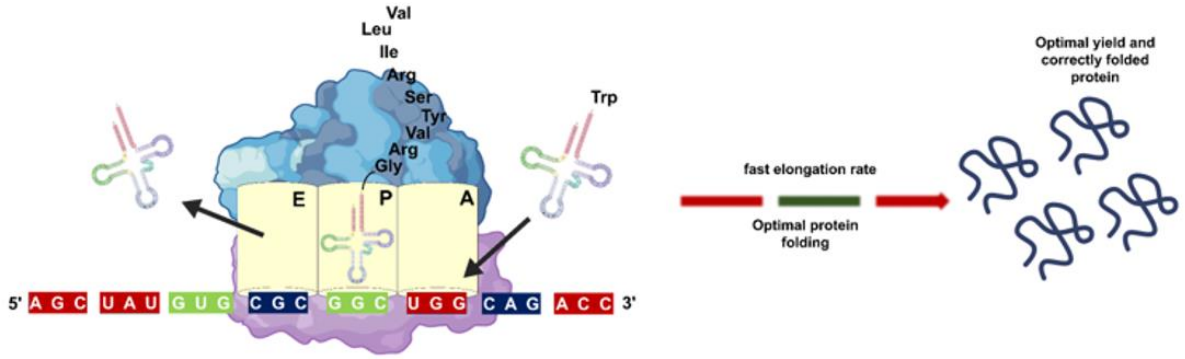
Impact of different synonymous codon substitution strategies on SARS-CoV-2 nucleocapsid protein expression in *Escherichia coli*

Adebayo J. Bello[1,2*], Ayomide O. Omotuyi[2], Olajumoke B. Oladapo[2,3], Adegbayi A. Adekunle[4], Kingsley U. Udechime[2], Olufunke T. Olugbenga[2], Abdulzamad B. Akinwande[2], Adeyemi F. Odewale[2], Eban Davis[2], Olawale Adamson[4], Onikepe Folarin[1], Joy Okpuzor[2], Joseph, B. Minari[2*]

1. Department of Biological Sciences, Redeemer's University, Ede, Osun State, Nigeria.
2. Department of Cell Biology and Genetics, Faculty of Science, University of Lagos, Akoka, Lagos, Nigeria.
3. Present Address: Stephenson School of Biomedical Engineering, University of Oklahoma.
4. Department of Oral and Maxillofacial Surgery, Lagos University Teaching Hospital/College of Medicine University of Lagos, Idi-Araba, Lagos, Nigeria.

Corresponding Author: Adebayo J. Bello: bayulinks@gmail.com;
Joseph B. Minari: jminari@unilag.edu.ng

Graphical Abstract



Abstract

Synonymous codon substitution, a gene engineering approach in synthetic biology, has been effective in improving the codon composition of recombinant genes of interest based on various criteria without altering the amino acid sequence. The SARS-CoV-2 virus nucleocapsid (N) protein is a stable, conserved and highly immunogenic that is less prone to mutation during infection, making it a key antigen in *in vitro* diagnosis, vaccine development, immunological and structural studies. While reports have focused on applying optimized N protein for different applications, the basic parameters used by different optimization tools for choosing the best approach for the N gene synonymous codon substitution are often neglected. Here, we analyzed the influence of different synonymous codon substitution strategies on SARS-CoV-2 N-protein expression in *E. coli*. Using different codon optimization (CO) and harmonization (CH) tools, we predicted and compared how parameters such as GC content, Codon Adaptation Index, codon quality and number of rare codons present in these sequences affect the N-protein expression. Our results also show that Minimum Free Energy (MFE) and RNA structure of N-term and C-tail of the N-protein coding sequence influence protein folding. We then predicted that the SR-rich region of the N-protein may contribute to slowing down the elongation rate during translation. This work presents a fundamental analysis of how different optimization tools affect SARS-CoV-2 N-protein expression and folding and suggests a basic approach to choosing the best strategy for optimal expression and folding of the protein for further studies.

Keywords: Genetic code; codon optimization; codon harmonization; protein expression; protein folding; SARS-CoV-2; nucleocapsid.

Highlights

Codon substitution affects SARS-CoV-2 nucleocapsid (N) protein expression.

SARS-CoV-2 N-protein expression varies with different codon optimization tools.

RNA structures of N- and C-term impact RNA stability, protein expression and folding.

SR-rich region of the N-protein may slow down elongation rate during translation.

Introduction

Severe acute respiratory syndrome coronavirus (SARS-CoV-2) is a fast-transmissible virus among humans, and it is responsible for coronavirus disease (COVID-19) [1]. Since the outbreak of the disease in the city of Wuhan, Hubei Province, China in December 2019, there have been more than 700 million confirmed cases and over 7 million deaths globally as of July 2025. (https://www.worldometers.info/coronavirus/). This disease poses an incomparable threat to global public health, and new theranostic discoveries are needed to combat the disease. Despite new vaccine development and drug candidates that provide a saving route to reduce the burden of the coronavirus disease, there is a need for continuous studies on the virus and its effect on host-pathogen interaction.

The SARS-CoV-2 nucleocapsid (N) protein is one of the four structural proteins with 419 bp amino residues. The N-protein is the most abundant, relatively conservative protein in coronaviruses [2], and it is one of the most studied proteins because of its importance as an antiviral target [3]. While serving as RNA binding proteins (RBPs), the N-protein is responsible for packaging the viral RNA genome into ribonucleotide complex (RNP) through protein-RNA interaction [4], determines pathogenicity and virulence [5], regulates progression of the host cell cycle, plays an important role in interaction of the virus with the host, as well as the apoptosis [6]. The protein has been reported to be highly expressed during infection [7]. These characteristics make the protein a key factor in the viral life cycle.

Most studies on SARS-CoV-2 involve expression of the viral proteins either *in vitro* or *in vivo* to develop drug candidates, vaccines, and diagnostic biomolecules to provide safe working environments [8], [9] [10]. Likewise, *in silico* studies have formed an integral part of SARS-CoV-2 research either by designing new bioinformatics tools or making use of existing algorithms to understand the virus structural architecture, design new drugs and vaccines against the virus [11], [12], [13]. However, for efficient viral protein expression, it has been proven that synonymous codon substitution [14] is an important factor to consider when expressing the protein in a heterologous host such as *Escherichia coli*. Because multiple codons can code for the same amino acid because of the genetic code degeneracy, which can thus influence protein expression, preferential usage of codons termed codon usage bias however varies in organisms [15]. Thus, different organisms are codon-biased during protein expression based on the abundance of corresponding intracellular aminoacyl-tRNAs during translation [16].

Production of synthetic nucleic acids and recombinant proteins is common in most laboratories for research purposes and industrial applications. One strategy that has improved this process over time is optimizing codons of genes of interest to fit into the codon usage bias of the

heterologous host [17], [18]. This process of codon optimization or synonymous codon substitution for efficient production of recombinant proteins in heterologous hosts is one of the major bases of modern biotechnology. This approach involves engineering gene sequences to enhance protein expression without changing the amino acid letters and employs translation elongation rate as a means of improving the expressed protein.

