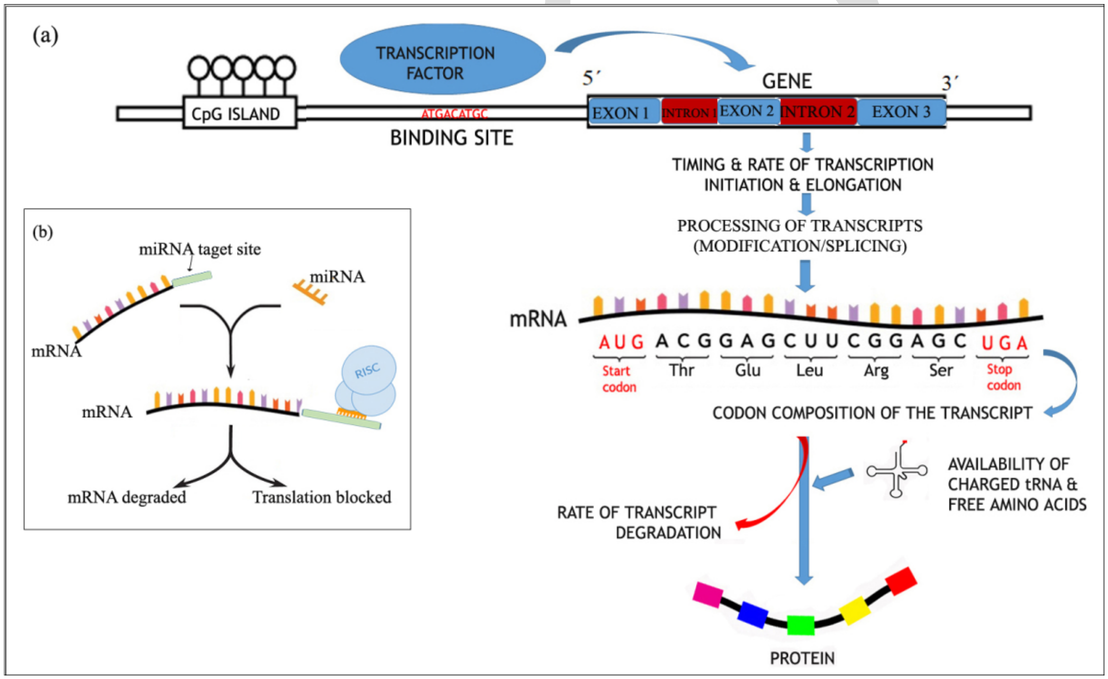
(a)
TRANSCRIPTION FACTOR
CpG ISLAND
ATGACATGC
BINDING SITE
5′
GENE
3′
EXON 1
INTRON 1
EXON 2
INTRON 2
EXON 3
TIMING & RATE OF TRANSCRIPTION INITIATION & ELONGATION
PROCESSING OF TRANSCRIPTS (MODIFICATION/SPLICING)
mRNA
AUG ACG GAG CUU CGG AGC UGA
Start codon
Thr
Glu
Leu
Arg
Ser
Stop codon
CODON COMPOSITION OF THE TRANSCRIPT
AVAILABILITY OF CHARGED tRNA & FREE AMINO ACIDS
RATE OF TRANSCRIPT DEGRADATION
PROTEIN
(b)
miRNA taget site
miRNA
mRNA
mRNA
RISC
mRNA degraded
Translation blocked

Figure 1: (a) Role of various factors that operate at the various stages of gene expression, (b) MicroRNA mediated gene silencing

Factors affecting gene expression level:
CpG Islands and transcription factors:
CpG islands are typically several hundred bases to several kilo bases long interspersed DNA sequences, rich in GC-rich, and predominantly non-methylated **[4].** CpG islands play a significant role in the regulation of genes by epigenetic changes **[5].** *In situ* hybridization on meta-phase chromosomes from blood lymphocytes revealed that the GC-rich genes localized in the GC-rich isochores express at a rate higher than others **[6].** The lengths of CpG islands vary and show relevant variation in their functions **[7].** CpG islands commonly found in the region where vertebrate genes begin transcription, perhaps in the case of all housekeeping genes and the genes expressed frequently in a cell. Approximately 70% promoters identified from the vertebrate genome were found to be associated with the CpG islands. CpG islands located over the transcription start point are known as start CpGs, whereas other islands are known as non-start CpGs. The promoter-associated CpG islands are structurally different from others **[8]**. CpG island regulates gene expression by influencing the local chromatin structure **[4]**, often lacks TATA boxes and displays heterogeneous transcriptional start sites enriched with transcription factor-binding motifs (E**2**F, Sp**1**, ETS, and Nrf-**1**) **[9]**. Methylation or demethylation of CpG islands results in repression or activation of gene expression, respectively. Altered CpG island methylation regulates gene expression by influencing the physical access of transcriptional factors at CpG islands [**1**0]. Several reports on CpG methylation showed that altered methylation at CpG attract several other methyl-CpG-binding-domain, which eventually recruits histone deacetylases (HDAC) and stops the transcription processes **[11]**. Previous reports showed that the regulatory CpGs (usually at the promoter region) on several protooncogenes/oncogenes are inappropriately methylated **[12]**. Antigens of the cancer-testis (CT) shows difference in methylation between normal and cancer cells, the former one has methylated CpGs at the CT antigens promoter region **[13]**. Long CpG island promoters are generally found in a few tissues. Furthermore, the genes associated with promoter CpG islands were more frequently expressed as compared to the genes without CpG islands **[7]**. Transcriptional factors and co-factors can vary for each gene. The difference in the sequence length of the island and their varied role in gene expression are thought to be due to the accessibility of the larger region for added transcriptional factors.

