Received: 2 October 2023 | Revised: 10 January 2024 | Accepted: 30 January 2024 | Published online: 5 February 2024

RESEARCH ARTICLE

Medinformatics
2024, Vol. 1(4) 184–194
DOI: 10.47852/bonviewMEDIN42021822

BON VIEW PUBLISHING

In Silico Optimization of DNA Codons in Genes Encoded by Various Variants of Ebola Virus

Anshu Mathuria[1], Mehak Ahmed[1] and Indra Mani[1],*

[1]*Department of Microbiology, University of Delhi (Gargi College), India*

Abstract: Ebola hemorrhagic fever is a serious and often deadly illness that affects primates, especially humans. Ebola virus disease (EVD) is very deadly and has no widespread treatment, making it one of the most lethal zoonotic infections. The viral genome is approximately 19 kb in length, non-segmented, linear, and negative single-stranded (−SS) RNA. The Ebola virus (EBOV) genome contains seven genes: glycoprotein (GP), nucleoprotein (NP), L, and VP (VP30, VP24, VP35, VP40), which encode GP, NP, RNA polymerase, and viral proteins. When we modified the DNA sequence by codon adaptation, we noted considerable increases in the codon adaptation index (CAI) and GC content when compared with that of the natural-type strain. On average, the NP gene in the modified DNA exhibited a 3.14-fold increase (equivalent to 213.5%) in CAI and a 1.2-fold increase (19.17%) in GC content. Similarly, the GP gene showed a 3.57-fold increase (257.14%) in CAI and a 1.16-fold increase (16.56%) in GC content. Furthermore, the modified DNA resulted in a 3.44-fold increase (244.8%) in CAI and a 1.22-fold increase (22.5%) in GC content for the VP35 gene, a 4.34-fold increase (334.8%) in CAI and a 1.26-fold increase (26.04%) in GC content for the VP30 gene, and a 3.84-fold increase (284.6%) in CAI and a 1.2-fold increase (21.2%) in GC content for the VP40 gene. The VP24 gene exhibited a 3.84-fold increase (284.61%) in CAI and a 1.23-fold increase (23.3%) in GC content, while the L gene showed a 3.84-fold increase (284.61%) in CAI and a 1.23-fold increase (23%) in GC content. The results obtained illustrate that modified genes can boost expression in a host organism without producing truncated proteins. Furthermore, GP and NP are considered as promising candidates for an EBOV vaccine, as they possess immunogenic properties and can stimulate an immune response.

Keywords: CAI, GC, Ebola virus, codon optimization, Ebola virus disease, vaccines

1. Introduction

The Ebola virus (EBOV) genome is approximately 19 kb in length, linear, non-segmented, and negative single-stranded (−SS) RNA [1]. EBOV is a zoonotic filovirus that possesses an envelope. It has been classified into five species: Zaire ebolavirus (ZEBOV), Tai Forest ebolavirus (formerly known as Cote d'Ivore), Bundibugyo ebolavirus, Sudan ebolavirus, and Reston ebolavirus (it is widespread in the Western Pacific region and is extremely deadly to primates other than humans) [2]. Furthermore, two other genera are included in the Filoviridae family, Cuevavirus and Marburgvirus (http://www.ictvonline.org/virusTaxonomy.asp). Ebola hemorrhagic fever (Ebola HF) is primarily spread among humans via direct physical contact with animal tissues or bodily fluids. Engaging in activities like processing and eating animal meat, as well as consuming contaminated water and bat droppings, is often associated with infections transmitted to humans. EBOV is usually spread by direct contact with contaminated bodily fluids through skin breaches or mucous membrane exposure [3]. The primary source of infection during an epidemic is coming in contact with either sick individuals or human corpses. However, fruit bats, which are asymptomatic carriers of the virus, are thought to be a naturally occurring reservoir and hence considered the animal reservoir of EBOV. Isolates of filoviruses obtained from fruit bats exhibit a higher level of genetic diversity [4, 5]. On November 21, 2014, the WHO stated that the continuing Ebola HF outbreak has resulted in 15,351 confirmed or probable cases, with 5,459 recorded deaths. Liberia, Guinea, and Sierra Leone are among the nations that have been most severely affected by the pandemic (http://apps.who.int/iris/bitstream/10665/144117/1/roadmapsitrep_21Nov2014_eng.pdf?ua=1).

The EBOV's virion maintains a consistent diameter of approximately 80 nm, while its length varies within the range of 970–1200 nm. However, when grown in cell culture, it exhibits a marked pleomorphism that can extend up to 14,000 nm [6]. The EBOV genome contains seven structural proteins encoding genes. These proteins are as follows, in 3'–5' order: nucleoprotein (NP), glycoprotein (GP), matrix protein (VP24), protein VP30, polymerase cofactor (VP35), matrix protein (VP40), and RNA-dependent RNA polymerase (L). The virion core contains an un-segmented, linear, and negative-sense RNA molecule. This RNA is coiled and interacts with the NP, VP30, VP35, and polymerase (L) proteins. The nucleocapsid is helical and encircled by a layer of unique GP spikes measuring 10 nm in length. The aforementioned GPs play an important function in the virus's infectivity because they facilitate virus entrance and contribute to its immunogenicity. They are targeted by immune cells and are therefore considered important in vaccine development. The viral matrix proteins VP24 and VP40 are located in between EBOV's outer envelope and nucleocapsid. The viral proteins VP35 and VP24

***Corresponding author:** Indra Mani, Department of Microbiology, University of Delhi (Gargi College), India. Email: indra.mani@gargi.du.ac.in

are key factors in virulence because they act as antagonists to type I interferon (IFN), inhibiting its action [7, 8].

The initial stage of viral replication involves adhering to the cell membrane of the host and entering the cell's interior. Although the exact mechanisms are not completely understood, it has been proven that GP spikes play a function in aiding virions' entrance into host cells, potentially through processes similar to macropinocytosis [9]. Additionally, VP30 has been identified as crucial for the re-initiation of transcription of subsequent genes during viral replication [10]. Due to its essential role in these processes, VP30 presents an intriguing target for potential antiviral therapies [11]. The assembly of viral nucleocapsids requires the availability of matrix proteins VP24, VP35, and NP. Silencing the expression of matrix protein (VP24) prevents the discharge of viruses, and it also leads to diminished transcription and translation of VP30 in VP24-deficient viral particles [12]. Furthermore, the highly produced VP40 matrix protein contributes significantly to the production of new virus particles. It is closely linked to the endosomal pathway and the process of virus development from the host cell [13].

In regions with tropical climates, where several febrile illnesses can present similar symptoms to Ebola virus disease (EVD), it is very important to consider testing for or providing empirical treatment for parasitic diseases (such as *Plasmodium* spp.), viral diseases, and bacterial diseases (such as *Salmonella typhi*) [14, 15]. Given the present COVID-19 pandemic caused by the SARS-CoV-2 virus, it is becoming increasingly important to research the specific characteristics of the EBOV that may heighten its potential for causing a worldwide pandemic in the future.

Currently, two licenced vaccines, a two-dose combination of Zabdeno (Ad26.ZEBOV) and Mvabea (MVA-BN-Filo), and Ervebo (recombinant vesicular stomatitis virus (rVSV)-ZEBOV) are being utilized for EBOV [16]. Codon optimization of DNA is a valuable technology used for enhancing the production of foreign proteins. This technique involves modifying the nucleotide sequence of a gene of interest to improve its translational efficiency in a different species, such as transforming a plant sequence into a human sequence or a human sequence into bacterial or yeast sequences [17].

