Acta Scientific MEDICAL SCIENCES (ISSN: 2582-0931)

Volume 8 Issue 10 October 2024

Research Article

Correlation between Eukaryotic Noncoding QED Genetic Codes and Cis-Regulatory Elements

Rama Shankar Singh*

Gen-Med R&D, LLC, Orlando, FL 32817, USA

***Corresponding Author:** Rama Shankar Singh, Gen-Med R&D, LLC, Orlando, FL 32817, USA. Email: gen-med@comcast.net

DOI: 10.31080/ASMS.2024.08.1929

Received: July 25, 2024

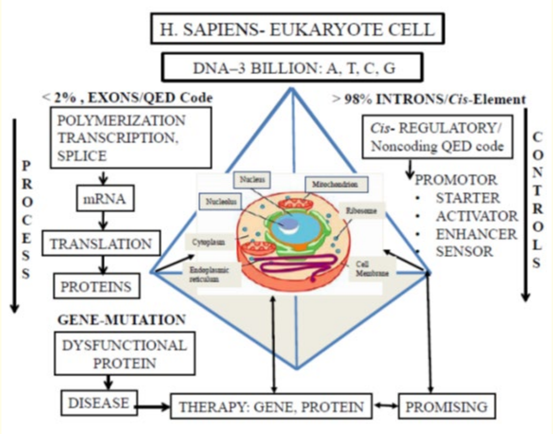
Published: September 17, 2024

© All rights are reserved by **Rama Shankar Singh.**

Abstract

The eukaryotic genome contains protein-encoding genes and cis-regulatory elements in the promoter regions of genes that control protein synthesis to maintain cellular homeostasis. Gene variants and protein synthesis errors produce dysfunctional proteins that cause disease. Thus, the genetic code and regulatory elements are closely linked to diseases. Eukaryotic nuclei contain the material responsible for heredity (DNA), but protein synthesis occurs elsewhere; genetic information is transmitted via transcription and splicing to the cytoplasm prior to protein synthesis. Cis-regulatory elements are found in eukaryotes but not in prokaryotes. The triplet genetic coding system proposed in 1963 was based on cells lacking introns, as introns were unknown until 1977, and lacking the control systems found in eukaryotes. The eukaryotic QED (quadruplet expanded DNA) genetic code model [1] is the first and only distinct genetic model for eukaryotes. The QED code has twenty independent encoding codons and thirty-five independent noncoding codons. Among the thirty-five noncoding QED codons are TATA, (CG) (CG), GC-rich, YCAY (Y-T or C), UAGG, GCAU, AT-rich, CG-rich, ATCG, START, and STOP. Here, we establish the applicable regulatory roles of noncoding QED codons like cis-regulatory elements by comparing their sequences. Some cis-regulatory elements, such as TATA and CAAT, located in the upstream promotor region are known to be involved in the initiation of transcription. These bases are identical to QED noncoding bases. More than twenty cis-regulatory elements and QED noncoding codon bases closely coincide and are listed in the text. Thus, it is reasonable to accept the application of the QED noncoding codon system to eukaryotic regulatory elements. Consequently, the noncoding QED code provides opportunities to correct dysfunctional proteins and identify cures for human diseases. A case in point is the tandem repeat (TR) neurodegenerative Huntington's disease. The triplet codons CAA and CAG encode glutamine (Gln); however, only TR CAG causes Huntington's disease without forming polyglutamine. Huntington's disease is explained by the QED code, in which CAA encodes Gln, but CAG is noncoding and does not promote polyglutamine formation but causes the disease.