Different tools have been designed to improve protein expression in heterologous hosts. Some reports have also documented the codon usage pattern and codon bias of the SARS-CoV-2 genome [19], [20], [21] while the virus N-protein optimization and expression for different downstream studies such as diagnostic [22], vaccines [23], immunological assessment [24], and structural studies [25] have also been reported. However, while these reports focused on applying optimized protein, the basic parameters for synonymous codon substitution used by different optimization tools are often neglected. The unreported comparison which usually leads to differences in the protein yield, solubility, and folding may be critical to SARS-CoV-2 nucleocapsid proteins for vaccine production, diagnostics, and structural and immunological studies. Thus, determining the most suitable tool for desirable optimized protein is vital, as different tools have unique parameters for optimizing nucleotide sequences [26]. Also, no report to the best of our knowledge has investigated how SAR-CoV-2 N-protein expression is affected by codon harmonization, a strategy to adapt the gene of interest into heterologous for optimal protein folding [27].

In this study, we performed synonymous codon substitution and compared the underlining parameters affecting SARS-CoV-2 nucleocapsid protein expression using JCat and Genewiz codon optimization algorithms, and codon harmonization tool. These optimization tools were chosen based on frequent usage for protein expressions in reported studies [27], [28], [29]. Overall, the genetic code substitution strategy predicted variation in expression analysis which may affect the protein yield and folding properties.

2.1. Methods

2.1. Retrieval of Gene Coding Sequences from Database

The gene coding sequence for SARS-CoV-2 nucleocapsid (N) protein was obtained from nucleotide database of National Center for Biotechnology Information (NCBI) website www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?from=28274&to=29533&report=fasta). The FASTA format of nucleotide sequence from 28274-29533 bp with Gene ID; 43740575 was used for this study.

2.2. Synonymous codon substitution strategies

Two codon optimization algorithms (JCat and Genewiz) and one codon harmonization algorithm [29] were used for the synonymous codon substitution of the nucleocapsid gene sequence. The coding sequence for the N protein (wild type) was obtained from the NCBI database and pasted into each algorithm while selecting *Escherichia coli* as the heterologous expression host for the N protein. Then, the tools were used to generate different variants of the N protein coding sequence (Suppl. File 1). All other parameters were set at default.

2.3. *In silico* analysis of the nucleocapsid gene codon sequence

Analysis of the gene coding sequences of the wildtype and the variants for the GC content, Codon Adaptation Index (CAI), rare codons frequency, and codon quality class were calculated on the Biologics Corp's Bioinformatics Center (https://www.biologicscorp.com/tools/) using *E. coli* as a heterologous host. The rare codon clustering of the substituted synonymous codons was calculated using the percentage MinMax Calculator [30].

2.4. Cloning and protein expression

All the *E. coli* cells were grown and maintained in Luria-Bertani (LB) medium with constant shaking at 220 rpm at 37°C. Routinely, 25 µg/ml chloramphenicol (Cm) and 50 µg/ml kanamycin (Km) when added when necessary. The different optimized sequences were flanked by NdeI and XhoI restriction enzymes and synthesized by Integrated DNA Technologies (IDT). The cDNA sequence was cloned into PET28a+ derivative plasmid between NdeI and XhoI sites. NEB10 *E. coli* strain was transformed with the plasmids pJB30_JCat, pJB31_Genewiz, and pJB33_CH through electroporation. The clones were selected on plates containing both Cm and Km and grown in liquid culture overnight at 37°C. Plasmid DNA was isolated from the overnight culture using QIAprep® Spin Miniprep Kit (Qiagen) and the precision and orientation of the sequence were confirmed by sequencing.

BL21(DE3)pLysS *E. coli* strains were chemically transformed with the plasmids and grown overnight in media containing Cm and Km. Expression of the transformed cells with relevant plasmids was carried out under the control of an inducible T7 promoter, as described in [31]. Fresh overnight culture was added to a sterile LB medium, and the cells were induced with IPTG at a final concentration of 1 mM after 100 minutes of initial growth and further incubated for 135 minutes post-induction before harvesting. Cell pellets from 5ml culture were first suspended in 10mM Tris-HCl pH 7.5 before adding 500 µl CelLytic® B Lysis buffer (Sigma, UK). The mixture was vortexed briefly and incubated at room temperature with constant shaking (60 rpm; 5 minutes) before centrifuging (13,000 rpm; 5 minutes). Then, protein loading samples from the supernatant were prepared in 4X Laemmli buffer containing 50mM final concentration

of Dithiothreitol (DTT). The sample mixture was incubated at 95°C for 5 minutes and spun down shortly. Sodium dodecyl-polyacrylamide gel electrophoresis (SDS-PAGE) was used to separate the protein bands on a 4–20% Mini-PROTEAN® TGX Stain-Free™ Protein Gels (Bio-Rad). The gel image was visualized and analyzed using ChemiDoc™ Imaging System (Bio-Rad) and ImageJ software v1.5j respectively.

2.5 RNA stability

The RNA stability of the N terminal region of the sequences was determined using RNAfold [32]. The secondary structures of their RNA were generated in FORNA format, and the Minimum free energy (MFE) was generated for each sequence.

2.6. Statistical analysis

All descriptive values were calculated using Microsoft Excel version 2402 (2024). Significant differences were calculated using one-way ANOVA where $*P<0.05$.

3. Results and discussion

As one of the basic principles in synthetic biology, codon optimization can conveniently enable high expression of toxic proteins in microbial hosts [33]. As the research on SARS-CoV-2 continues to evolve, it is important to determine the best-fit parameters for expressing its toxic protein in microbial hosts.

3.1. Prediction of the effect of different synonymous codon substitution strategies

In this study, we compared the impact of different codon optimization strategies on the expression of SARS-CoV-2 nucleocapsid protein in *E. coli*. Using *In Silico* analysis, we showed that codon optimization is an important tool for SARS-CoV-2 N-protein and compared the different parameters responsible for increased protein expression. The optimized sequences differ from the wildtype by 24.44, 26.98, and 23.97 % for JCat, Genewiz, and CH optimized sequence respectively as calculated from the Percentage Index Matrix (Supplemental Figure S1) after performing sequence alignment using ClustalW. The percentage differences between the optimized sequence and the wildtype sequence show that the codon substitution strategy has significant impact on the protein expression.

The presence of rare codons can lead to a reduction in the protein yield in microbial hosts. Several reports have shown that too many rare codons do not only delay translation, but they can contribute to early termination leading to production of premature proteins [34], [35], [36], [37]. Our result shows that the low rare codons in the Genewiz optimized sequence can contribute to increased elongation process during translation, as against the low frequency JCat, and CH (Fig 1a). Most of the optimized codons also fall within the codon quality class of 91-100 (Fig 1b), indicating a strong influence on the N-protein expression in *E. coli*.