The objective of our study was to optimize the codon level of all seven genes of EBOV in *Escherichia coli* (*E. coli*) using computational methods. This optimization aimed to achieve higher expression levels of the desired proteins while maintaining their antigenicity and functional activity, which were identical to their native counterparts. By optimizing the DNA codons of the studied genes, we can increase the expression of these proteins, ensuring their efficient and increased production for purposes such as immunodiagnostics and immunotherapy, without introducing changes to their amino acid sequences.

2. Materials and Method

2.1. Collection of sequence

The nucleotide sequences of different variants of the EBOV were obtained from the NCBI-GenBank. The sequences were identified by their accession numbers: KY786027.1, MG572235.1, FJ968794.1, MH121167.1, KY008770.1, and MT742157.1 (http://www.ncbi.nlm.nih.gov).

2.2. Analysis and optimization of codons

The online PHP application called Optimizer (http://genomes.urv.es/OPTIMIZER/) [18] is a useful tool to anticipate and enhance the expression of a gene in a heterologous gene expression host. Using *E. coli* str. K-12 substr. MG1655 as a reference host, Optimizer was used to optimize and calculate the codon adaptation index (CAI), G & C composition, and A &T composition of the generated sequences of DNA. This strain of *E. coli* is often used for heterologous expression of genes. CAI values were calculated for each gene in the six different variants.

2.3. Data evaluation

The statistical evaluation of the genes was conducted using Microsoft Excel 2021 software. Mean, standard deviation, and range calculations were performed on the gene data. The data were grouped in a table, and a graph was generated to illustrate the comparison between CAI values of natural-type and enhanced gene sequences for various EBOV variants.

2.4. Nucleotide sequence alignment

The seven genes' nucleotide sequences from variants KY786027.1, MG572235.1, FJ968794.1, MH121167.1, KY008770.1, and MT742157.1 were aligned using Optimizer. The alignment was performed between the natural-type sequences and the corresponding modified sequences for each gene.

3. Results

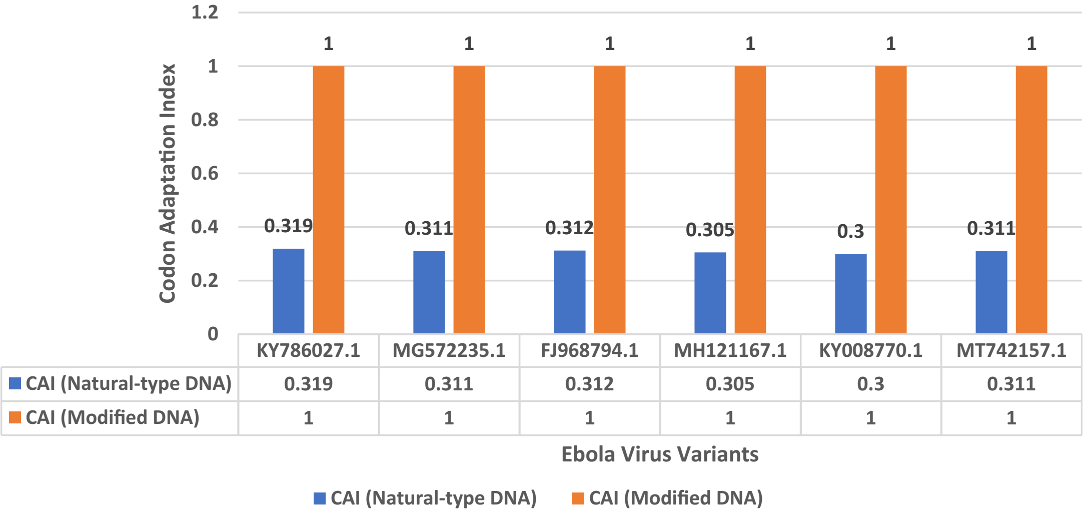
The present study aimed to address this need by employing DNA codon optimization to produce sufficient quantities of proteins in the desired host. In our study, the CAI of the codon-modified sequences of DNA was found to be greater than that of the sequences of natural-type DNA. The CAI, GC% (percentage of guanine and cytosine), and AT% (percentage of adenine and thymine) for the natural-type NP gene varied among the six different variants, ranging from 0.3 to 0.319, 45.4 to 46.9, and 53.1 to 54.6, respectively. The average values (± standard deviation) for CAI, GC%, and AT% were 0.309 (±0.0065), 46.15 (±0.668), and 53.83 (±0.668), respectively. On the other hand, the GC% and AT% frequencies in the modified DNA ranged from 53.8 to 55.9 and 44.1 to 46.2, respectively, with average values (± standard deviation) of 55.01 (±0.685) and 44.98 (±0.685). Notably, the CAI for the modified DNA was 1 for all six variants, indicating an enhanced translational efficiency of the codon-modified sequences.

Upon comparing the average values of the CAI, AT content, and GC content of the NP gene across all six variants, notable differences were observed between the modified DNA and the natural-type sequences. The modified DNA exhibited significantly higher values for GC and CAI content, with increases of 1.2 times (19.17%) and 3.14 times (213.5%), respectively, compared to the mean values of the natural-type sequences. Conversely, the modified DNA displayed a reduction of 16.4% in the mean AT content compared to the natural-type sequences (Table 1). To visually represent the variations in CAI values among the studied variants, a graph was generated, plotting CAI numeric representations along the y-axis and the number of variants along the x-axis (Figure 1). Furthermore, an alignment was performed on the nucleocapsid gene sequences of both the codon-modified sequences and natural-type sequences, as illustrated in Supplementary Figure 1.

In the case of the VP35 gene, the CAI, AT content, and GC content in the natural-type sequences of the six different variants ranged from 0.269 to 0.309, 41 to 46.8, and 53.2 to 59, respectively. The average values (±SD) were 0.29 (±0.013) for CAI, 44.78 (±1.996) for GC content, and 55.21 (±1.996) for AT content (Table 2). Upon

Table 1. The expression intensity of the NP gene from the Ebola virus in *E. coli* for the natural-type and codon-modified sequences

Ebola virus variants	Natural-type DNA			Modified DNA		
(GenBank accession no.)	CAI	GC%	AT%	CAI	GC%	AT%
KY786027.1	0.319	46.9	53.1	1	55.2	44.8
MG572235.1	0.311	46.9	53.1	1	55.2	44.8
FJ968794.1	0.312	46.4	53.6	1	54.9	45.1
MH121167.1	0.305	45.4	54.6	1	53.8	46.2
KY008770.1	0.3	45.5	54.5	1	55.9	44.1
MT742157.1	0.311	45.9	54.1	1	55.1	44.9
N	6	6	6	6	6	6
Min.	0.3	45.4	53.1	1	53.8	44.1
Max.	0.319	46.9	54.6	1	55.9	46.2
Mean ± SD	0.309 ± 0.0065	46.15 ± 0.668	53.83 ± 0.668	1.00 ± 0.00	55.01 ± 0.685	44.98 ± 0.685



	KY786027.1	MG572235.1	FJ968794.1	MH121167.1	KY008770.1	MT742157.1
■ CAI (Natural-type DNA)	0.319	0.311	0.312	0.305	0.3	0.311
■ CAI (Modified DNA)	1	1	1	1	1	1

Figure 1. Graph showing a comparison between the natural-type and modified DNA for nucleoprotein (NP) gene

optimization, the frequencies of GC and AT content in the respective DNA sequences ranged from 51.4 to 56.4 and 43.6 to 48.6, with average values (±SD) of 54.86 (±1.949) and 45.13 (±1.949), respectively. The CAI for the modified DNA was 1 for all six variants.