Keywords: QED Code; Eukaryote; Prokaryote; Encoding; Noncoding; Cis-Elements; TATA; CAAT; CAG; Tandem Repeat; Human Diseases



Graphical abstract

Introduction

Humans contain more than one trillion eukaryotic cells, which each contain DNA heredity material and three billion (A, T, C, and G) bases. The genetic code translates eukaryotic cell genes into proteins for maintaining homeostasis. In 1963, triplet gene coding was proposed [2] for translating a gene lacking introns in a prokaryote. In 1977, Split Gene Discovery [3,4] established the existence of introns and found that genes are not distributed continuously. Furthermore, out of three billion DNA bases, less than 2% encode proteins (exons), and more than 98% are noncoding (introns). The introns were previously considered junk but are now described as a cis-regulatory element in the upstream promoter area that controls polymerization during protein synthesis. In eukaryotes, protein synthesis occurs in the cytoplasm rather than the nucleus. Several processes are needed to generate mRNAs in the cytoplasm for protein synthesis at the ribosome, and these processes include unique transcription [5] for selecting genes, splicing [6] for separating exons from introns, and alternate splicing [7] for synthesizing multiple genes. The mRNA is translated into proteins at the ribosome in both eukaryotes and prokaryotes. No special processing is needed for prokaryotes because their DNA is already in the cytoplasm. In prokaryotes, gene control is accomplished by operons [8,9]. Viruses contain reverse transcriptase rather than DNA, as bacteria use the host's tool to synthesize proteins and cause disease.

Cis-regulatory elements initiate and control eukaryotic protein synthesis and gene expression [10]. Gene mutations, transcription errors [11], splicing errors [12], exon/intron interface errors [13] and tandem repeats (TRs) [14-16] can lead to dysfunctional proteins that cause disease. There are more than 7000 monogenic rare diseases that cannot be cured, and patients with these diseases can only manage their symptoms. For example, multigene cancers cannot be cured, and the goal in treating these cancers is to extend patients' lifespans by five years; in addition, treatments for the least understood neurodegenerative disease remain at the early stage. Consequently, eukaryotic cis-regulatory elements are anticipated to serve as a method to correct dysfunctional proteins and find cures for diseases.

The triplet genetic code only contains START and STOP controls. Without eukaryotic genetic codes, using the triplet gene code to establish a cis-regulatory element road map [17] or to search for its encoding in DNA [18-20] is similar to force–fitting a square peg into a triangular hole. Approximately ten generic classes of promoters that contain AT-rich, CG-rich and ATCG-rich promoters have been reported [21], which were found through analysis of tissue-specific gene expression and regulation data obtained for Arabidopsis thaliana, Drosophila melanogaster, Homo sapiens, and Oryza sativa using (TiGER) [22]. The first and only eukaryote QED genetic code model [1] developed with twenty encoding codons and thirty-five noncoding codons exhibits these attributes and has a high degree of commonality with cis-regulatory elements.

Method

Prokaryote and eukaryotic protein synthesis

Cis-regulatory elements control protein synthesis in eukaryotic cells. As the nuclei contain DNA heredity material and require gene selection, complex transcription and alternate splicing are needed to transport mRNA in the cytoplasm and synthesize multiple proteins at the ribosome. Thus, one gene can encode multiple proteins in eukaryotes. In prokaryotes, the protein synthesis of either one or multiple genes is continuous due to a lack of gene selection control.

Prokaryote protein synthesis

During 1961-63, F.H.C. Crick introduced the concept of central dogma of biology (DNA is transcribed into mRNA, which is translated to proteins) and proposed the triplet genetic code [2].

The sixty-four triple codes (4x4x4) were verified and accepted, as the 1968 Nobel Prize in Medicine [23] was awarded to R.W. Halley, H. G. Khorana, and M.W. Nirenberg. The verified triplet code has one START, three STOPS, and 60 protein-encoding degenerate codons, in which multiple codons encode the same amino acids and one tRNA encodes several amino acids.

In 1963, the existence of introns was unknown until the “Split Gene” was discovered in 1977 [3]; subsequently, the 1993 Medicine Nobel Prize was awarded to Richard J. Roberts and Phillip A. Sharp [4]. Thus, triplet coding applies to cells lacking introns, prokaryotes, and genes with one uninterrupted protein synthesis, as illustrated in Figure 1(a).

Viruses violate Crick’s proposed triplet coding sequence (DNA is transcribed into mRNA and translated into proteins) because viruses begin the process with mRNA followed by reverse transcriptase to cDNA to mRNA and proteins. The process was recognized, as the 1975 Medicine Nobel Prize was awarded to David Baltimore, Renato Dulbecco, and Howard Martin Temin [24]. Virus protein synthesis is illustrated in Figure 1(b).