The calculated GC contents for the optimized sequences are 53, 58, and 50% for JCat, Genewiz, Geneart, and CH respectively (Fig 1c). The CAI are 0.81, 0.92, and 0.57 for JCat, Genewiz, and CH respectively (Fig 1d). GC content and CAI have been good predictive parameters for increased protein expression. Reports have shown that higher GC content can increase mRNA levels [38] and influence protein expression [39], [40]. Although some reports argued that CAI is not a comprehensive measure to determine the level of protein expression, however, Lin *et al.*, (2022) showed that an increase in CAI is a good predictive value for experimental approach [41]. In our study, Genewiz optimized sequence showed the highest values for the calculated parameters. Generally, the optimized sequence influenced the predictive parameters; increased the GC content and CAI, reduced rare codons frequency, and improved codon quality class compared to JCat and CH. Fulfilling all these parameters positively, we suggest the possibility of higher expression of the SARS-CoV-2 N-protein in *E. coli* using Genewiz codon optimization tool.

3.2. protein expression

To validate the experiment, we cloned and expressed the plasmids carrying each coding sequence in BL21(DE3)pLysS *E. coli* cells. The cloned codon-optimized genes flanked by NdeI and XhoI were confirmed after sequencing to show the correct sequence and their orientation. The N-protein expressed after IPTG induction (Fig 1E) shows that the synonymous substitution to the coding sequence did not affect the cell growth condition by inducing toxicity [42], [43]. The Genewiz recombinant N-protein was more highly expressed than JCat and CH (Fig 1F) using protein band quantification. However, the expression of JCat codon optimization was significantly lower than CH by forming inclusion bodies [44] despite fulfilling all the *in-silico* analyses positively and having the highest number of relative codon frequency (Table 1).

Table 1: codon usage frequency for each optimized sequence calculated from https://www.biologicscorp.com/tools/CodonUsageCalculator

		Jcat		Genewiz		CH	
Amino acid		**Codon**	**Frequency**	**Codon**	**Frequency**	**Codon**	**Frequency**
Ala	**A**	GCU	1	GCG	0.89	GCC	0.38
Arg	**R**	CGU	0.93	CGC	0.52	AGA	0.41
Asn	**N**	AAU	1	AAC	0.86	AAC	1
Asp	**D**	GAC	1	GAU	0.96	GAC	1
Cys	**C**						
Glu	**E**	GAA	1	GAA	1	GAG	1
Gln	**Q**	CAG	0.97	CAG	0.54	CAA	0.8
Gly	**G**	GGU	1	GGC	0.74	GGG	0.93
His	**H**	CAC	1	CAU	1	CAC	0.75
Ile	**I**	AUC	1	AUU	1	AUA	1
Leu	**L**	CUG	0.96	CUG	0.93	CUA	0.85
Lys	**K**	AAA	0.77	AAA	0.84	AAG	1
Met	**M**	AUG	1	AUG	1	AUG	1
Phe	**F**	UUC	1	UUU	0.83	UUC	1
Pro	**P**	CCG	1	CCG	0.96	CCC	0.36
Ser	**S**	UCU	1	AGC	0.81	UCU	0.41
Thr	**T**	ACC	1	ACC	0.84	ACA	0.53
Trp	**W**	UGG	1	UGG	1	UGG	1
Tyr	**Y**	UAC	1	UAU	1	UAC	1
Val	**V**	GUU	1	GUG	0.75	GUA	0.5
STOP		UAA		UAA		UAA	
Average			**0.98**		**0.87**		**0.79**

3.3. Impact of RNA structure

To further analyze the impact of synonymous codon substitution we predicted the secondary RNA structure to evaluate how they affect N-protein expression and folding. RNAs fold into 3D structures that range from simple helical elements to complex tertiary structures and quaternary ribonucleoprotein assemblies, and the functions of many regulatory RNAs depend on how their 3D structure changes in response to a diverse array of cellular conditions [45]. In this study, RNA secondary structure analysis of the N-term and C-tail (Fig 2A) was carried out to evaluate

the structural differences and minimum free energy (MFE). The three RNA sequences have structures with different numbers of hairpin loops at the N-term (Fig 2B-D) and C-tail (Fig 2F-H). The "linear" structure of RNA in Genewiz at the N-term and C-tail may contribute to its stability and high protein expression, which agrees with Leppek et al., (2022) [46], that reducing the number of hairpin loops can enhance mRNA stability and increase protein expression. The linear pattern in Genewiz suggests an optimal translation rate to allow the N-term of the protein to fold properly before the C-terminal part is formed [47]. Conspicuously, the JCat optimized sequence resulted in an MFE of -41.6 kcal/mol, while Genewiz and CH optimized sequences yielded -40.2 and -5.4 kcal/mol respectively at N-term (Fig 1E). At the C-term, Jcat, Genewiz, and CH have -44.1, -38.0, and -23.7 kcal/mol respectively. The more negative the MFE, the more stable the RNA. However, highly stable RNA structure can interfere with ribosome access leading to low translation [48]. We interpreted this report to be the case of JCat with the highest MFE at both ends of the N-protein. Consequently, the Genewiz-optimized sequence appears to be the most advantageous choice in terms of RNA stability, considering the optimal MFE and the structural linearity.

3.4. Percentage MiniMax analysis and protein folding

Similarly, throughout translation, the accurate pairing of codons and anticodons guides the incorporation of amino acids into the growing polypeptide chain. Adequate codon usage ensures that the translation machinery has access to a diverse pool of tRNAs, minimizing the occurrence of ribosomal pauses and misincorporations that could lead to protein misfolding [49]. Fig 3 shows the relative codon usage of the wildtype and optimized sequences using %MiniMax calculator with Genewiz showing a higher codon usage compared to the other sequences. The %MiniMax calculator, a freely accessible web-based portal and downloadable script, evaluates relative usage frequencies of synonymous codons within a protein sequence of interest [50]. They compare these results against a rigorous null model, providing valuable insights into codon usage patterns and optimization [30]. In Fig. 3A, the three sequences showed similar codon usage patterns, notably from amino residues 168 to 203 which may suggest similar folding patterns. The 36 amino acid residues majorly fall within the SR-rich IDR region of the nucleocapsid protein structure (Fig 3B). This may suggest the importance of serine and arginine bonds in protein expression and folding. However, the synonymous mutation resulted in 47.2% of the three RNA sequences (Fig 3C). Previous studies have experimentally confirmed the effect of codon substitution on protein folding [51], [52]. Liu et al 2021 [36] and Mauro and Chappell, 2014 [53] also provided extended reviews on the impact of codon optimization on protein

folding. However, more experimental studies are needed to investigate the effect of SR-rich region and how synonymous mutation in the region may contribute to the overall protein expression and folding. Given these accounts, it is important to balance high protein yield with folding. Generally, the Genewiz pattern in this study may suggest higher protein expression and a well-folded protein.