When the mean values of CAI, GC content, and AT content for the VP35 gene in all six variants were compared, the modified DNA had considerably higher values. The modified DNA's average CAI and GC content was calculated to be 3.44 (244.8%) and 1.22 (22.5%) times greater, respectively, than the natural-type sequences' corresponding mean values. The average AT content in the modified DNA, on the other hand, was reduced by 18.25% when contrasted with the natural-type sequences (Table 2). A graph was created to show these data (Figure 2). Both the natural-type and codon-modified VP35 gene sequences were aligned, as shown in Supplementary Figure 2. For the VP40 gene, the CAI, GC content, and AT content in the natural-type sequences of the six different variants ranged from 0.246 to 0.278, 45.7 to 48.6, and 51.4 to 54.3, respectively. The average values (±SD) were 0.263 (±0.014) for CAI, 47.13 (±1.169) for GC content, and 52.86 (±1.169) for AT content (Table 3). After codon optimization, GC and AT frequencies in the respective DNA sequences ranged from 56.6 to 57.7 and 42.3 to 43.4, with average values (±SD) of 57.12 (±0.36) and 42.88 (±0.360), respectively. The CAI for the modified DNA was 1 for all six variants.

Upon comparing the average values of CAI, GC content, and AT content for the VP40 gene in all six variants, it was observed that the values of the modified DNA were considerably greater. The modified DNA's mean CAI and GC content was calculated to be 3.84 (284.6%) and 1.2 (21.2%) times greater, respectively, than the natural-type sequences' corresponding mean values. The average AT content in the modified DNA, on the other hand, was reduced by 18.88% when compared to the natural-type sequences (Table 3). To visualize these findings, a graph was created (Figure 3). Both the natural-type and codon-modified VP40 gene sequences were aligned, as shown in Supplementary Figure 3. It is worth mentioning that codon optimization did not result in any changes to the VP40 gene's amino acid sequence.

The natural-type GP gene exhibited a range of 0.26–0.295 for CAI, 46–46.8 for GC content, and 53.2–54 for AT content across the six different variants. The average values (±SD) were 0.28 (±0.0123) for CAI, 46.48 (±0.279) for GC content, and 53.52 (±0.279) for AT content (Table 4). After codon optimization, the GC and AT frequencies in the respective DNA sequences ranged from 53 to 55.4 and 44.6 to 47, with average values (±SD) of 54.18 (±0.813) and 45.82 (±0.813), respectively. The CAI for the modified DNA was 1 for all six variants.

Upon comparing the average values of CAI, GC content, and AT content for the GP gene in all six variants, it was observed that the values of the modified DNA were considerably greater. The modified DNA's mean CAI and GC content was found to be 3.57 (257.14%) and 1.16 (16.56%) times higher, respectively, than the natural-type sequences' corresponding average values. The average AT content in the

Table 2. The expression intensity of the VP35 gene from the Ebola virus in *E. coli* for the natural-type and codon-modified sequences

Ebola virus variants	Natural-type DNA			Modified DNA		
(GenBank accession no.)	CAI	GC%	AT%	CAI	GC%	AT%
KY786027.1	0.299	44.8	55.2	1	56.1	43.9
MG572235.1	0.3	45.5	54.5	1	56.4	43.6
FJ968794.1	0.289	44.8	55.2	1	53.8	46.2
MH121167.1	0.297	46.8	53.2	1	56.2	43.8
KY008770.1	0.269	41	59	1	51.4	48.6
MT742157.1	0.309	45.8	54.2	1	55.3	44.7
N	6	6	6	6	6	6
Min.	0.269	41	53.2	1	51.4	43.6
Max.	0.309	46.8	59	1	56.4	48.6
Mean ± SD	0.29 ± 0.013	44.78 ± 1.996	55.21 ± 1.996	1.00 ± 0.00	54.86 ± 1.949	45.13 ± 1.949

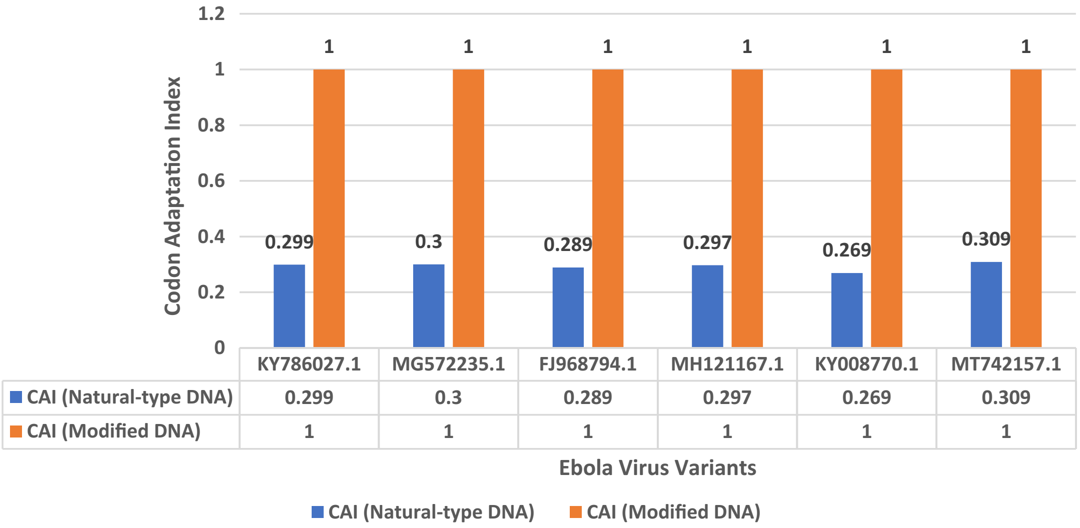


Figure 2. Graph showing a comparison between the natural-type and modified DNA for VP35 gene

Table 3. The expression intensity of the VP40 gene from the Ebola virus in *E. coli* for the natural-type and codon-modified sequences

Ebola virus variants	Natural-type DNA			Modified DNA		
(GenBank accession no.)	CAI	GC%	AT%	CAI	GC%	AT%
KY786027.1	0.278	48.2	51.8	1	57	43
MG572235.1	0.27	48.6	51.4	1	57	43
FJ968794.1	0.266	47.6	52.4	1	56.6	43.4
MH121167.1	0.275	46.5	53.5	1	57.2	42.8
KY008770.1	0.246	46.2	53.8	1	57.7	42.3
MT742157.1	0.246	45.7	54.3	1	57.2	42.8
N	6	6	6	6	6	6
Min.	0.246	45.7	51.4	1	56.6	42.3
Max.	0.278	48.6	54.3	1	57.7	43.4
Mean ± SD	0.263 ± 0.014	47.13 ± 1.169	52.86 ± 1.169	1.00 ± 0.00	57.12 ± 0.36	42.88 ± 0.360

modified DNA, on the other hand, was reduced by 14.38% when contrasted with the natural-type sequences (Table 4). To visualize these findings, a graph was created (Figure 4). As demonstrated in Supplementary Figure 4, the GP gene sequences of both the natural-type and codon-modified sequences were aligned. The goal of codon optimization is to improve translational efficiency and thereby boost the immunogenicity of epitope-based vaccinations. Modifying the codon bias of gene sequences thus has potential as a method for controlling gene expression.