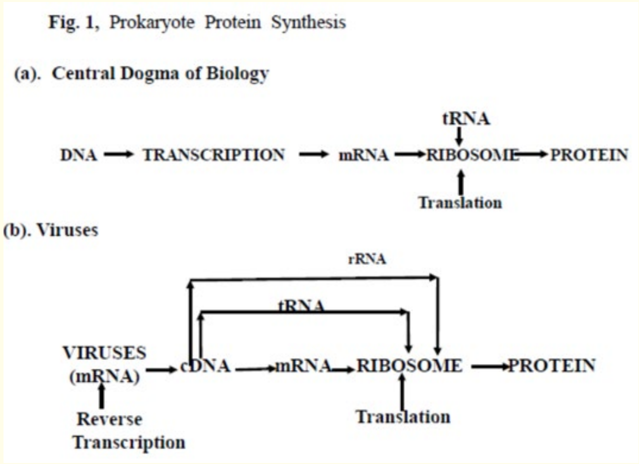


Figure 1

Prokaryotes’ signals are induced by the environment and metabolic processes, requiring gene control for multiple continuous gene protein syntheses; however, triplet coding lacks this information. In 1961, F. Jacob and J. Monod developed gene regulatory elements [8] called operons to control protein synthesis. The method was recognized when the 1965 Medicine Nobel Prize was awarded to F. Jacob, A. Lwoff, and J. Monod [9]. The operons have a regulator, repressor, operator, and promotor to control the polymerization and protein synthesis of genes. When a metabolite signal is received, either (present) ON or (absent) OFF, the regulator initiates the process by generating a repressor protein [25]; this protein combines with the metabolite molecule to turn the operator ON or OFF, allowing the promoter to initiate polymerization for protein synthesis.

Two well-known operons are lac and trp. The lac operon default is negative, i.e., when the metabolite signal is absent, the repressor blocks the operator (OFF) and promotor, and no protein synthesis occurs. When the metabolite lactose signal is present, the regulator generates a repressor, and lactose turn the operator ON, allowing the promotor to initiate the polymerization process and synthesize the three enzymes. The trp operon default is positive, i.e., when a metabolite signal is present, the regulator-generated repressor and metabolite keep the operator and promotor blocked (OFF), and no tryptophan synthesis occurs. When the metabolite tryptophan signal is absent, the generated repressor has no effect. Thus, the operator remains ON for polymerization and five enzymes used for tryptophan production are synthesized. The operon operation is binary (ON or OFF), and enzyme production is uninterrupted. The enzyme production of the lac and trp operons is shown in Figure 2 lac and trp operon gene control.

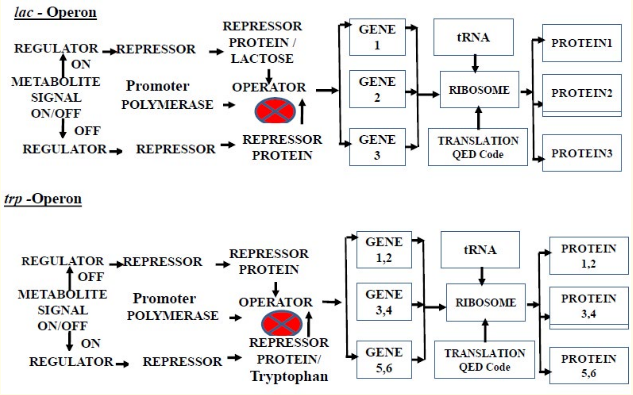


Figure 2

Eukaryote protein synthesis

Human species contain more than a trillion eukaryotic cells, and each cell contains heredity DNA material inside its nucleus, making protein synthesis complex. Furthermore, cell-to-cell communication signals transduce the timing and location of a specific cell for protein synthesis. Transcription and splicing are needed for protein synthesis to transport mRNA from the inside of the nucleus to the cytoplasm. Roger Kornberg [5] elucidated the detailed transcription process involving cis-regulatory elements in the promoter region, starting with the TATA box, transcription factor binding proteins, mediators, promotors, activators, and other controlling elements; these elements were used to control RNA polymerization and synthesize the protein. The 2006 Chemistry Nobel Prize was awarded to Kornberg for his "fundamental studies on the molecular basis of eukaryotic transcription".