3.5. Elongation rhythm prediction

Synonymous code substitution finetunes codon bias usage to fit the abundance of tRNA available for protein production [54]. Therefore, we predict and represent how elongation rhythm is influenced by the three strategies used in this study (Fig 4). The decrease or increase in the translation rate influences the tRNA binding to the cognate mRNA as the ribosome performs its function which may eventually affect the protein yield and folding. A combination of “slow” and “rapid” tRNA:mRNA binding would generate a correctly folded protein but low yield (Fig4a). While a combination of “very rapid” and “rapid” with some “slow” tRNA:mRNA binding would generate optimal yield and correctly folded protein (Fig4b), an all through “very rapid” would lead to production of the high yield protein forming inclusion bodies (Fig4c)

Conclusion

This study shows that the optimized sequence of SARS-CoV-2 nucleocapsid protein for *E. coli* using Genewiz produced an optimal yield which could influence the folding of the protein compared to other optimized sequences. This work can provide a fundamental study for expressing viral proteins in microbial hosts to ascertain the best codon optimization strategies suitable for further studies such as designing new drugs and vaccine candidates, and structural studies.

Competing interest

Authors declare no competing interest.

References

1. Lamers, M.M. & Haagmans, B.L., 2022. SARS-CoV-2 pathogenesis. Nature Reviews Microbiology, 20(5), pp.270-284. doi:10.1038/s41579-022-00660-8
2. Zhang, P., Gao, Q., Wang, T., Ke, Y., Mo, F., Jia, R., Liu, W., Liu, L., Zheng, S., Liu, Y. & Li, L., 2020. Evaluation of recombinant nucleocapsid and spike proteins for serological diagnosis of novel coronavirus disease 2019 (COVID-19). MedRxiv, pp.2020-03. doi:10.1101/2020.03.17.20037713
3. Morse, M., Sefcikova, J., Rouzina, I., Beuning, P.J. & Williams, M.C., 2023. Structural domains of SARS-CoV-2 nucleocapsid protein coordinate to compact long nucleic acid substrates. Nucleic Acids Research, 51(1), pp.290-303. doi:10.1093/nar/gkad114
4. Kang, S., Yang, M., Hong, Z., Zhang, L., Huang, Z., Chen, X., He, S., Zhou, Z., Zhou, Z., Chen, Q. & Yan, Y., 2020. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. Acta Pharmaceutica Sinica B, 10(7), pp.1228-1238. doi:10.1016/j.apsb.2020.04.009
5. Moghaddar, M., Radman, R. & Macreadie, I., 2021. Severity, pathogenicity and transmissibility of delta and lambda variants of SARS-CoV-2, toxicity of spike protein and possibilities for future prevention of COVID-19. Microorganisms, 9(10), p.2167. doi:10.3390/microorganisms9102167
6. Gioti, K., Kottaridi, C., Voyiatzaki, C., Chaniotis, D., Rampias, T. & Beloukas, A., 2021. Animal coronaviruses induced apoptosis. Life, 11(3), p.185. doi:10.3390/life11030185
7. Pan, P., Shen, M., Yu, Z., Ge, W., Chen, K., Tian, M., Xiao, F., Wang, Z., Wang, J., Jia, Y. & Wang, W., 2021. SARS-CoV-2 N protein promotes NLRP3 inflammasome activation to induce hyperinflammation. Nature Communications, 12(1), p.4664. doi:10.1038/s41467-021-24924-9
8. Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B.J. & Jiang, S., 2009. The spike protein of SARS-CoV—a target for vaccine and therapeutic development. Nature Reviews Microbiology, 7(3), pp.226-236. doi:10.1038/nrmicro2090
9. Weiss, C., Carriere, M., Fusco, L., Capua, I., Regla-Nava, J.A., Pasquali, M., Scott, J.A., Vitale, F., Unal, M.A., Mattevi, C. & Bedognetti, D., 2020. Toward nanotechnology-enabled approaches against the COVID-19 pandemic. ACS Nano, 14(6), pp.6383-6406. doi:10.1021/acsnano.0c03697
10. Zhou, Y., Liu, Y., Gupta, S., Paramo, M.I., Hou, Y., Mao, C., Luo, Y., Judd, J., Wierbowski, S., Bertolotti, M. & Nerkar, M., 2023. A comprehensive SARS-CoV-2–human protein–protein interactome reveals COVID-19 pathobiology and potential host

therapeutic targets. Nature Biotechnology, 41(1), pp.128-139. doi:10.1038/s41587-022-01542-2

11. Robson, B., 2020. Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. Computers in Biology and Medicine, 119, p.103670. doi:10.1016/j.compbiomed.2020.103670.
12. Hufsky, F., Lamkiewicz, K., Almeida, A., Aouacheria, A., Arighi, C., Bateman, A., Baumbach, J., Beerenwinkel, N., Brandt, C., Cacciabue, M. & Chuguransky, S., 2021. Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research. Briefings in Bioinformatics, 22(2), pp.642-663. doi:10.1093/bib/bbaa232
13. Bukhari, S.N.H., Jain, A., Haq, E., Mehbodniya, A. & Webber, J., 2022. Machine learning techniques for the prediction of B-cell and T-cell epitopes as potential vaccine targets with a specific focus on SARS-CoV-2 pathogen: A review. Pathogens, 11(2), p.146. doi:10.3390/pathogens11020146
14. Lin, B.C., Kaissarian, N.M. & Kimchi-Sarfaty, C., 2022. Implementing computational methods in tandem with synonymous gene recoding for therapeutic development. Trends in Pharmacological Sciences. doi:10.1016/j.tips.2022.01.004
15. Kim, Y.A., Mousavi, K., Yazdi, A., Zwierzyna, M., Cardinali, M., Fox, D., Peel, T., Coller, J., Aggarwal, K. & Maruggi, G., 2023. Computational design of mRNA vaccines. Vaccine. doi:10.1016/j.vaccine.2022.12.058
16. Su, Z., Wilson, B., Kumar, P. & Dutta, A., 2020. Noncanonical roles of tRNAs: tRNA fragments and beyond. Annual Review of Genetics, 54, pp.47-69. doi:10.1146/annurev-genet-021120-105406
17. Gustafsson, C., Govindarajan, S. & Minshull, J., 2004. Codon bias and heterologous protein expression. Trends in Biotechnology, 22(7), pp.346-353. doi:10.1016/j.tibtech.2004.04.006
18. Parvathy, S.T., Udayasuriyan, V. & Bhadana, V., 2022. Codon usage bias. Molecular Biology Reports, 49(1), pp.539-565. doi:10.1007/s11033-021-06864-2
19. Hou, W., 2020. Characterization of codon usage pattern in SARS-CoV-2. Virology Journal, 17(1), p.138. doi:10.1186/s12985-020-01413-1
20. Das, J.K. & Roy, S., 2021. Comparative analysis of human coronaviruses focusing on nucleotide variability and synonymous codon usage patterns. Genomics, 113(4), pp.2177-2188. doi:10.1016/j.ygeno.2021.04.023