Transcription factors (TFs), control gene expression by binding to regulatory sequences, such as enhancer sequences, recruit co-activators and RNA polymerase to target genes. The biophysical interaction between DNA and protein structure determines the ability of the TFs to bind to a regulatory sequence **[14]**. TFs recognize unique cis-regulatory regions of the core promoter elements, which recruit the transcription machinery **[15]**. Many TFs are involved in the transcription of different promoters while some are very selective to a few promoters **[16]**. Their influence can be positive or negative, depends largely on the presence of several other functional domains and the overall impact of the entire TF complex. Differences in the concentration of TFs and co-factors influence the timing as well as the rate of transcription, thereby the expression of a gene **[17][18]**. For the mitochondrial genes of vertebrates the transcription factor A (mtTFA) plays the key role for the regulation of gene expression **[19]**. In the year **2**000, Batlle et al., identified that a transcription factor (snail protein) plays a crucial role in epithelial tumour progression by down-regulating the adhesion protein E-cadherin [**2**0]. They observed that the inhibition of this TF restores the expression of the E-cadherin gene. GATA-**3** is essential for Th**2** cytokine gene expression in CD**4** T cells, a major target for modifying the immune responses in many immunological conditions **[21]**. Under dehydration and low temperature, dehydration-responsive element binding protein **1** and **2** acts as *trans*-acting factors in Arabidopsis **[22]**. Another TF named NF-κB was found to be involved in human diseases gets activated by oncoproteins to induce cellular transformation. It has been reported that HIV and many other viruses induce NF-κB activation, because of its ability to regulate cell cycle, DNA replication, and apoptosis **[23]**. The transcription factor NKX**2**-**5** is needed for the maturation and maintenance of atrioventricular node function and any mutation in this TF encoding gene leads to congenital heart disease **[24]**. Therefore, the development of ligand binding approaches to address these transcription factors may provide a new biomedical avenue to treat or prevent diseases.

Variation in G+C content:
Genomic GC content varies significantly from small single cellular prokaryotes to multicellular eukaryotes due to two major factors: natural selection and mutation pressure **[25]**. Advancement in the field of sequencing technology and sequence-based analysis makes it possible to identify and compare genomes based on their sequence characteristics. The key cause of heterogeneity in GC content (location-specific) in prokaryotes is due to the presence of differentially expressed genes with varying level of nucleotide bias **[26]**. GC content variation between different genomic regions of the eukaryotes is an adaptation to higher body temperature, and it plays a fundamental role in genome organization. Some genomic regions of the eukaryotes are found to be very GC rich as well as gene rich, whereas some other regions are GC poor **[27]**. This variation in the nucleotide composition in different genomic region makes use of the codons with varying composition in different genes present at those particular genomic positions **[28]**.

Furthermore, several studies have demonstrated that GC content correlates with the gene expression level **[29]**. Newman *et al*., (2016) disclosed that the increased production of their studied proteins was not due to increased translation but for the change in transcription of GC-optimized codons **[**30**]**. This suggests that evolutionary force is responsible for maintaining the conserved genomic regions and for the GC rich codon usage of genes in eukaryotes. Guo *et al.* 2007, reported that the cereal crops, particularly maize provide a plausible evolutionary mechanism for genes with a strong GC bias **[31]**. The telomeric region shows an unusual pattern of GC substitution different from the rest of the genome. In addition to the role on gene expression level the genomic GC content influences many other important genomic features: (i) gene density; (ii) recombination rate; (iii) distribution of repetitive elements (SINE and LINE); (iv) replication timing; (v) methylation pattern, and (vi) distribution of the transposable elements **[32]**.