The eukaryotic splicing process was well established [4,6,7] by 2010.

Figure 3 shows the synthesis of eukaryotic proteins with noncoding codons, such as TATA, AT-rich, CG-rich, CAAT, and ATCG, in the upstream promotor area, such as ACTIVATOR, ENHANCER, REPRESSOR, and SENSOR.

Eukaryotic protein synthesis is not a binary process and is triggered by cell–cell communication and the needs of specific cells. The noncoding QED eukaryotic code contains nearly all the cis-regulatory elements listed in Figure 3.

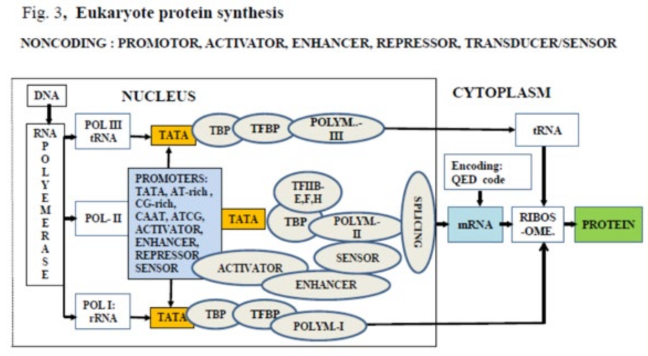


Figure 3

Common bases between cis-regulatory elements and the noncoding eukaryotic QED (quadruplet expanded DNA) code [1].

The key features of the QED model are the following: 1. All four DNA (A, T, C, and G) bases are involved; 2. The base positions are independent-symmetric; 3. The self-complementarity forming adjacent base pairs with any two adjacent NN (A, T, C, or G) bases is noncoding. The third feature extends Nobel Laureate Khorana's observation of the noncoding attributes of two adjacent bases forming natural pairs (poly-rAU) to quadruplets, leading to twenty independent encoding and thirty-five independent noncoding codons.

All cis-regulatory elements should be noncoding. The noncoding QED code meets these criteria and is restated here from [1] Table 1(b) in Table 1. The rows and columns of thirty-five noncoding QED codons are the same, the red font was changed to italics, and U was replaced by T for rows 11 to 35. The subscript p was also dropped for each case for convenience; for example, (TA) (TA) is (TPAP) (TPAP) and P a subscript.

The protein-encoding processes in the QED and triplet code are similar. Since the triple code has only two translational control elements, START and STOP, the prediction of QED START and STOP noncoding codons was done using the triplet START and STOP codes as a guide in [1], and are listed in Table 1.

Cis-regulatory elements and eukaryotic noncoding QED codon bases have a high degree of coincidence in eukaryotic transcription and splicing.

The cis-regulatory elements in the eukaryotic promoter region have been observed [5] to start, activate, enhance, and sense/moderate to control transactions and splicing processes in the nucleus; this process transports mRNA to the cytoplasm for protein synthesis at the ribosome. Whether these cis-regularity bases are noncoding has yet to be established. However, the eukaryotic noncoding QED code model meets the necessary conditions.

The cis-regulatory elements and noncoding QED bases coincidence are discussed and agreements are summarized in Table 2.

Transcription

Roger Kornberg [5] established the transcription process in eukaryotes that starts with a TATA box, several transcription factor-binding proteins, mediators, promotors, activators, and a moderator or a sensor. He was awarded the 2006 Chemistry Nobel Prize for his "fundamental studies of the molecular basis of eukaryotic transcription." The TATA box in the promoter region initiates the transcription process.