21. Tyagi, N., Sardar, R. & Gupta, D., 2022. Natural selection plays a significant role in governing the codon usage bias in the novel SARS-CoV-2 variants of concern (VOC). PeerJ, 10, p.e13562. doi:10.7717/peerj.13562Li, G., Li, W., Fang, X., Song, X., Teng, S., Ren, Z., Hu, D., Zhou, S., Wu, G. and Li, K., 2021. Expression and purification of recombinant SARS-CoV-2 nucleocapsid protein in inclusion bodies and its application in serological detection. Protein expression and purification, 186, p.105908.
22. Feng, Y., Zhang, Y., He, Z., Huang, H., Tian, X., Wang, G., Chen, D., Ren, Y., Jia, L., Wang, W. & Wu, J., 2022. Immunogenicity of an inactivated SARS-CoV-2 vaccine in people living with HIV-1: a non-randomized cohort study. EClinicalMedicine, 43. doi:10.1016/j.eclinm.2021.101227
23. Di, D., Dileepan, M., Ahmed, S., Liang, Y. & Ly, H., 2021. Recombinant SARS-CoV-2 nucleocapsid protein: Expression, purification, and its biochemical characterization and utility in serological assay development to assess immunological responses to SARS-CoV-2 infection. Pathogens, 10(8), p.1039. doi:10.3390/pathogens10081039
24. Hodge, C.D., Rosenberg, D.J., Grob, P., Wilamowski, M., Joachimiak, A., Hura, G.L. & Hammel, M., 2021. Rigid monoclonal antibodies improve detection of SARS-CoV-2 nucleocapsid protein. MAbs, 13(1), p.1905978. doi:10.1080/19420862.2021.1905978
25. Huang, C.Y., Draczkowski, P., Wang, Y.S., Chang, C.Y., Chien, Y.C., Cheng, Y.H., Wu, Y.M., Wang, C.H., Chang, Y.C., Chang, Y.C. and Yang, T.J., 2022. In situ structure and dynamics of an alphacoronavirus spike protein by cryo-ET and cryo-EM. Nature communications, 13(1), p.4877. https://doi.org/10.1038/s41467-022-32588-3
26. Bello, A.J., Igwo-Ezikpe, M., Toye, E.T. & Okpuzor, J., 2019. The expression level of a recombinant lipase predicted in silico by different codon optimization algorithms. World Scientific News, (137), pp.31-41.
27. Punde, N., Kooken, J., Leary, D., Legler, P.M. & Angov, E., 2019. Codon harmonization reduces amino acid misincorporation in bacterially expressed P. falciparum proteins and improves their immunogenicity. AMB Express, 9(1), pp.1-14. doi:10.1186/s13568-019-0888-1
28. Diez, M., Medina-Muñoz, S.G., Castellano, L.A., da Silva Pescador, G., Wu, Q. & Bazzini, A.A., 2022. iCodon customizes gene expression based on the codon composition. Scientific Reports, 12(1), p.12126. doi:10.1038/s41598-022-16269-4
29. Rehbein, P., Berz, J., Kreisel, P. & Schwalbe, H., 2019. “CodonWizard”–An intuitive software tool with graphical user interface for customizable codon optimization in

protein expression efforts. Protein Expression and Purification, 160, pp.84-93. doi:10.1016/j.pep.2019.04.003

30. Rodriguez, A., Wright, G., Emrich, S. & Clark, P.L., 2018. % MinMax: A versatile tool for calculating and comparing synonymous codon usage and its impact on protein folding. Protein Science, 27(1), pp.356-362. doi:10.1002/pro.3337
31. Summers, D.K. & Sherratt, D.J., 1988. Resolution of ColE1 dimers requires a DNA sequence implicated in the three-dimensional organization of the cer site. EMBO Journal, 7, pp.851-858. doi:10.1002/j.1460-2075.1988.tb02884.x
32. Olorunniji, F.J., Merrick, C., Rosser, S.J., Smith, M.C.M., Stark, W.M. & Colloms, S.D., 2017. Multipart DNA assembly using site-specific recombinases from the large serine integrase family. In Site-specific Recombinases, Methods in Molecular Biology; Eroshenko, N., ed.; Humana Press, New York, USA, pp.303-323.
33. Hofacker, I.L., 2003. Vienna RNA secondary structure server. Nucleic Acids Research, 31(13), pp.3429-3431. doi:10.1093/nar/gkg601
34. Kaur, J., Kumar, A. & Kaur, J., 2018. Strategies for optimization of heterologous protein expression in E. coli: Roadblocks and reinforcements. International Journal of Biological Macromolecules, 106, pp.803-822. doi:10.1016/j.ijbiomac.2017.08.124
35. Yang, Q., Yu, C.H., Zhao, F., Dang, Y., Wu, C., Xie, P., Sachs, M.S. & Liu, Y., 2019. eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. Nucleic Acids Research, 47(17), pp.9243-9258. doi:10.1093/nar/gkz638
36. Nieuwkoop, T., Finger-Bou, M., van der Oost, J. & Claassens, N.J., 2020. The ongoing quest to crack the genetic code for protein production. Molecular Cell, 80(2), pp.193-209. doi:10.1016/j.molcel.2020.09.025
37. Liu, Y., Yang, Q. & Zhao, F., 2021. Synonymous but not silent: the codon usage code for gene expression and protein folding. Annual Review of Biochemistry, 90, pp.375-401. doi:10.1146/annurev-biochem-081320-093601
38. Kudla, G., Lipinski, L., Caffin, F., Helwak, A. & Zylicz, M., 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biology, 4(6), p.e180. doi:10.1371/journal.pbio.0040180
39. Piovesan, A., Pelleri, M.C., Antonaros, F., Strippoli, P., Caracausi, M. & Vitale, L., 2019. On the length, weight, and GC content of the human genome. BMC Research Notes, 12(1), pp.1-7. doi:10.1186/s13104-019-4130-x