Codon usage bias:

Codon usage bias (CUB) is the phenomenon where some codons encoding a specific amino acid are used more preferentially over others. It is a complex process influenced by several factors like GC content, protein abundance, mRNA folding etc. A balance between mutation, selection, and genetic drift ensures the strength of CUB **[33]**. Molecular evolutionary investigations suggest that codon usage represents a characteristic pattern of preference in each organism, and it also differs within in a single genome at different locations **[34]**. The codon bias is the most prominent among highly expressed genes. The influence of codon bias on gene expression is a topic of active debate. Codon bias, first studied on the model organism *Escherichia coli (E. coli)*, revealed strong bias in codon usage for genes encoding abundant proteins. CUB causes variation in GC content, particularly at the synonymous third codon position. CUB has been studied extensively in the past two decades but the relative contribution of the two major determinants of codon usage bias *i.e.*, mutation and selection is still mysterious. Genetic engineering approach utilizes the information generated from CUB analysis to increase protein production. Research on CUB has shown its role in regulating protein expression by affecting elongation speed. However, recent studies have shown that translation efficiency is mainly affected by the efficiency of the initiation of translation, where CUB plays a minor role **[35]**. The approach of codon optimization used in the process of genetic engineering affects protein conformation, and function, and also increases immunogenicity, but reduces efficacy **[36]**. Heterologous protein expression for increased protein production is a multidimensional optimization problem **[37]**. Therefore, while optimizing codons in a gene, other factors like mRNA folding, bendability and stability should be taken into consideration, particularly for *in vivo* applications **[38]**. Furthermore, it was observed that the first 30-40 nucleotides of a gene show a different pattern of codon usage from the rest of the gene sequence **[39]**. Further, the anterior region of coding sequence undergoes selection pressure and regulates gene expression through mRNA folding during translation initiation **[**40**]**. Previous studies suggest that synonymous sites are functionally neutral, but some recent findings contradict it, *i.e.* synonymous mutations are associated with diseases **[41]**. Mendelian disease-causing synonymous SNPs, for example, has a comparable effect size as non-synonymous SNPs in association studies of human disease **[42]**. In addition, substantial associations have been identified in the case of Alzheimer's disease between genetic variants containing a favored codon in minor alleles and a rare codon in the major allele **[43]**. These results provide an insight into the roles of CUB in forming different protein structures and can be used in rare variant association research to boost detection efficiency.

mRNA folding energy:

The folding energy of an mRNA depends on the coding sequence of a gene. With the high folding score, mRNAs fold more strongly. Intra-molecular hydrogen bonds and base stacking interactions between nucleotide pairs determine the secondary structure of the mRNA molecule. Furthermore, the function of an mRNA closely resembles its structure. The secondary structure of mRNA can be predicted with the aid of bioinformatics, since mRNA molecules are more conserved in secondary structure than their primary base sequences **[44]**. Enzymatic methods can also be used for structural inferences **[45]**. A given mRNA may have a different structural conformations at different locations depending on the sequence properties. Furthermore, the mRNA folding pattern is environment-dependent **[46]**. Therefore, during the rapid cell growth a single gene might be folded somewhat differently to allow faster translation and elongation than steady state condition. The correlation between mRNA folding and translation is complicated and has opposite impacts **[47]**. Highly expressed mRNAs' folding structures undergo strong selection pressure to reduce ribosome sequestration for accelerated elongation. Researchers proposed that the folding of an mRNA influences ribosome binding and hence plays a crucial role in determining the gene expression level **[48]**. Selection pressure acts near the initiator region to decrease the folding energy, slow down the ribosomes and decrease translational efficiency. Kudla *et al*. 2009 analyzed the green fluorescent protein (GFP) from *E. coli* genome **[49]**. For the same gene they designed different coding sequences without altering the native amino acid composition. They observed that the folding energy of the initial 40 nucleotides from each mRNA showed strong correlation with the protein abundance values. Tuller *et al*., 2010 found a clear link between the genomic profile of the folding energy and the ribosome density demonstrated by the study of *S. cerevisiae* and *E. coli* transcriptomes [48]. This relationship means that highly structured mRNA holds back the velocity of ribosomal motion on mRNA, as the density of ribosomes is higher for lower ribosomal velocity assumed by constant ribosomal flux [50]. Genes having the tendency to express at high rate undergo strong selection pressure for stronger folding which results in the slow evolution of these genes [45]. These results altogether suggest that folding energy influences the global translation efficiency (translational initiation plus translation elongation). An error in protein folding and the accumulation of these misfolded proteins leads to amyloid diseases.