TATA Box: The TATA Box is located -30 bp upstream of the promoter to the start of the eukaryotic transcription process. Thus, the first noncoding QED code (TA) (TA) element in Table 1 can reasonably be verified and assigned under the QED code column for the start of the transcription process like start cis-element TATA. The coincidence is listed in Table 2.

CAAT Box: CCAAT is located -75-85 bp upstream of the promoter region and is referred to as a CAAT Box before the eukaryotic transcription process begins. The 11th element of Table 1, (AC) (TA) ≡ (CA) (AT) per rules #2 and # 3 of the QED model, meets this criterion and is assigned as a CAAT box in Table 1. The coincidences between cis-element and noncoding QED code are listed in Table 2.

Splicing

Splicing is a critical pre-mRNA process that occurs after transcription in the nucleus. The cis-regulatory elements in the promoter region also regulate the splicing process [6]. However, the process is more complex than transcription due to different exon/intron (3'-exon end/5'-intron start) and intron/exon (3'-intron end/5'-exon start) interfaces. Additionally, introns must be separated to transport exons (mRNAs) from the nucleus to the cytoplasm for protein synthesis. Furthermore, alternate splicing [7] allows multiple genes to be synthesized from one gene. The triplet coding gene has only START and STOP controls and cannot control transcription, splicing, RNA binding proteins, activation, enhancers or suppressors, and sensors needed for eukaryotes. The noncoding QED eukaryote codes can meet these requirements.

CG and GC-rich regions occur at both exon/intron and intron/exon interfaces and promote the splicing process. The second noncoding QED code (CG) (CG) has features enabling the splicing process, which is shown in Table 1. The enhancers and suppressors are located at exon/intron and intron/exon interfaces. GC-rich and YCAY (Y-T or C), i.e., TCAT, CCAT, TCAC, CCAC, UAGG, UGCAUG, and others, have been reported as enhancers and suppressors [6].

Approximately ten generic classes of promoters consisting of AT-rich, CG-rich and ATCG-rich promoters have been reported [21] through the analysis of tissue-specific gene expression and regulation data obtained for Arabidopsis thaliana, Drosophila melanogaster, Homo sapiens, and Oryza sativa using (TiGER) [22].

Based on the above-reported results, thirty-five noncoding eukaryotic QED genetic codes, TATA, CAAT, AT-rich (9 of them), CG-rich (14 of them), and ATCG-rich, have been correlated, as shown in Table 1, as promotors and cis-regulatory elements. The coincidence between cis-elements and noncoding QED codes is listed in Table 2.

Tandem repeats (TR)

TRs of short and variable lengths have been reported to either tune gene expression [10] or cause diseases [11-13] in eukaryotes.

CAG TR also affects other neurodegenerative diseases [15], such as fragile X disorders (CGGs).

Advances in genomic technologies have accelerated the clinical use [16] of the TR. The GC-rich Penta, AT-rich Penta, Polyalanine-encoding GCN, GC-trinucleotide, and CAG/CTG/CCTG TR have been identified to cause more than sixty-seven diseases. In all reported cases, ATX or ATXX and CGX or CGXX are common and cause the disease. While verifying the triplet code, Nobel Laurate H. G. Khorana observed [23] that adjacent bases that form natural pairs, namely, underlined bases, do not promote polypeptide formation, which is a noncoding attribute. Thus, ATXX and GCXX are noncoding regions of introns. When the distribution of ATXX and CGXX in introns deviates, exon/intron and intron/exon interfaces are affected and exon-encoding regions are disrupted, yielding a dysfunctional protein that causes disease.

The noncoding QED eukaryote code explains the uncertainties regarding the causality of TR diseases. Noncoding conditions are necessary to prevent the average TR from interfering with the protein-encoding process. The QED genetic model is based on ATN or ATNN and CGN or CGNN, with N being any DNA (A, T, C, and G) base that does not promote polypeptide formation, a noncoding feature.