The natural-type VP30 gene exhibited a range of 0.225–0.261 for CAI, 39.4–45.3 for GC content, and 54.7–60.6 for AT content across the six different variants. The average values (±SD) were 0.236 (±0.012) for CAI, 43.05 (±2.617) for GC content, and 56.95 (±2.617) for AT content (Table 5). After codon optimization, the

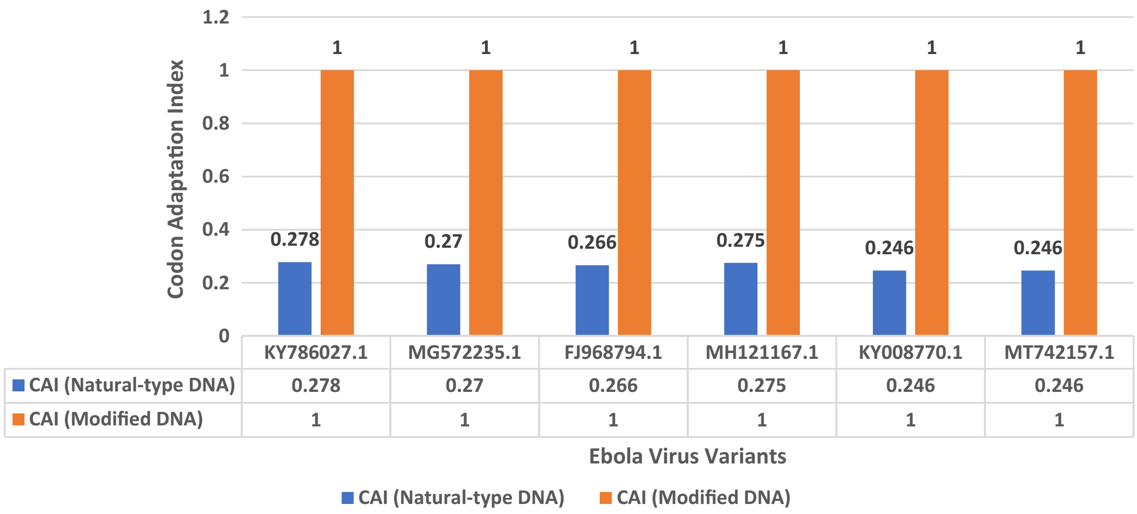


Figure 3. Graph showing a comparison between the natural-type and modified DNA for VP40 gene

Table 4. The expression intensity of the GP gene from the Ebola virus in *E. coli* for the natural-type and codon-modified sequences

Ebola virus variants	Natural-type DNA			Modified DNA		
(GenBank accession no.)	CAI	GC%	AT%	CAI	GC%	AT%
KY786027.1 (sGP)	0.275	46.8	53.2	1	53	47
MG572235.1	0.276	46	54	1	54.7	45.3
FJ968794.1	0.286	46.7	53.3	1	55.4	44.6
MH121167.1	0.26	46.5	53.5	1	53.8	46.2
KY008770.1	0.288	46.5	53.5	1	54.1	45.9
MT742157.1	0.295	46.4	53.6	1	54.1	45.9
N	6	6	6	6	6	6
Min.	0.26	46	53.2	1	53	44.6
Max.	0.295	46.8	54	1	55.4	47
Mean ± SD	0.28 ± 0.0123	46.48 ± 0.279	53.52 ± 0.279	1.00 ± 0.00	54.18 ± 0.813	45.82 ± 0.813

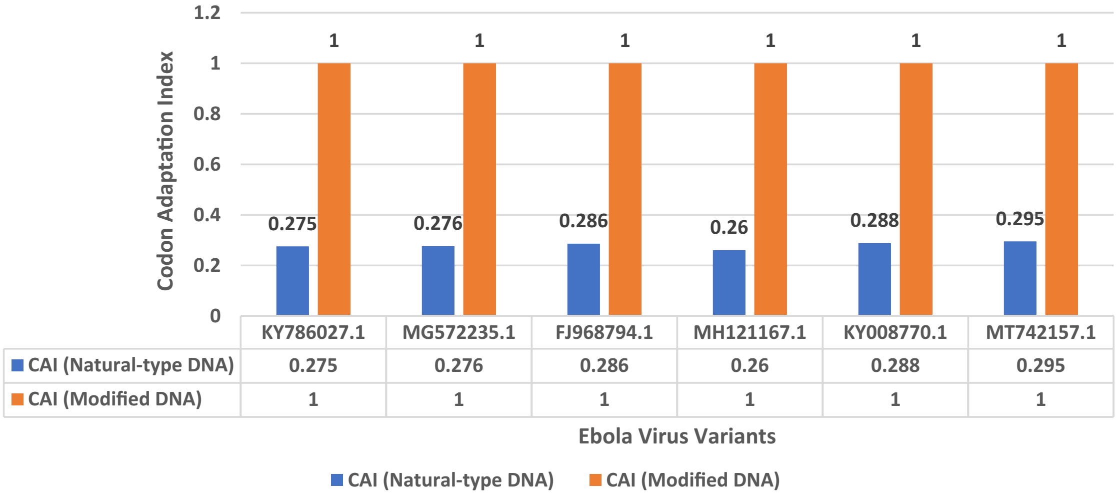


Figure 4. Graph showing a comparison between the natural-type and modified DNA for the glycoprotein (GP) gene

GC and AT frequencies in the respective DNA sequences ranged from 48.5 to 57.1 and 42.9 to 51.5, with average values (±SD) of 54.26 (±3.322) and 45.73 (±3.322), respectively. The CAI for the modified DNA was 1 for all six variants. When comparing the mean values of CAI, GC content, and AT content for the VP30 gene in all six variants, it was evident that the values of the modified DNA were considerably higher. The modified DNA's average CAI and GC content was calculated to be 4.34 (334.8%) and 1.26 (26.04%) times greater, respectively, than the natural-type sequences' corresponding mean values. The mean AT content in the modified DNA, on the other hand, was reduced by 19.7% when compared to the natural-type sequences (Table 5). Both the natural-type and codon-modified VP30 gene sequences were aligned, as shown in Supplementary Figure 5. To visualize these findings, a graph was created (Figure 5).

The natural-type VP24 gene demonstrated a range of 0.256–0.289 for CAI, 41.9–44.3 for GC content, and 55.7–58.1 for AT content across the six different variants. The average values (±SD) were 0.26 (±0.012) for CAI, 43.06 (±0.973) for GC content, and 56.93 (±0.973) for AT content (Table 6). Following codon optimization, the GC and AT frequencies in the respective DNA sequences ranged from 52.4 to 54.2 and 45.8 to 47.6, with average values (±SD) of 53.1 (±0.745) and 46.9 (±0.745), respectively. The CAI for the modified DNA was 1 for all six variants. When contrasting with the mean values of CAI, GC content, and AT content for the VP24 gene in all six variants, it was observed that the values of the modified DNA were considerably greater. The modified DNA's average CAI and GC content was found to be 3.84 (284.61%) and 1.23 (23.3%) times higher, respectively, than the natural-type sequences' corresponding average values. The average AT content in the modified DNA, on the other hand, was reduced by 17.61% when compared to the natural-type sequences (Table 6). Both the natural-type and codon-modified VP24 gene sequences were aligned, as shown in Supplementary Figure 6. To visualize these findings, a graph was created (Figure 6). As a result of codon optimization, the VP24 gene's amino acid arrangement remained unchanged.