Gene length:

The length of genes varies from gene to gene relating to their proper folding and function. The average gene length is usually longer in eukaryotes than prokaryotes. Gene length is a dynamic property. Length of the gene increases during the evolutionary process due to transposable elements. Therefore, the increased length of the eukaryotic genes indicates the evolutionary complexity associated with gene length [51]. Eukaryotic genes are rich in introns. Introns are the sites where transposable elements are introduced and also provide a mechanism for generating transcriptional complexity within the multicellular genomes via the preferable mixture of exons. Researchers found that the genes expressed at high level are shorter in size and tend to have shorter introns [52]. Large scale analysis on prokaryotic genes proved the relationship between gene length and expression level. Long gene requires substantial time to be expressed following its activation, for example, the largest known gene human dystrophin (primary transcript of 2.3Mb) requires approximately 16h for transcription. Although the gene dystrophin has a longer gene length required for encoding the proper number of amino acid to make it fully functional, it seems to be under selection pressure since the gene dystrophin is very long but has fewer introns [53]. Gene lengths show significant correlations with both gene duplication (negative) and alternative splicing (positive) [52]. Grishkevich and Yanai (2014) have shown that the relationship between gene duplication and alternative splicing is regulated by two key genetic properties, *i.e.* gene length and level of expression [52]. In the cortical neurons, longer genes associated with neuronal development and synapses were down-regulated by topoisomerase inhibitors via impairing transcription elongation [54], might impair neuronal function and lead to neurological disorder [55]. In addition, recent studies showed gene-length mediated shift in the expression of Rett syndrome (RTT) neurodevelopmental disorder [56]. Transcriptional timing is inherently influenced by gene length, provides a mechanism for temporal regulation of gene expression [57]. A research on Drosophila has shown that the gene length mediates developmental timing of gene expression [58]. These findings altogether suggest that gene length is an important factor influencing virtually all aspects of molecular evolution.

Amino Acid Composition:

Proteins vary in their amino acid compositions, which depends largely on the locations where the proteins are destined to function. It is anticipated that natural selection could play a role in preserving or enhancing the protein activity, specificity or stability by favoring specific codons encoding corresponding amino acids at critical positions in the protein's primary structure. But in less constrained positions of the protein, a combination of both mutation pressure and genetic drift might act on the coding sequences to encode the specific amino acids in the protein [59]. Metabolic constraints on protein crystal structure include biosynthesis cost of amino acids, complexity of synthetic pathways, nutrients, protein synthesis accuracy and speed [59]. The correlations between the composition of amino acids and protein function are well documented for three lineages of life: prokaryotes, archaea and eukaryotes [60]. However, less attention was given to the relationship between the efficiency of protein biosynthesis and its primary structure. The degree to which the composition of amino acids is skewed to minimize metabolic costs ought to be a good measure of the amount of proteins synthesized from each gene per generation. Akashi *et al*., 2002 demonstrated that the frequency of certain amino acids varies in wide functional protein categories as a result of translation rate estimation [59]. The cost of amino acid biosynthesis ranges from 11.7 PO_4 for less complex amino acids (*eg:* glycine and proline) to 74 PO_4 for extremely complex amino acids (*eg*: tryptophan). The use of these less costly amino acids in highly expressed genes has an energetic benefit that can surpass 0.025% of the total energy expenditure. Costly amino acids namely Tryptophan, Phenylalanine, Histidine, Cysteine, and Leucine are found in less frequency in highly expressed genes, whereas the frequency of the less costly amino acids such as Glutamine, Asparagine, and Glycine are usually more in highly expressed genes. These results suggest the effect of natural selection to enhance metabolic efficiency by increasing the use of more costly amino acids in lowly expressed genes but the use of less costly amino acid in highly expressed genes, respectively. Moreover, a significant number of genes across all living species encodes proteins with amino acid repeats of different length and composition that play important role in overall protein structure and function [61]. In addition to its role as a substrate for protein synthesis, recently it was reviewed that amino acids in concert with hormones modulate various signal transduction pathways, which regulate mRNA translation **[62]**. The utilization of amino acids and its demand varies between healthy and disease conditions **[63]**. Any abnormality in the metabolic pathways of a specific amino acid leads to the accumulation of that amino acid, can evoke a toxicity syndrome which usually extends to central nervous syndrome (e.g. hypoglycemia) **[64]**.

tRNA abundance:

A tRNA (transfer ribonucleic acid) molecule typically contains **76** to **90** nucleotides that help decipher a mRNA sequence into a protein. Each tRNA recognizes specific amino acid and carries only one amino acid attached to its end at a time. When a tRNA binds to ribosome with its matching codon, it transfers a corresponding amino acid to the expanding polypeptide chain. The translation rate or the decoding rate of a codon depends on the speed of delivery of its translationally competent tRNA to the ribosome **[65]**. We all know about the redundancy of the genetic code *i.e.*, **61** codons code for **20** amino acids. Hence, there must be an equal number of tRNA molecules for each of these codons but there is a large variation in the tRNA gene copy number per cell by up to **10** fold **[66]**. Soon after the experimental documentation of the correlation between the codon usage bias and tRNA abundance researchers tried to find out the relationship between the codon adaptation and the gene expression level **[67]**. Extensive research on gene expression found strong correlation with codon adaptation. Highly expressed genes that are enriched with optimal codons are recognized by abundant tRNAs and translated faster than codons read by low-abundance tRNAs **[4**0]. The initial region at the **5′** end of a gene has somewhat different nucleotide composition pattern generally recognized by tRNA species with lower intracellular abundance and provides several physiological benefits **[68]**. Individual tRNA expression varies in different tissues, and tRNAs decoding amino acids with specific chemical properties showed structured expression in multiple tissue types. Tissue-specific expressed gene and coordinated expression of tRNAs implicate its function in controlling translation and probably secondary processes in mammals **[69]**. Goodarzi *et al.*, (**2016**) showed that specific tRNAs are up-regulated in human breast cancer cells as they gain metastatic activity **[7**0]. Gorochowski *et al.*, **2015** analyzed the role of tRNA abundance in mRNA folding and translation elongation **[71]**. They observed that the gene regions enriched with codons having more abundant tRNA has the propensity to form strong secondary structure. This structure eventually influences the translation elongation dynamics and enhances protein translation and leads to increased protein yield. Mutation in tRNA genes and its processing enzymes leads to a variety of complicated clinical phenotypes, for example, mutation in mitochondrial tRNA (mt-tRNA) causes mitochondrial myopathies **[72]**.

Presence of the correct 5′- cap and poly (A) tail in 3′-end region and the role of miRNA:

Soon after the post-transcriptional modification, all mRNA (except the replication dependent histone transcript) molecules in eukaryotes acquire a **5**′- cap (m^7GpppN) and a poly (A) tail at their **3**′-ends. The **5**'-cap structure in eukaryotes regulates the overall quality of mRNA products by inducing the translation activation frequency **[73]**. The mRNA poly (A) tail at the **3**′-end region has a profound effect on mRNA bendability, translation rate, cell viability, growth, and development **[74]**. During the process of translation, synergistic effect of the poly (A) tail and the **5**′- cap of the mRNA direct the ribosome to bind to the initiator region. The **5′** m**7**G cap of eukaryotic mRNA recruits cellular proteins and mediates cap-related biological functions. Decades of research have established the importance of a proper cap structure for the optimal translation of functional messenger RNA **[75]**. It is an important regulation point of gene expression that protects mRNA from degradation, promotes transcription and nuclear export **[76]**. Researchers working in a cell free translation system revealed that poly (A) tail independently promotes the binding of the small ribosomal subunit **[77]**. Thus, poly (A) tails are also known as translational enhancer. Polyadenylation signals (PAS) are often considered as a distinguishing characteristic of eukaryotic genes. A highly conserved motif AAUAAA is present in almost all eukaryotic polyadenylated mRNAs found **10** to **30** nucleotides upstream of the cleavage site, essential for both cleavage and poly (A) addition as well as for promoting downstream transcriptional termination **[78]**. Approximately **30** to **40** nt downstream of the AAUAAA motif there is another additional region (less conserved, either U or GU rich or both) and the distance between the AAUAAA motif and these additional motifs determines the cleavage site *i.e.* site of poly (A) addition. Preiss and Hentze, **1998**, showed that independently both the **5**′-cap and the **3**′-tail can promote translation but not enough to promote efficient translation **[79]**. These results suggested the need of proper cap and tail (closed –loop model) for efficient translation of an mRNA molecule.

MicroRNAs are short non-coding RNA molecules, which regulate gene expression at the transcription and post transcription level, generally bind to their target mRNAs **3** prime untranslated region. Recently, the structure and functions of this essential intracellular genomic regulator have been highlighted. MicroRNA binds with mRNA and inhibits protein translation or destabilizes target transcript. The length of the **3** 'UTRs decides the density of miRNAs binding to the mRNAs **[8**0]. Extensive research on miRNA revealed that the seed region is responsible for target recognition, which pairs fully with the target region **[81]**. They play crucial role in key developmental processes by regulating the expression of some important genes. Recently, it was observed that animal miRNAs show minimal sequence complementarity with the target sequence; thereby a single miRNA can possibly interact with many genes with similar sequence composition **[82]**. They are assembled with Argonaute into multiprotein effector complexes, called RNA-induced silencing complexes (RISCs) **[83]**. MicroRNA can upregulate and downregulate the expression of a gene and in some specific conditions a single gene could encounter both regulation direction **[84]**. Human miR-**373** was the first miRNA to be identified as an activator of gene expression **[85]**. Corresponding work showed that miRNAs have extensive gene regulatory mechanisms **[86-87]**. Similarly, several other researches showed the inhibition / downregulation of gene expression by miRNA through perfect binding with their target genes **[88]**. MicroRNA mediated regulation of gene expression is selective, specific to the sequence, and depends on the miRNP factors and other RNA binding proteins **[89]**.