40. Mauger, D.M., Cabral, B.J., Presnyak, V., Su, S.V., Reid, D.W., Goodman, B., Link, K., Khatwani, N., Reynders, J., Moore, M.J. & McFadyen, I.J., 2019. mRNA structure regulates protein expression through changes in functional half-life. Proceedings of the National Academy of Sciences, 116(48), pp.24075-24083. doi:10.1073/pnas.1909569116
41. Lin, B.C., Kaissarian, N.M. & Kimchi-Sarfaty, C., 2022. Implementing computational methods in tandem with synonymous gene recoding for therapeutic development. Trends in Pharmacological Sciences. doi:10.1016/j.tips.2021.11.001
42. Rump, A., Risti, R., Kristal, M.L., Reut, J., Syritski, V., Lookene, A. & Boudinot, S.R., 2021. Dual ELISA using SARS-CoV-2 nucleocapsid protein produced in E. coli and CHO cells reveals epitope masking by N-glycosylation. Biochemical and Biophysical Research Communications, 534, pp.457-460. doi:10.1016/j.bbrc.2020.10.114
43. Royster, A., Ren, S., Ma, Y., Pintado, M., Kahng, E., Rowan, S., Mir, S. & Mir, M., 2023. SARS-CoV-2 nucleocapsid protein is a potential therapeutic target for anticoronavirus drug discovery. Microbiology Spectrum, 11(3), pp.e01186-23. doi:10.1128/spectrum.01186-23
44. Di, D., Dileepan, M., Ahmed, S., Liang, Y. & Ly, H., 2021. Recombinant SARS-CoV-2 nucleocapsid protein: Expression, purification, and its biochemical characterization and utility in serological assay development to assess immunological responses to SARS-CoV-2 infection. Pathogens, 10(8), p.1039. doi:10.3390/pathogens10081039
45. Tauber, D., Tauber, G. & Parker, R., 2020. Mechanisms and regulation of RNA condensation in RNP granule formation. Trends in Biochemical Sciences, 45(9), pp.764-778. doi:10.1016/j.tibs.2020.06.005
46. Ganser, L.R., Kelly, M.L., Herschlag, D. & Al-Hashimi, H.M., 2019. The roles of structural dynamics in the cellular functions of RNAs. Nature Reviews Molecular Cell Biology, 20(8), pp.474-489. doi:10.1038/s41580-019-0149-1
47. Leppek, K., Byeon, G.W., Kladwang, W., Wayment-Steele, H.K., Kerr, C.H., Xu, A.F., Kim, D.S., Topkar, V.V., Choe, C., Rothschild, D. & Tiu, G.C., 2022. Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. Nature Communications, 13(1), p.1536. doi:10.1038/s41467-022-29294-w
48. Sander, I.M., Chaney, J.L. & Clark, P.L., 2014. Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design. Journal of the American Chemical Society, 136(3), pp.858-861. doi:10.1021/ja409554d

49. Kim, D.S., Challa, S., Jones, A. & Kraus, W.L., 2020. PARPs and ADP-ribosylation in RNA biology: from RNA expression and processing to protein translation and proteostasis. Genes & Development, 34(5-6), pp.302-320. doi:10.1101/gad.334730.119
50. Zhou, M., Wang, T., Fu, J., Xiao, G., & Liu, Y., 2015. Nonoptimal codon usage influences protein structure in intrinsically disordered regions. Molecular Microbiology, 97(5), pp.974-987. doi:10.1111/mmi.13073
51. Clarke IV, T.F., & Clark, P.L., 2008. Rare codons cluster. PLoS ONE, 3(10), p.e3412. doi:10.1371/journal.pone.0003412
52. Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M.V., & Komar, A.A., 2016. Synonymous codons direct cotranslational folding toward different protein conformations. Molecular Cell, 61(3), pp.341-351. doi:10.1016/j.molcel.2016.01.008
53. Shen, X., Song, S., Li, C., & Zhang, J., 2022. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. Nature, 606(7915), pp.725-731. doi:10.1038/s41586-022-04739-4
54. Mauro, V.P., & Chappell, S.A., 2014. A critical analysis of codon optimization in human therapeutics. Trends in Molecular Medicine, 20(11), pp.604-613. doi:10.1016/j.molmed.2014.09.003
55. Komar, A.A., 2021. A code within a code: how codons fine-tune protein folding in the cell. Biochemistry (Moscow), 86, pp.976-991. doi:10.1134/S0006297921080087

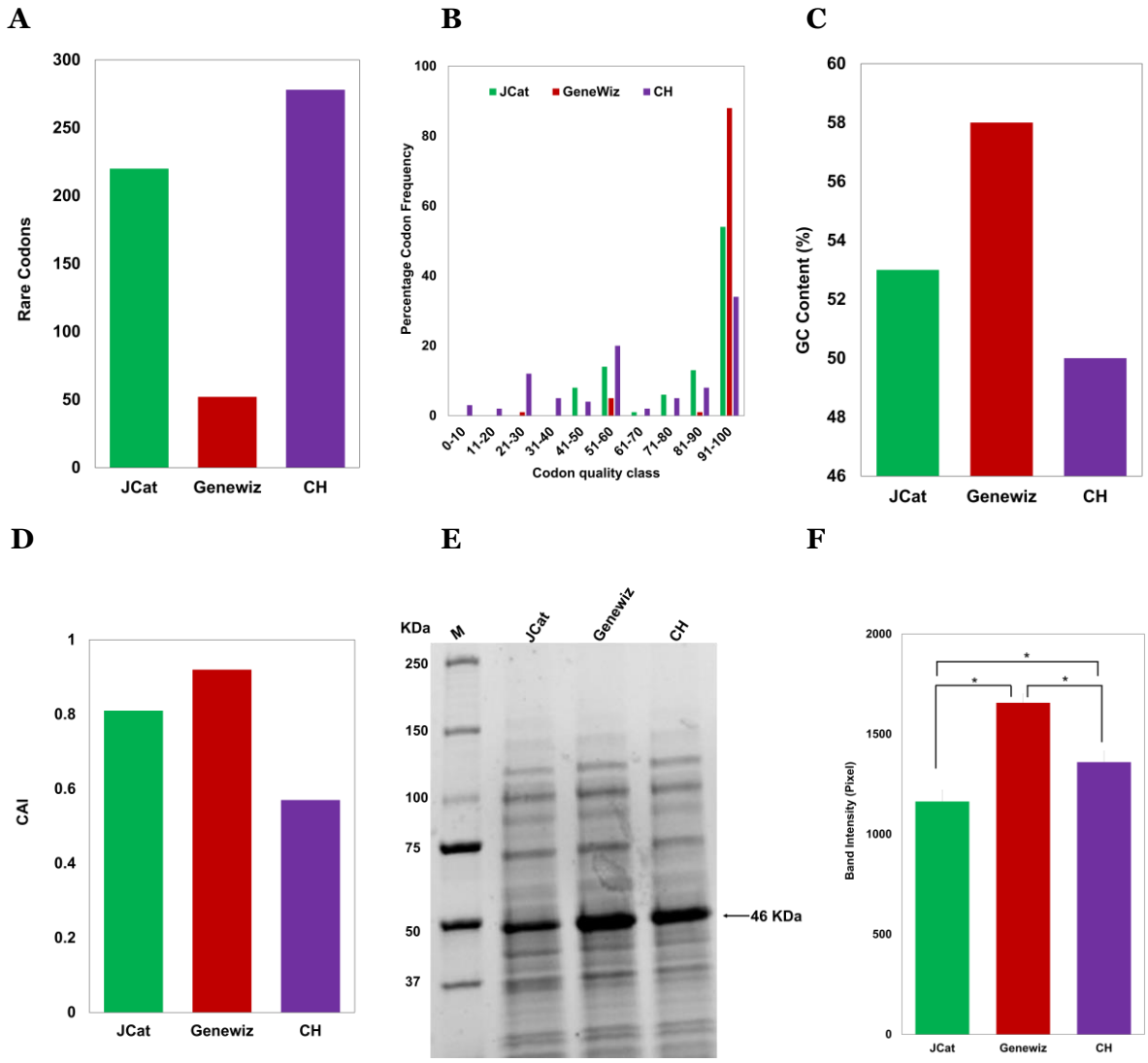


Fig 1. Effects of different parameters influenced by the genetic code substitution. (A) Rare codon. (B) Codon quality class. (C) percentage GC content. (D) Codon Adaptation Index. (E) gel image of the expressed protein, M – protein marker. (F) average band intensity and standard error bar calculated from the gel image in triplicate. The asterisk shows significant at $P<0.05$.