The natural-type L gene exhibited a range of 0.23–0.382 for CAI, 39.6–40.6 for GC content, and 59.4–60.4 for AT content across the six different variants. The average values (±SD) were 0.26 (±0.057) for CAI, 39.98 (±0.386) for GC content, and 60 (±0.430) for AT content (Table 7). Upon codon optimization, the GC and AT frequencies in the respective DNA sequences ranged from 32.3 to 53.7 and 46.3 to 67.7, with average values (±SD) of 49.18 (±8.374) and 50.816 (±8.374), respectively. The CAI for the modified DNA was 1 for all six variants.

When the mean values of CAI, GC content, and AT content for the polymerase gene in all six variants were compared, the modified DNA had considerably higher values. The modified DNA's average CAI and GC content was calculated to be 4.016 (301.6.0%) and 1.34 (34.673%), respectively, times higher than the natural-type sequences' respective mean values. The average AT content in the modified DNA, on the other hand, was reduced by 22.924% as compared to the natural-type sequences (Table 7). As illustrated in Figure 7, a graph reflecting these comparisons was created. As shown in

Table 5. The expression intensity of the VP30 gene from the Ebola virus in *E. coli* for the natural-type and codon-modified sequences

Ebola virus variants	Natural-type DNA			Modified DNA		
(GenBank accession no.)	CAI	GC%	AT%	CAI	GC%	AT%
KY786027.1	0.234	45.3	54.7	1	57.1	42.9
MG572235.1	0.235	44.6	55.4	1	56.4	43.6
FJ968794.1	0.23	45.1	54.9	1	56.1	43.9
MH121167.1	0.261	39.4	60.6	1	48.5	51.5
KY008770.1	0.231	40.1	59.9	1	52.1	47.9
MT742157.1	0.225	43.8	56.2	1	55.4	44.6
N	6	6	6	6	6	6
Min.	0.225	39.4	54.7	1	48.5	42.9
Max.	0.261	45.3	60.6	1	57.1	51.5
Mean ± SD	0.236 ± 0.012	43.05 ± 2.617	56.95 ± 2.617	1.00 ± 0.00	54.26 ± 3.322	45.73 ± 3.322

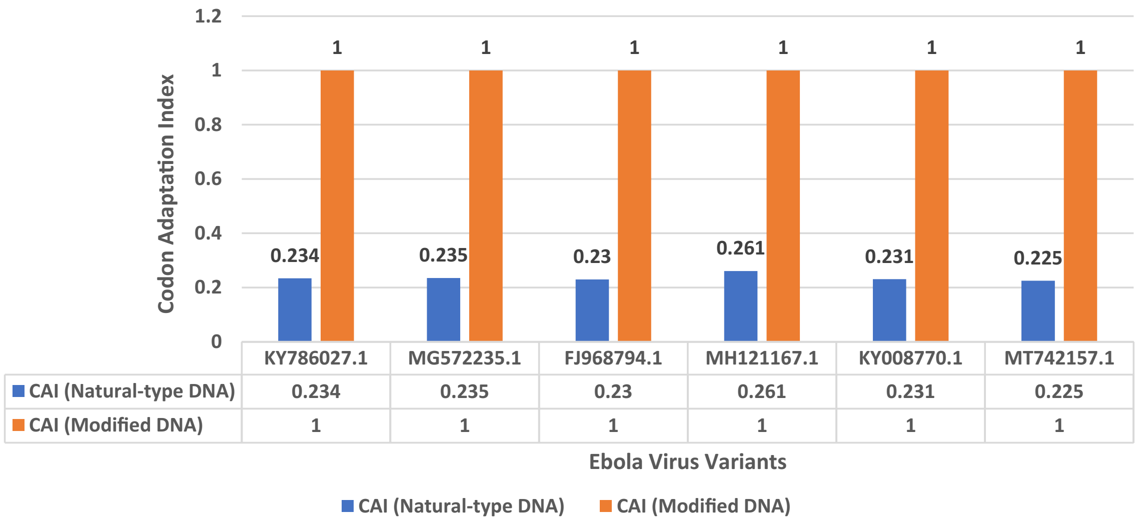
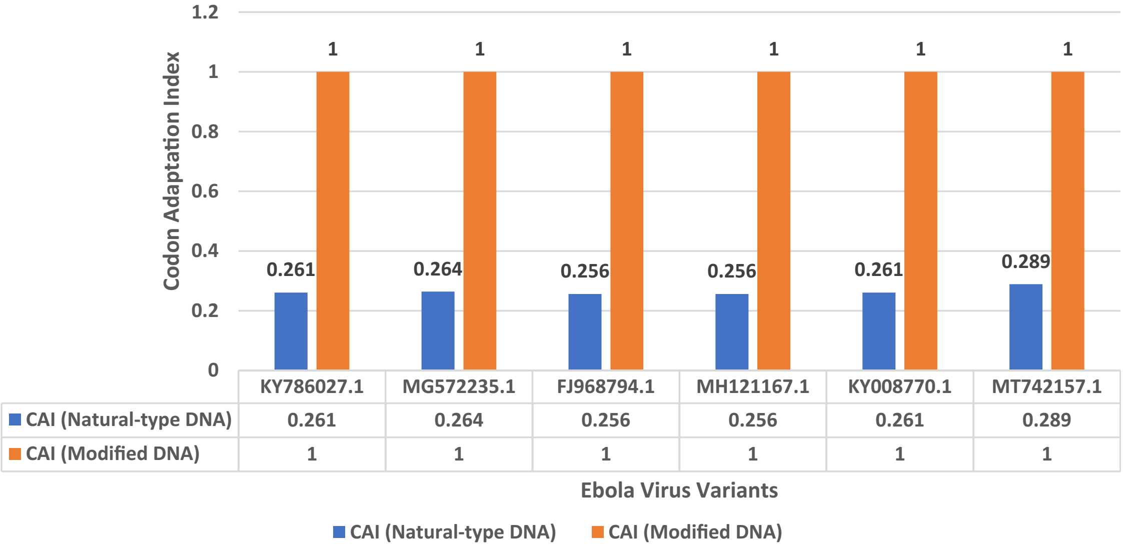


Figure 5. Graph showing a comparison between the natural-type and modified DNA for the VP30 gene

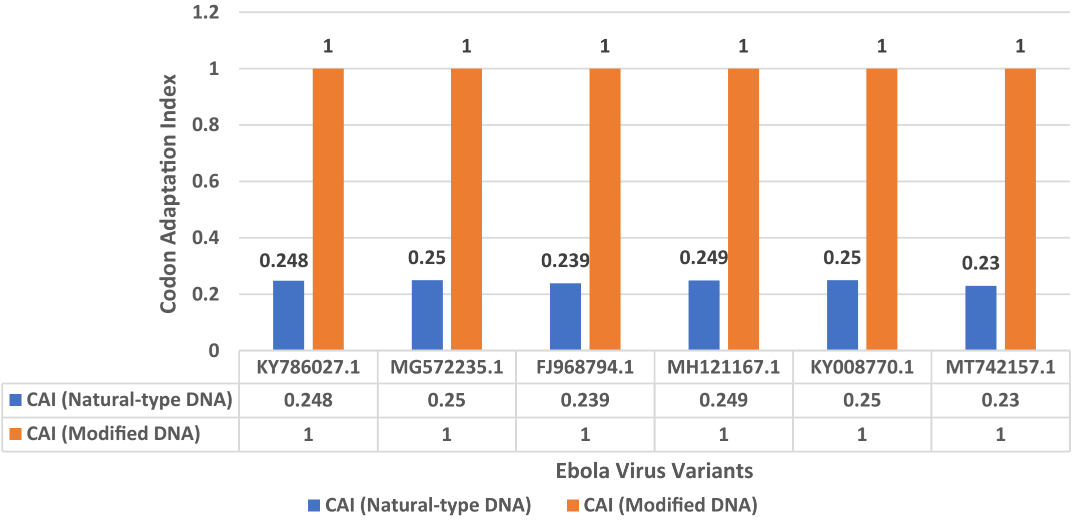
Table 6. The expression intensity of the VP24 gene from the Ebola virus in *E. coli* for the natural-type and codon-modified sequences