Conclusion:

Transcription, mRNA folding, CpG islands, translation, gene length, GC composition, codon usage bias, amino acid composition and tRNA abundance are essential processes during eukaryotic gene expression, but their relative global contributions to steady-state protein concentrations in multi-cellular eukaryotes are largely unknown. These factors influence gene expression through their interaction with cellular machinery either individually or in combination of these sequence-based factors. Sequence features alone can explain >**5**0% of protein abundance variation. Therefore, while optimizing a gene sequence for heterologous expression a single base pair change can show high degree of co-variation and complex interdependence. Several of these features have hardly been used in synthetic gene design and require more attention in future attempts. Hence, a systematic assessment of all relevant variables is essential to ensure the desired level of protein production. We document factors that need to be examined in details for increasing gene expression in eukaryotes. An understanding of the intricate relationships of the factors in a coordinated approach through establishment of protein expression system is relevant. Targeted study of these constraints in specific disease condition will certainly give novel insights for gene therapies and make significant innovations to ameliorate the specific diseases.

Acknowledgments: We are thankful to Indian Institute of Integrative Medicine, CSIR, Canal Road, Jammu-**1**80001 for providing the necessary facilities to carry out the review work**.**

Conflicts of interest:

The authors declare that no conflict of interest exists for this work.

References:

[1] Klasberg S *et al. Bioinform Biol Insight*, 2016 **10:**121. [PMID: 27493475]
[2] Song Y *et al. Journal of clinical microbiology* 2003 **41**:1363. [PMID: 12682115]
[3] Haimovich, G., *et al*.,. *Cell*, 2013 **153(5)**:**1**000. [PMID: 23706738]
[4] Deaton, A.M. and A. Bird,. *Genes & development*, 2011 **25**(**10**): **1**010. [PMID: 21576262]
[5] Wachter, E., *et al*.,. *Elife*, 2014. **3**: e03397. [PMID: 25259796]
[6] Arhondakis, S., *et al*., *Gene*, 2004. **325**: 165. [PMID: 14697521]
[7] Elango, N. and V.Y. Soojin, *Genetics*, 2011: **11**0.126094. [PMID: 21288871]
[8] Ponger, L., L. Duret, and D. *Genome research*, 2001. **11**(**11**): 1854. [PMID: 11691850]
[9] Landolin, J.M., *et al. Genome research*, **2**010. [PMID: 20501695]
[10] Watt, F. and P.L. Molloy, *Genes & development*, 1988. **2(9):**1136. [PMID: 3192075**]**
[11] Long, M.D et al, *Biomolecules*, 2017. **7**(**1**):15. [PMID: 28216563]
[12] Baylin, S.B., *et al. Cancer research*, 1986. **46**(**6**): 2917. [PMID: **3**009002]
[13] Meklat, F., *et al. British journal of haematology*, 2007. **136**(**6**): 769. [PMID: 17223912]
[14] Todeschini, A.-L et al, *Trends in genetics*, 2014. **30**(**6**): 211. [PMID: 24774859]
[15] Maricque, B.B et al, *Nucleic Acids Res*, 2017. **45**(**4**): e16. [PMID: 28204611]
[16] Wuttke, D.S., et al, *J Mol Biol*, 1997. **273**(**1**): 183. [PMID: 9367756]
[17] Lee, T.I. and R.A. Young, *Cell, 2013*. **152**(6): 1237. [PMID: 23498934]
[18] Cheng, C., *et al*., *Genome research*, 2012. **22**(**9**): 1658. [PMID: 22955978]
[19] Virbasius, J.V. and R.C. Scarpulla, *Proceedings of the National Academy of Sciences*, 1994. **91**(**4**):1309. [PMID: 8108407]
[20] Batlle, E., *et al*., *Nature cell biology*, **2**000. **2**(**2**): **84**. [PMID: 10655587]
[21] Zheng, W.-p. and R.A. Flavell, *Cell*, 1997. **89**(**4**): 587. [PMID: 9160750]

[22] Liu, Q., *et al.*, *The Plant Cell*, ***1998***. **10**(**8**):1391. [PMID: 9707537]