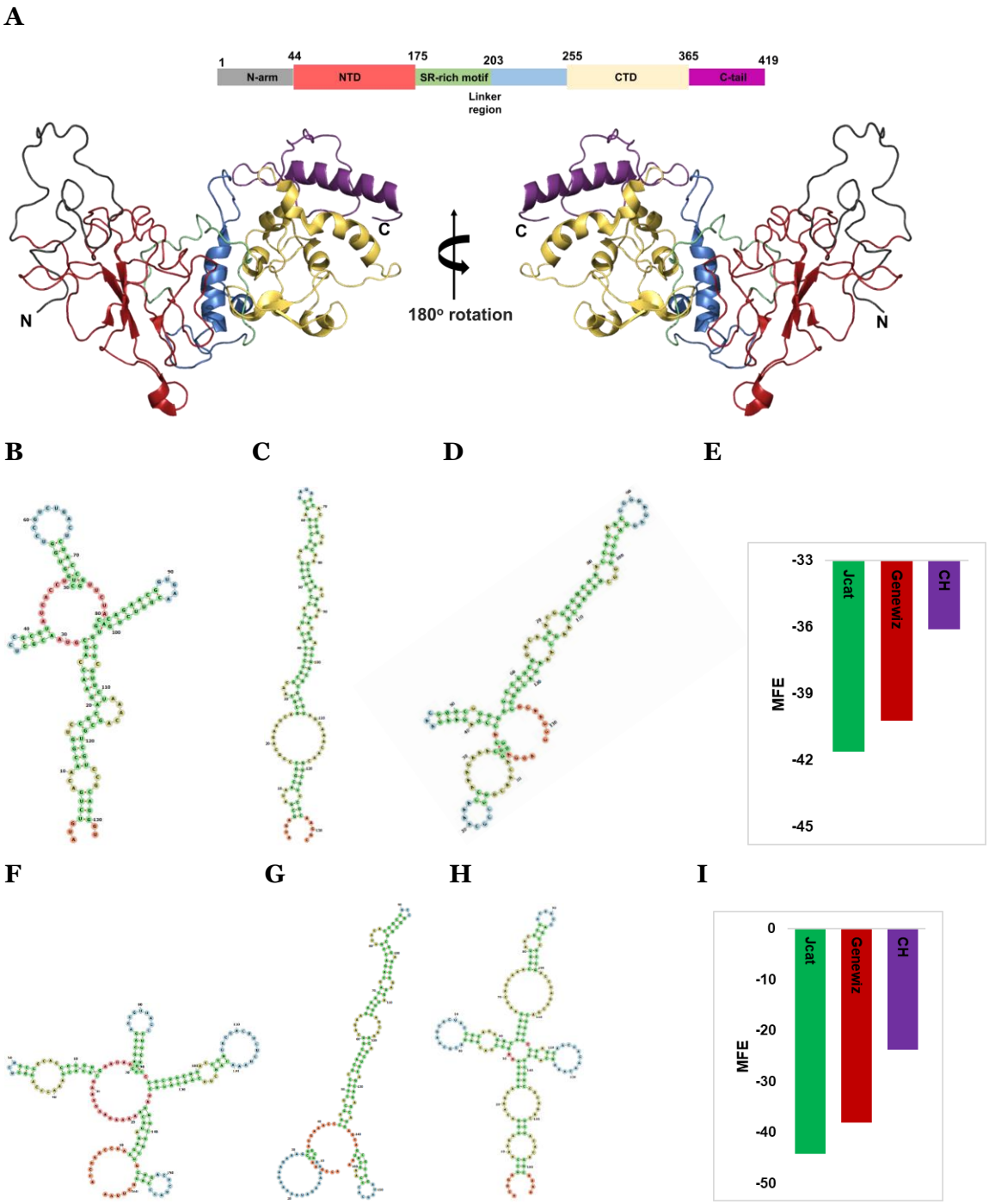
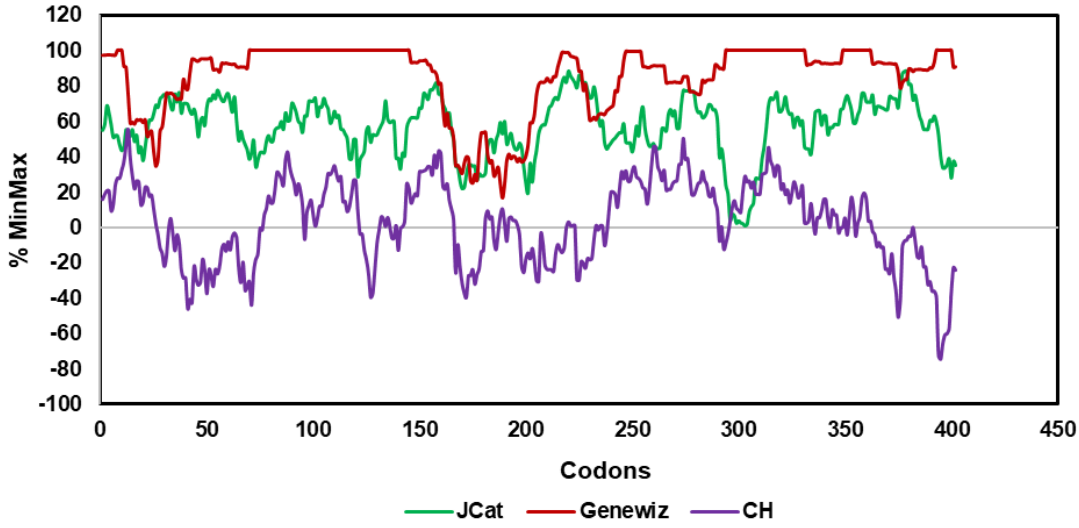
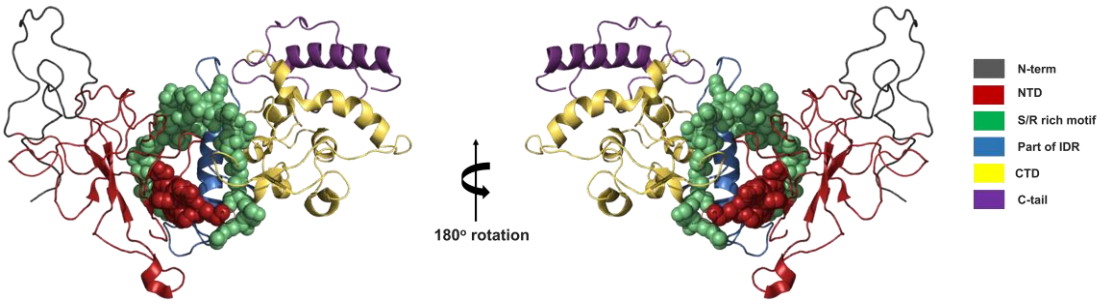