Ebola virus variants (GenBank accession no.)	Natural-type DNA			Modified DNA		
	CAI	GC%	AT%	CAI	GC%	AT%
KY786027.1	0.261	42.9	57.1	1	52.6	47.4
MG572235.1	0.264	42.5	57.5	1	52.4	47.6
FJ968794.1	0.256	42.6	57.4	1	52.4	47.6
MH121167.1	0.256	44.3	55.7	1	54.2	45.8
KY008770.1	0.261	41.9	58.1	1	53.4	46.6
MT742157.1	0.289	44.2	55.8	1	53.6	46.4
N	6	6	6	6	6	6
Min.	0.256	41.9	55.7	1	52.4	45.8
Max.	0.289	44.3	58.1	1	54.2	47.6
Mean ± SD	0.26 ± 0.012	43.06 ± 0.973	56.93 ± 0.973	1.00 ± 0.00	53.1 ± 0.745	46.9 ± 0.745



	KY786027.1	MG572235.1	FJ968794.1	MH121167.1	KY008770.1	MT742157.1
CAI (Natural-type DNA)	0.261	0.264	0.256	0.256	0.261	0.289
CAI (Modified DNA)	1	1	1	1	1	1

Figure 6. Graph showing a comparison between the natural-type and modified DNA for VP24 gene



	KY786027.1	MG572235.1	FJ968794.1	MH121167.1	KY008770.1	MT742157.1
CAI (Natural-type DNA)	0.248	0.25	0.239	0.249	0.25	0.23
CAI (Modified DNA)	1	1	1	1	1	1

Figure 7. Graph showing a comparison between the natural-type and modified DNA for polymerase (L) gene

Table 7. The expression intensity of the L gene from the Ebola virus in *E. coli* for the natural-type and codon-modified sequences

Ebola virus variants (GenBank accession no.)	Natural-type DNA			Modified DNA		
	CAI	GC%	AT%	CAI	GC%	AT%
KY786027.1	0.248	39.9	60.1	1	53.2	46.8
MG572235.1	0.25	39.6	60.4	1	53.1	46.9
FJ968794.1	0.239	40.3	59.7	1	53.7	46.3
MH121167.1	0.249	39.8	60.2	1	53.6	46.4
KY008770.1	0.25	39.7	60.3	1	52.8	47.2
MT742157.1	0.23	40.6	59.4	1	50	50
N	6	6	6	6	6	6
Min.	0.23	39.6	59.4	1	50	46.3
Max.	0.25	40.6	60.4	1	53.7	50
Mean ± SD	0.243 ± 0.0081	39.98 ± 0.386	60 ± 0.430	1.00 ± 0.00	52.73 ± 1.379	47.26 ± 1.379

Supplementary Figure 7, the polymerase gene sequences of both the natural-type and codon-modified sequences were aligned.

4. Discussions

Codon optimization is a widely used technique aimed at optimizing protein expression by overcoming constraints related to codon usage. This approach involves modifying the DNA sequence to optimize codon selection, thereby enhancing gene functionality, improving protein expression levels, reducing production costs, and facilitating drug development. Codon optimization finds application in various contexts, including animal testing, the removal of stop codons, and the improvement of gene functionality and protein expression levels. It is a common expression technique employed to enhance protein levels. This is achievable because of the degeneracy of the genetic code, which permits multiple synonymous codons to encode most amino acids. The choice of codons significantly impacts protein expression levels, leading to the widespread use of codon optimization in bioproduction and *in vivo* nucleic acid medicinal applications.

Studies have demonstrated the substantial potential of codon optimization to significantly increase protein expression, with some reports indicating improvements exceeding 1000-fold [19]. However, the majority of studies show more modest increases in protein expression levels. Unexpectedly, synonymous codon mutations have also been leveraged to finely adjust expression by de-optimizing. For instance, in the case of bispecific antibodies, de-optimizing one of the light chain genes resulted in improved antibody production [20]. DNA-based immunity has shown promise in human trials for HIV infection. In one study, adapting the codon usage of the HIV gag protein through a DNA vaccination resulted in a gene expression increment by 10-fold compared to the natural-type sequence [21]. Gene optimization has proven beneficial in various therapeutic applications, particularly when a protein is produced *in vivo* following gene delivery. This method is currently widely employed across various applications [18]. Gene optimization affects numerous molecular mechanisms, encompassing transcription and translation, resulting in elevated product yield and activity [22]. In addition to enhancing protein expression levels, codon usage analysis has found applications in metagenomic studies. For example, frequency analysis of codon usage has been employed to ascertain the host range of RNA virus genomes in high-temperature acidic metagenomes, irrespective of their bacterial, archaeal, or eukaryotic origins [23].

In this study, we calculated the mean values of CAI, AT, and GC content for all EBOV variants. We compared these values with the corresponding values obtained for the modified DNA sequences. The results concluded that the sequences of modified DNA had noticeably different and greater values compared to their respective natural-type variants for all genes. Upon analyzing the NP gene of all six variants, the modified DNA sequences exhibited significantly higher CAI values. In the modified DNA, the CAI of the NP gene was increased by 3.14-fold (213.5%). Similarly, the CAI values in the modified DNA were enhanced by 3.44-fold (244.8% increase) for the VP35 gene, 3.84-fold (284.6% increase) for the VP40 gene, and 3.57-fold (257.14% increase) for the GP gene. The VP30 gene showed an increase of 4.34-fold (334.8% increase), while the VP24 gene showed an increase of 3.84-fold (284.61% increase). Additionally, the L gene exhibited an increase of 3.84-fold (284.61% increase).

These findings indicate that the optimization of DNA sequences led to significantly higher CAI values for various EBOV genes, indicating improved codon usage and potentially enhanced gene expression. Additionally, the study observed a rise in the proportion of GC content in modified DNA sequences compared to natural-type sequences. The GC content of the NP gene in the modified DNA showed an average increase of 1.2-fold (an increase of 19.17%). Likewise, the GC content in the modified DNA of VP35 and VP40 increased by 1.22-fold (an increase of 22.5%) and 1.2-fold (an increase of 21.2%), respectively. The VP24 gene exhibited an increase of 1.23-fold (an increase of 23.3%), and the L gene demonstrated a similar increase of 1.23-fold (23% increase). However, the AT content in all genes remained notably lower compared to natural-type sequences. Likewise, the in silico codon optimization approach has been applied across various microorganisms, including *Mycobacterium tuberculosis* [24], influenza A virus [25], SARS-COV-2 [26], Nipah virus [27], and others. EBOV survivors' humoral responses largely target the surface GP, and the existence of anti-GP neutralizing antibodies has been linked to protection against EBOV infection. The EBOV surface GPs GP1, 2 are important for host cell adhesion and fusion and are the principal target of neutralizing antibodies.