CAG TRs [24] are known to cause Huntington's (HT) disease. Under triplet coding, CAA and CAG encode Glutamine (Gln), but only CAG causes this disease. What prevents the gene from encoding Glen? Additionally, the number of repeats is critical for HT. Individuals with fewer than 36 repeats are disease-free, approximately 36 repeats represent the threshold, and more than 45-50 repeats lead to disease. The eukaryotic noncoding QED codons CAG or CAGN, where N is any base (T, A, C, and G), is noncoding and prevents polypeptide glutamine formation, which is not observed in Huntington's disease tests.

The thirty-five noncoding QED eukaryotic codons are involved in many TR-related diseases. Furthermore, the noncoding QED code bases and cis-regulatory element bases are highly correlated.

Number *	Noncoding codons	Noncoding Synonymous codons			H.B.	Cis-Regulation
1	*(TA)(TA)*	*(TA)(AT)*	*(AT)(TA)*	*(AT)(AT)*	8	TATA -Trans. Start
2	*(CG)(CG)*	*(CG)(GC)*	*(GC)(CG)*	*(GC)(GC)*	12	(CG)(CG)-Intron
3	*(AU)GG*	*GG(AU)*	*GG(UA)*	*(UA)GG*	10	(AU)GG- START
5	*(UG)(AG)*	*(GU)(AG)*	*(UG)(GA)*	*(GU)(AG)*	10	(UG)(AG)-STOP
8	*(UA)(GA)*	*(AG)UA)*	*(UA)(AG)*	*(GA)(AU)*	9	(UA)(GA)-STOP
10	*(UA)AA*	*AA(UA)*	*(AU)AA*	*AA(AU)*	8	(UA)AA- STOP
6	*(UG)AA*	*AA(UG)*	*(GU)AA*	*AA(GU)*	9	*AA (UG)* Promotor&
7	*(TA)(GT)*	*(GT)(TA)*	*(TA)(TG)*	*(GT)(AT)*	9	*(TA) (TG)* Promotor&
11	*(TA)(AC)*	*(AC)(TA)*	*(TA)(CA)*	*(AC)(AT)*	9	*CAAT Box,* Promotor
15	*TT(AC)*	*(AC)TT*	*(CA)TT*	*TT(CA)*	9	*TT(AC),* Promotor&
16	*TT(AG)*	*(GA)TT*	*(AG)TT*	*TT(GA)*	9	*TT(AG),* Promotor&
22	*AA(CT)*	*(CT)AA*	*(TC)AA*	*AA(TC)*	9	*AA(TC),* Promotor&
30	*(CT)(TA)*	*(TC)(TA)*	*(CT)(AT)*	*(TC)(AT)*	9	*(TA)(CT),* Promotor&
12	*(TT)(AA)*	*(AA)(TT)*			8	*TT AA,* Promotor&
14	*TT(TA)*	*(TA)TT*	*(AT)TT*	*TT(AT)*	8	*TT(TA),* Promotor&
4	*(UG)(AC)*	*(AC)(UG)*	*(UG)(CA)*	*(AC)(GU)*	10	*(UG)(CA), Promotor&*
9	*(UA)(GC)*	*(UA)(CG)*	*(CG)(UA)*	*(CG)(AU)*	10	*(GC)(AU),* Promotor&
17	*TT(CG)*	*(CG)TT*	*TT(GC)*	*(GC)TT*	10	*TT(CG),* Promotor&
18	*CC(TA)*	*(TA)CC*	*(AT)CC*	*CC(AT)*	10	*(CC)(AT),* Promotor&
23	*AA(CG)*	*(GC)AA*	*(CG)AA*	*AA(GC)*	10	*AA(CG),* Promotor&
28	*(AC)(AG)*	*(AC)(GA)*	*(CA)(GA)*	*(CA)(AG)*	10	*(AC)(AG),* Promotor&
32	*(CT)(AC)*	*(TC)(AC)*	*(CT)(CA)*	*(TC)(CA)*	10	*(TC)(AC),* Promotor&
33	*(CT)(AG)*	*(TC)(AG)*	*(CT)(GA)*	*(TC)(GA)*	10	*(CT)(AG),* Promotor&
34	*(CT)(TG)*	*(TC)(TG)*	*(CT)(GT)*	*(TC)(GT)*	10	*(CT)(AG),* Promotor&
19	*CC(TG)*	*(TG)CC*	*(GT)CC*	*CC(GT)*	11	*CC(TG),* Promotor&
20	*CC(AG)*	*(AG)CC*	*(GA)CC*	*CC(GA)*	11	*CC(AG),* Promotor&
24	*GG(CT)*	*(CT)GG*	*(TC)GG*	*GG(TC)*	11	*GG(CT),* Promotor&
26	*GG(AC)*	*(AC)GG*	*(CA)GG*	*GG(CA)*	11	*GG(AC),* Promotor&