Fig 2. Genomic and structural arrangement of SARS-CoV-2 nucleocapsid protein and RNA. (A) top: schematic arrangement of nucleocapsid protein, highlighting the N-terminal domain (NTD) and C-terminal domain (CTD) flanking the SR-rich intrinsically disordered regions (IDRs). The NTD and CTD are flanked by other two IDRs, N-term and C-tail as illustrated. Down: The structure of the nucleocapsid protein (protein ID: 8FD5) visualized using PyMoL v3.0. RNA secondary structure of the N-term (132 nucleotide bases) for JCat (B), Genewiz (C), and CH (D), and minimum free energy (MFE) (E). RNA secondary structure of the C-term (165 nucleotide bases) for JCat (F), Genewiz (G) and CH (H), and minimum free energy (MFE) (I).

A



B



C

```
Genewiz     CCGAAAGGCUUCUAUGCCGAAGGUUCGCGCGGAGGGAGCCAAGCGUCAUCACGCAGCAGC     60
JCat        CCGAAAGGUUUCUACGCUGAAGGUUCUCGUGGUGGUUCUCAGGCUUCUUCUCGUUCUUCU     60
CH          CCCAAGGGAUUCUACGCAGAGGGGUCAAGGGGAGGGUCUCAAGCUAGCUCAAGAUCCAGU     60
            ** ** ** ***** ** ** ** **  * ** **    ** **    **  *

Genewiz     AGCCGUUCGAGGAACAGCAGCCGCAACAGCACUCCUGGUAGUUCACGU   108
JCat        UCUCGCUCUCGUAACUCUUCUCGCAACUCUACCCCGGGUUCUUCUCGU   108
CH          UCUAGGUCUAGAAACUCUUCGCGGAACUCUACACCCGGGUCAAGUCGG   108
                * **  * ***      ** ***   ** ** **       **
```

Fig 3. Predicted effect of the genetic code substitution on protein folding. (A) %MiniMax showing codon usage patterns of the three optimized sequences. (B) The structure of the nucleocapsid protein (protein ID: 8FD5) visualized using PyMoL v3.0. The spherical regions are SR rich region (green) and few residues P168 – E174 (red) upstream of the SR rich region. (C) sequence alignment of the optimized nucleotides at the regions in (B). The alignment was performed using ClustalW2 Multiple Sequence Alignment.

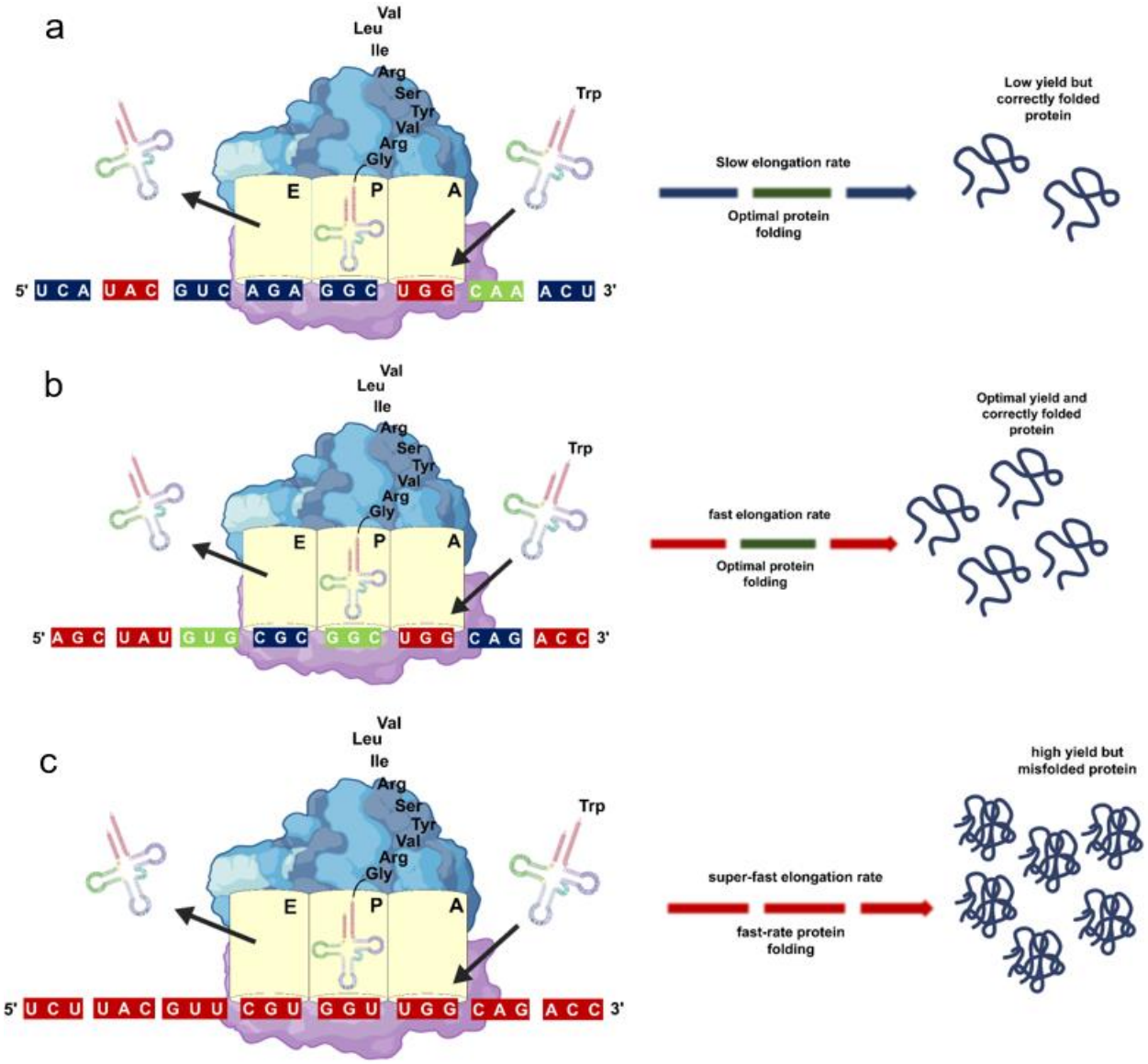


Fig 4. Effects of elongation rhythm pattern on the expression and folding of the protein from the three optimized tool. (A) codon harmonization (CH). (B) Genewiz. (C) JCat.