Several vaccine candidates are currently being developed to prevent EVD by utilizing the EBOV GP and NP to elicit a defensive immune response in preclinical animal models. Two vaccine strategies that have shown promise in preventing EVD include rVSV-based vaccine, which uses a modified vesicular stomatitis virus (VSV) vector to express the ZEBOV surface GP. The vaccination is called rVSVG-ZEBOV-GP. It has been used in preclinical research and has been granted WHO prequalification. This vaccine has demonstrated effectiveness in preventing EBOV infection. Adenovirus and modified vaccinia Ankara (MVA) vector-based vaccine. Adenovirus and modified vaccinia Ankara (MVA) vector-based vaccine employs an adenovirus type 26-vectored vaccine containing the ZEBOV GP

(Ad26.ZEBOV), followed by a booster dose of a MVA virus strain (MVA-BN-Filo). This sequential immunization strategy has also gained WHO prequalification status and has demonstrated promising effects in EVD prevention [28–31]. These vaccine candidates have been used in response to recent Ebola outbreaks and are considered important tools in controlling and preventing the spread of the EBOV. In clinical trials conducted by the Partnership for Research on Ebola Vaccinations (PREVAC) collaboration, promising results were observed with the rVSVG-ZEBOV-GP vaccine and the Ad26.ZEBOV-MVA-BN-Filo combination. The rVSVG-ZEBOV-GP vaccine, a single-dose option, is recommended for reactive ring vaccination among individuals at high risk of EBOV infection during epidemics. The Ad26.ZEBOV-MVA-BN-Filo combination, a two-dose regimen, is recommended for individuals at low to moderate risk of EBOV infection [32]. GP and NP, known for their immunogenic properties, are suitable vaccine candidates against EBOV.

Utilizing codon optimization can enhance gene expression in cells, yielding high titers for large-scale vaccine production. Nevertheless, this process may pose challenges, necessitating a thorough understanding of codon usage patterns within the target host or expression system. To overcome these challenges, conducting *in vitro* analyses and experimental validation is crucial, complementing in silico studies to guarantee the safety and efficacy of biotech therapeutics.

5. Conclusions

In conclusion, the development of vaccines based on GP or NP, combined with codon optimization and immunology-based studies, offers great promise in combating EBOV infection. This approach can facilitate the production of effective vaccines at an industrial scale. However, it is essential to address the challenges associated with codon optimization to ensure the safety and efficacy of resulting vaccines. While there are several applications for codon optimization in the development of recombinant protein therapeutics and nucleic acid treatments, it is vital to examine potential obstacles. Reports suggest that synonymous codon changes introduced during optimization may impact protein conformation, stability, and function, as well as increase in immunogenicity. Such alterations could have unintended consequences, including reduced efficacy or association with certain diseases. To overcome the constraints brought on by codon bias, the current study, however, focuses on optimizing codons to improve the expression of seven EBOV genes in *E. coli*. Increased GC and CAI content in the modified sequences suggested that *E. coli* might overexpress them. Based on this study and previous research on immunogenicity, GP and NP are promising candidates for an EBOV vaccine, as they possess immunogenic properties and can stimulate an immune response. The subsequent phase involves *in vitro* validation of the findings from the in silico study, assessing the degree of overexpression attained by the modified sequences, and evaluating their safety and efficacy in eliciting an immune response. Successful validation could lead to further development of these modified genes on an industrial scale for immunodiagnostic tools and immunotherapeutic. It is crucial to emphasize that rigorous testing is essential to confirm the findings and ensure the safety and potency of potential vaccines. These endeavors will aid in the advancement of efficient immunodiagnostic and immunotherapeutic approaches to combat EBOV infections, offering optimism for managing this lethal illness.

Acknowledgment

The authors are thankful to the principal of Gargi College for providing the infrastructural support.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

Indra Mani is an Editorial Board Member for *Medinformatics*, and was not involved in the editorial review or the decision to publish this article. The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in supplementary figures at https://doi.org/10.47852/bonviewMEDIN42021822.

Supplementary Information

The supplementary figures are available at https://doi.org/10.47852/bonviewMEDIN42021822.

Author Contribution Statement

Anshu Mathuria: Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Mehak Ahmed:** Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Indra Mani:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Jain, S., Martynova, E., Rizvanov, A., Khaiboullina, S., & Baranwal, M. (2021). Structural and functional aspects of Ebola virus proteins. *Pathogens*, *10*(10), 1330. https://doi.org/10.3390/pathogens10101330

[2] Goeijenbier, M., van Kampen, J. J., Reusken, C. B., Koopmans, M. P., & van Gorp, E. C. (2014). Ebola virus disease: A review on epidemiology, symptoms, treatment and pathogenesis. *The Netherlands Journal of Medicine*, *72*(9), 442–448.

[3] Kanapathipillai, R., Henao Restrepo, A. M., Fast, P., Wood, D., Dye, C., Kieny, M. P., & Moorthy, V. (2014). Ebola vaccine—An urgent international priority. *New England Journal of Medicine*, *371*(24), 2249–2251. https://doi.org/10.1056/NEJMp1412166

[4] Carroll, S. A., Towner, J. S., Sealy, T. K., McMullan, L. K., Khristova, M. L., Burt, F. J., . . . , & Nichol, S. T. (2013). Molecular evolution of viruses of the family *Filoviridae* based on 97 whole-genome sequences. *Journal of Virology*, *87*(5), 2608–2616. https://doi.org/10.1128/jvi.03118-12

[5] Towner, J. S., Amman, B. R., Sealy, T. K., Carroll, S. A. R., Comer, J. A., Kemp, A., . . . , & Rollin, P. E. (2009). Isolation of genetically diverse Marburg viruses from Egyptian fruit bats. *PLOS Pathogens*, *5*(7), e1000536. https://doi.org/10.1371/journal.ppat.1000536

[6] Geisbert, T. W., & Jahrling, P. B. (1995). Differentiation of filoviruses by electron microscopy. *Virus Research*, *39*(2–3), 129–150. https://doi.org/10.1016/0168-1702(95)00080-1

[7] Feldmann, H., Sanchez, A., & Geisbert, T. (2013). *Filoviridae*: Marburg and Ebola viruses. In D. M. Knipe & P. M. Howley (Eds.), *Fields virology: Sixth edition* (Vol. 1). Lippincott Williams & Wilkins.

[8] Mateo, M., Carbonnelle, C., Martinez, M. J., Reynard, O., Page, A., Volchkova, V. A., & Volchkov, V. E. (2011). Knockdown of Ebola virus VP24 impairs viral nucleocapsid assembly and prevents virus replication. *The Journal of Infectious Diseases*, *204*, S892–S896. https://doi.org/10.1093/infdis/jir311

[9] Aleksandrowicz, P., Marzi, A., Biedenkopf, N., Beimforde, N., Becker, S., Hoenen, T., . . . , & Schnittler, H. J. (2011). Ebola virus enters host cells by macropinocytosis and clathrin-mediated endocytosis. *The Journal of Infectious Diseases*, *204*, S957–S967. https://doi.org/10.1093/infdis/jir326

[10] Biedenkopf, N., Hartlieb, B., Hoenen, T., & Becker, S. (2013). Phosphorylation of Ebola virus VP30 influences the composition of the viral nucleocapsid complex: Impact on viral transcription and replication. *Journal of Biological Chemistry*, *288*(16), 11165–11174. https://doi.org/10.1074/jbc.M113.461285

[11] Ascenzi, P., Bocedi, A., Heptonstall, J., Capobianchi, M. R., Di Caro, A., Mastrangelo, E., . . . , & Ippolito, G. (2008). Ebolavirus and Marburgvirus: Insight the *Filoviridae* family. *Molecular Aspects of Medicine*, *29*(3), 151–185. https://doi.org/10.1016/j.mam.2007.09.005