27	*(AC)(CG)*	*(CA)(CG)*	*(CA)(GC)*	*(AC)(GC)*	11	*(AC)(CG),* Promotor&
29	*(AG)(CG)*	*(GA)(CG)*	*(AG)(GC)*	*(GA)(GC)*	11	*(AG)(CG),* Promotor&
31	*(CT)(CG)*	*(TC)(CG)*	*(CT)(GC)*	*(TC)(GC)*	11	*(CG)(TC),* Promotor&
35	*(GT)(CG)*	*(TG)(CG)*	*(GT)(GC)*	*(TG)(GC)*	11	*(CG)(TG),* Promotor&
13	*(CC)(GG)*	*(GG)(CC)*			12	*(CC)(GG),* Promotor&
21	*CC(CG)*	*(CG)CC*	*(GC)CC*	*CC(GC)*	12	*CC(CG)* , Promotor&
25	*GG(CG)*	*(CG)GG*	*(GC)GG*	*GG(GC)*	12	*GG(CG),* Promotor&

Table 1: Thirty-five noncoding QED codes and Cis-Regulatory element predictions.

* These numbers are from Table 1 (b), (1). &: upstream.

Cis-regulatory	Noncoding QED code	Table 1 row #
TATA Box	*(TA)(TA)*	1
CAAT Box	*(CA)(TA)*	11
CG/GC	*(CG)(CG)*	2
YCAY	*(TC)(AT)*	30
(Y-T (U)Or C)	*CC(AT)*	18
	(TC)(AC)	32
UAGG	*(UA)GG*	3
UGCAUG	*(GC)(AU)*	9
UGCAUG	*(UG)(CA)*	4
AT-Rich	*AT-Rich*	7,14
GC-Rich	*CG or GC-* Rich	17,21,23,25
		27,29,31,35

Table 2: Correlation between cis-regulatory elements and noncoding QED code bases.

Summary

The eukaryotic QED genetic code and cis-regulatory elements control protein synthesis in eukaryotic cells. Cis-regulatory elements in the promoter region play a pivotal role in controlling gene selection, transcription, splicing, and polymerization in human eukaryotic cells during protein synthesis, thereby maintaining homeostasis. The thirty-five noncoding QED codons highly correlate with cis-regulatory elements, as shown in Table 2. Consequently, the QED code can encode proteins and control their processing. Any variant in the gene, errors in transcription, splicing, or variable length tandem repeats can yield dysfunctional proteins that cause disease. Thus, the QED genetic code provides a paradigm shift in the research and development of tools and techniques for correcting dysfunctional proteins and identifying paths for curing incurable human diseases.

Data Availability

N/A.

Code Availability

N/A.

Acknowledgments

I am deeply grateful to Dr. Nawin Mishra, Distinguished Professor Emeritus of Genetics & Genomics, University of South Carolina, Columbia, SC 29208, for his encouragement and valuable comments regarding this work. My daughter, Usha Singh's rare, incurable disease, was the catalyst to start the lifelong journey to find a gene-disease causal relationship. The outcome is the development of QED genetic coding for eukaryotic cells and its correlation with cis-regulatory elements. I am thankful to my grandchildren, Sonya Khan and Omar Khan, for their invaluable assistance in assembling the DNA mechanical model. This work would never have materialized without the unwavering full support of my family: my wife, Amrita Singh, and my children, Asha Singh and Om Prakash Singh (a.k.a. Tom Singh).

Author Contribution

Rama Shankar Singh - 100%.