[23] Baldwin, A.S., *The Journal of clinical investigation*, 2001. **107**(**1**): **3**. [PMID: 11134170]

[24] Schott, J.-J., *et al.*, *Science*, 1998. **281**(**5373**):**108**. [PMID**:** 9651244]

[25] Hildebrand, F et al, *PLoS genetics*, 2010. **6**(**9**): e1001107**.** [PMID: 20838593]

[26] Kudla, G et al, *Molecular biology and evolution*, 2004. **21**(**7**): 1438. [PMID: 15084682]

[27] Lander, E.S., *et al.*, *Nature*, 2001. **409**(**6822**): 860. [PMID: 11237011]

[28] Vetsigian, K. and N. Goldenfeld, *Proceedings of the National Academy of Sciences*, 2008: p. pnas. 0810122106. [PMID: 19116280]

[29] Song, H., *et al.*, *Scientific reports*, 2017. **7**(**1**): p. 14853. [PMID: 29093502]

[30] Newman, Z.R., *et al.*, *Proceedings of the National Academy of Sciences*, 2016. **113**(10): p. E1362. [PMID: 26903634]

[31] Guo, X et al, *FEBS letters*, 2007. **581**(**5**): 1015. [PMID: 17306258]

[32] Glémin, S., *et al.*, *Trends in Genetics*, 2014. **30**(**7**): 263. [PMID: 24916172]

[33] Trotta, E., *Nucleic acids research*, 2013. **41**(**20**): 9382. [PMID: 23945943]

[34] Butt, A.M., *et al.*, *Emerging microbes & infections*, 2016. **5**(10): **e107**. [PMID: 27729643]

[35] Zhao, F et al, *Nucleic acids research*, 2017. **45**(**14**): 8484. [PMID: 28582582]

[36] Mauro, V.P. and S.A. Chappell, *Trends in molecular medicine*, 2014. **20**(**11**): **604**. [PMID: 25263172]

[37] Schlegel, S., *et al.*, *Microbial cell factories*, 2013. **12**(**1**): **24**. [PMID: 23497240]

[38] Hancock, S.P., *et al.*, *PLoS One*, 2016. **11**(**3**): p. e0150189. [PMID: 26959646]

[39] Tuller, T. and H. Zur, *Nucleic acids research*, 2014. **43**(**1**): 13. [PMID: 25505165]

[40] Quax, T.E., *et al.*, *Molecular cell*, 2015. **59**(**2**): 149. [PMID: 26186290]

[41] Im, E.-H. and S.S. Choi, *Genomics & informatics*, 2017. **15**(**4**): 123. [PMID: 29307137]

[42] Li, M.-X., *et al.*, *PLoS genetics*, 2013. **9**(**1**): p. e1003143. [PMID: 23341771]

[43] Miller, J.E., *et al. Pacific Symposium on Biocomputing*. **2018**: World Scientific. [PMID: 29218897]

[44] Mathews, D.H *et al. Cold Spring Harbor perspectives in biology*, **2010**: p. a003665. [PMID: 20685845]

[45] Park, C., *et al.*, *Proceedings of the National Academy of Sciences*, ***2013***. **110**(**8**): E**678**. [PMID: 23382244]

[46] Cristofari, G. and J.-L. Darlix, *Prog Nucleic Acid Res Mol Biol*. **2002**. [PMID: 12206453**]**

[47] Faure, G., *et al.*, *Nucleic acids research*, 2016. **44**(**22**):10898. [PMID: 27466388]

[48] Tuller, T., *et al.*, *Proceedings of the National Academy of Sciences*, 2010. **107**(**8**): 3645. [PMID: 20133581]

[49] Kudla, G., *et al.*, *science*, 2009. **324**(**5924**): 255. [PMID: 19359587]

[50] Ingolia, N.T., *et al.*, *science*, 2009. **324**(**5924**): 218. [PMID: 19213877]

[51] Canapa, A., *et al.*, *Cytogenetic and genome research*, 2015. **147**(**4**): 217. [PMID: 26967166]

[52] Grishkevich, V. and I. Yanai, *Genome research*, 2014. **24**(9):1497. [PMID: 25015383]

[53] Jeffares, D.C et al, *Trends in genetics*, 2008. **24**(**8**): 375. [PMID: 18586348]

[54] King, I.F., *et al.*, *Nature*, 2013. **501**(**7465**): 58. [PMID: 23995680]

[55] Zylka, M.J et al, *Neuron*, 2015. **86**(**2**): 353. [PMID: 25905808]

[56] Barbash, S. and T.P. Sakmar, *Scientific reports*, 2017. **7**(**1**): 190. [PMID: 28298623]