[12] Hoenen, T., Groseth, A., Kolesnikova, L., Theriault, S., Ebihara, H., Hartlieb, B., . . . , & Becker, S. (2006). Infection of naÏve target cells with virus-like particles: Implications for the function of ebola virus VP24. *Journal of Virology*, *80*(14), 7260–7264. https://doi.org/10.1128/jvi.00051-06

[13] Stahelin, R. V. (2014). Membrane binding and bending in Ebola VP40 assembly and egress. *Frontiers in Microbiology*, *5*, 300. https://doi.org/10.3389/fmicb.2014.00300

[14] Carroll, M. W., Haldenby, S., Rickett, N. Y., Pályi, B., Garcia-Dorival, I., Liu, X., . . . , & Hiscox, J. A. (2017). Deep sequencing of RNA from blood and oral swab samples reveals the presence of nucleic acid from a number of pathogens in patients with acute Ebola virus disease and is consistent with bacterial translocation across the gut. *mSphere*, *2*(4), e00325-17. https://doi.org/10.1128/mspheredirect.00325-17

[15] Vernet, M. A., Reynard, S., Fizet, A., Schaeffer, J., Pannetier, D., Guedj, J., . . . , & Baize, S. (2017). Clinical, virological, and biological parameters associated with outcomes of Ebola virus infection in Macenta, Guinea. *JCI Insight*, *2*(6), e88864. https://doi.org/10.1172/jci.insight.88864

[16] Tomori, O., & Kolawole, M. O. (2021). Ebola virus disease: Current vaccine solutions. *Current Opinion in Immunology*, *71*, 27–33. https://doi.org/10.1016/j.coi.2021.03.008

[17] Graf, M., Schoedl, T., & Wagner, R. (2009). Rationales of gene design and *de novo* gene construction. In P. Fu & S. Panke (Eds.), *Systems biology and synthetic biology* (pp. 411–438). Wiley. https://doi.org/10.1002/9780470437988.ch12

[18] Puigbò, P., Guzmán, E., Romeu, A., & Garcia-Vallvé, S. (2007). OPTIMIZER: A web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Research*, *35*, W126–W131. https://doi.org/10.1093/nar/gkm219

[19] Gustafsson, C., Minshull, J., Govindarajan, S., Ness, J., Villalobos, A., & Welch, M. (2012). Engineering genes for predictable protein expression. *Protein Expression and Purification*, *83*(1), 37–46. https://doi.org/10.1016/j.pep.2012.02.013

[20] Magistrelli, G., Poitevin, Y., Schlosser, F., Pontini, G., Malinge, P., Josserand, S., . . . , & Fischer, N. (2017). Optimizing assembly and production of native bispecific antibodies by codon de-optimization. *mAbs*, *9*(2), 231–239. https://doi.org/10.1080/19420862.2016.1267088

[21] Menzella, H. G. (2011). Comparison of two codon optimization strategies to enhance recombinant protein production in *Escherichia coli*. *Microbial Cell Factories*, *10*(1), 15. https://doi.org/10.1186/1475-2859-10-15

[22] Edison, L. K., Dan, V. M., Reji, S. R., & Pradeep, N. S. (2020). A strategic production improvement of *Streptomyces* Beta glucanase enzymes with aid of codon optimization and heterologous expression. *Biosciences Biotechnology Research Asia*, *17*(3), 587–599. https://doi.org/10.13005/bbra/2862

[23] Stedman, K. M., Kosmicki, N. R., & Diemer, G. S. (2013). Codon usage frequency of RNA virus genomes from high-temperature acidic-environment metagenomes. *Journal of Virology*, *87*(3), 1919–1919. https://doi.org/10.1128/jvi.02610-12

[24] Mani, I., Chaudhary, D. K., Somvanshi, P., & Singh, V. (2010). Codon optimization of the potential antigens encoding genes from *Mycobacterium tuberculosis*. *International Journal of Applied Biology and Pharmaceutical Technology*, *1*(2), 292–301.

[25] Mani, I., Singh, V., Chaudhary, D. K., Somvanshi, P., & Negi, M. P. S. (2011). Codon optimization of the major antigen encoding genes of diverse variants of influenza a virus. *Interdisciplinary Sciences: Computational Life Sciences*, *3*, 36–42. https://doi.org/10.1007/s12539-011-0055-z

[26] Al Zamane, S., Nobel, F. A., Jebin, R. A., Amin, M. B., Somadder, P. D., Antora, N. J., . . . , & Moni, M. A. (2021). Development of an in silico multi-epitope vaccine against SARS-COV-2 by précised immune-informatics approaches. *Informatics in Medicine Unlocked*, *27*, 100781. https://doi.org/10.1016/j.imu.2021.100781

[27] Gupta, A., Gangotia, D., Vasdev, K., & Mani, I. (2021). In silico DNA codon optimization of the variant antigen-encoding genes of diverse variants of Nipah virus. *bioRxiv Preprint*. https://doi.org/10.1101/2021.04.23.441071

[28] Afolabi, M. O., Ishola, D., Manno, D., Keshinro, B., Bockstal, V., Rogers, B., . . . , & Watson-Jones, D. (2022). Safety and immunogenicity of the two-dose heterologous Ad26.ZEBOV and MVA-BN-Filo Ebola vaccine regimen in children in Sierra Leone: A randomised, double-blind, controlled trial. *The Lancet Infectious Diseases*, *22*(1), 110–122. https://doi.org/10.1016/S1473-3099(21)00128-6

[29] Barry, H., Mutua, G., Kibuuka, H., Anywaine, Z., Sirima, S. B., Meda, N., . . . , & Thiébaut, R. (2021). Safety and immunogenicity of 2-dose heterologous Ad26.ZEBOV, MVA-BN-Filo Ebola vaccination in healthy and HIV-infected adults: A randomised, placebo-controlled phase II clinical trial in Africa. *PLOS Medicine*, *18*(10), e1003813. https://doi.org/10.1371/journal.pmed.1003813

[30] Gsell, P. S., Camacho, A., Kucharski, A. J., Watson, C. H., Bagayoko, A., Nadlaou, S. D., . . . , & Kéïta, S. (2017). Ring vaccination with rVSV-ZEBOV under expanded access in response to an outbreak of Ebola virus disease in Guinea, 2016: An operational and vaccine safety report. *The Lancet Infectious Diseases*, *17*(12), 1276–1284. https://doi.org/10.1016/S1473-3099(17)30541-8

[31] Henao-Restrepo, A. M., Camacho, A., Longini, I. M., Watson, C. H., Edmunds, W. J., Egger, M., . . . , & Kieny, M. P. (2017).

Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: Final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). *The Lancet*, *389*(10068), 505–518. https://doi.org/10.1016/S0140-6736(16)32621-6

[32] Lévy, Y., Lane, C., Piot, P., Beavogui, A. H., Kieh, M., Leigh, B., . . . , & Yazdanpanah, Y. (2018). Prevention of Ebola virus disease through vaccination: Where we are in 2018. *The Lancet*, *392*(10149), 787–790. https://doi.org/10.1016/S0140-6736(18)31710-0

How to Cite: Mathuria, A., Ahmed, M., & Mani, I. (2024). In Silico Optimization of DNA Codons in Genes Encoded by Various Variants of Ebola Virus. *Medinformatics*, *1*(4), 184–194. https://doi.org/10.47852/bonviewMEDIN42021822