Competing Interests

The authors have no competing interests.

Additional Information

Correspondence and requests for materials should be addressed to: Rama Shankar Singh, email: gen-med @comcast.net

Reprints and permission information is available upon contacting the author.

Bibliography

1. Rama Shankar Singh. “Quadruplet Expanded DNA (QED) Genetic Code for Eukaryotic Cells”. *Acta Scientific Medical Sciences* 7.12 (2023): 70-82.
2. Crick F H. “On the genetic code”. *Science* 139 (1963): 461-464.
3. Berget SM., *et al*. “Spliced segments at the 5’ terminus of adenovirus 2 late mRNA”. *Proceedings of the National Academy of Sciences of the United States of America* 74 (1977): 3171-3175.
4. Roberts RJ and Sharp P A. “For their discoveries of split genes” The Nobel Prize in physiology or medicine 1993”. NobelPrize.org. Nobel Prize outreach AB 2022.
5. Kornberg RD. “The Nobel Prize in chemistry 2006”. NobelPrize.org. Nobel Prize (2022).
6. Wang Z and Burge C B. “Splicing regulation: From a parts list of regulatory elements to an integrated splicing code RNA”. 14 (2008): 802-813.
7. Nilsion TW., *et al*. “Expansion of the eukaryotic proteome by alternative splicing”. *Nature* 463 (2010): 457-463.
8. Jacob F and Monod J. “Genetic regulatory mechanisms in the synthesis of proteins”. *Journal of Molecular Biology* 3 (1961): 318-356.
9. Jacob F., *et al*. “The Nobel Prize in physiology or medicine 1965”. NobelPrize.org. Nobel Prize outreach AB 2022 (1965).
10. Horton CA., *et al*. “Short tandem repeats bind transcription factors to tune eukaryotic gene expression”. *Science* 381 (2023): 1304-1321.
11. Maston GA., *et al*. “Transcriptional regulatory elements in the human genome”. *Annual Review of Genomics and Human Genetics* 7 (2006): 29-59.
12. Novoyatleva T., *et al*. “Pre-mRNA missplicing as a cause of human disease”. *Progress in Molecular and Subcellular Biology* 44 (2006): 27-46.
13. Ward A J and Cooper TA. “The pathobiology of splicing”. *Journal of Pathology* 220 (2010): 152-163.
14. “Genetic Modifiers of Huntington’s disease (GeM -HD) Consortium, (lead contact -James F. Gusella), CAG Repeat Not Polyglutamine Length Determines Timing of Huntington’s Disease Onset”. *Cell* 178 (2019): 887-900.
15. Hannan A J. “Tandem repeats mediating genetic plasticity in health and disease”. *Nature Reviews* 19 (2018): 287-298.
16. Gall-Duncan T., *et al*. “Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences”. *Cold Spring Harbor Laboratory Press* 32 (2022): 1-27.
17. Carl G., *et al*. “Hold out the genome: a roadmap to solving the cis-regulatory code”. *Nature* 625 (2024): 41-50.
18. Wang J and Agarwal V. “How DNA encodes the start of transcription”. *Science* 384 (2024): 382-383.
19. Dudnyk K., *et al*. “Sequence basis of transcription initiation in the human genome”. *Science* 384 (2024): 409-425.
20. White MA. “Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed Sequences”. *Genomics* 106 (2015): 165-170.
21. Gagniuc P and Ionescu-Tirgoviste C. “Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters”. *BMC Genomics* 13 (2012): 512.
22. Liu X., *et al*. “TiGER: a database for tissue-specific gene expression and regulation”. *BMC Bioinformation* 9 (2008): 271.
23. Holley RW., *et al*. “The Nobel Prize in physiology or medicine 1968”. NobelPrize.org. Nobel Prize outreach AB 2022 (2022).
24. David Baltimore., *et al*. “The 1975 Nobel Prize in Physiology or Medicine “for their discoveries concerning the interaction between tumours viruses and the genetic material of the cell”.
25. Lewis M. “The lac repressor, C. R”. *Biologies* 328 (2005): 521-548.