[57] Kirkconnell, K.S., *et al.*, *Cell Cycle*, 2017. **16**(**3**): **259**. [PMID: 28055303]

[58] Artieri, C.G. and H.B. Fraser, *Molecular biology and evolution*, 2014. **31**(**11**): 2879. [PMID: 25069653]

[59] Akashi, H. and T. Gojobori, *Proceedings of the National Academy of Sciences*, 2002. **99**(**6**):3695. [PMID: 11904428]

[60] Williford, A. and J.P. Demuth, *Molecular biology and evolution*, 2012. **29**(**12**):3755. [PMID: 22826459]

[61] Barik, S., *Heliyon*, 2017. **3**(**12**): p. e00492. [PMID: 29387823]

[62] Kimball, S.R. and L.S. Jefferson, *Nutrition & metabolism*, 2004. **1**(**1**): p. **3**. [PMID: 15507151]

[63] Soeters, P.B., *et al.*, *The Journal of nutrition*, 2004. **134**(**6**): 1575S. [PMID: 15173433]

[64] Siegel, G.J., *Basic neurochemistry: molecular, cellular and medical aspects*. **1999**. https://www.ncbi.nlm.nih.gov/books/NBK**20385**/

[65] Kapur, M. et al, *Neuron*, 2017. **96**(**3**): 616. [PMID: 29096076]

[66] Du, M.-Z., *et al.*, *DNA Research*, 2017. **24**(**6**):623. [PMID: 28992099]

[67] Moriyama, E.N. and J.R. Powell, *Journal of molecular evolution*, 1997. **45**(**5**): 514. [PMID: 9342399]

[68] Misawa, K. and R.F. Kikuno, *BMC research notes*, 2011. **4**(**1**): p. **20**. [PMID: 21272306]

[69] Dittmar, K.A et al, *PLoS genetics*, 2006. **2**(**12**): p. e221. [PMID: 17194224]

[70] Goodarzi, H., *et al.*, *Cell*, 2016. **165**(**6**): 1416. [PMID: 27259150]

[71] Gorochowski, T.E., *et al.*, *Nucleic acids research*, 2015. **43**(**6**): **3**022. [PMID: 25765653]

[72] Abbott, J.A et al, *Frontiers in genetics*, 2014. **5**:**158**. [PMID: 24917879]

[73] Babendure, J.R., *et al.*, *Rna*, 2006. **12**(**5**): 851. [PMID: 16540693]

[74] Jalkanen, A.L *et al. Seminars in cell & developmental biology*. 2014: [PMID: 24910447]

[75] Ramanathan, A et al, *Nucleic acids research*, 2016. **44**(**16**): 7511. [PMID: 27317694]

[76] Topisirovic, I., *et al.*, *Wiley Interdisciplinary Reviews*: *RNA*, 2011. **2**(**2**): 277. [PMID: 21957010]

[77] Jackson, R.J et al, *Nature reviews Molecular cell biology*, 2010. **11**(**2**): 113. [PMID: 20094052]

[78] Laishram, R.S., *FEBS letters***,** 2014**. 588**(**14**): 2185. [PMID: 24873880]

[79] Preiss, T. and M.W. Hentze, *Nature,* 1998. **392**(**6675**): 516. [PMID: 9548259]

[80] Cheng, C et al, BMC *genomics*, 2009. **10**(**1**): 431. [PMID: 19751524]

[81] Catalanotto, C et al, *International journal of molecular sciences*, 2016. **17**(**10**): 1712. [PMID: 27754357]

[82] Shivdasani, R.A., *Blood*, 2006. **108**(**12**): 3646. [PMID: 16882713]

[83] Valinezhad Orang et al, *International journal of genomics*, 2014. [PMID: 25180174]

[84] Cordes, K.R., *et al.*, *Nature*, 2009. **460**(**7256**): p. 705. [PMID: 19578358]

[85] Place, R.F., *et al.*, *Proceedings of the National Academy of Sciences*, 2008. **105**(**5**): 1608. [PMID: 18227514]

[86] Matsui, M., *et al.*, *Nucleic acids research*, 2013. **41**(**22**): 10086. [PMID: 23999091]

[87] Zhang, Y., *et al.*, *Retrovirology*, 2014. **11**(**1**): **23**. [PMID: 24620741]

[88] Miao, L., *et al.*, *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 2016. **1859**(**4**): 650. [PMID: 26926595]

[89] Vasudevan, S et al, *Science*, 2007. **318**(**5858**): 1931. [PMID: 18048652]

Edited by P Kangueane

Citation: Singh & Sophiarani, Bioinformation 16(5): 422-431 (2020)