BIOGENESIS Jurnal Ilmiah Biologi

pISSN 2302-1616, eISSN 2580-2909
Vol 8, No. 2, December 2020, pp. 225-233
Available online http://journal.uin-alauddin.ac.id/index.php/biogenesis
DOI https://doi.org/10.24252/bio.v8i2.15805

Strategy for Designing the Synthetic Gene Encoding *Human papillomavirus* Major Capsid L1 Protein for Heterologous Expression in *Escherichia coli* System

Kartika Sari Dewi*, Wien Kusharyoto
Research Center for Biotechnology, Indonesian Institute of Sciences
JL. Raya Bogor Km. 46 Bogor, West Java, Indonesia. 16911
*Email: kart008@lipi.go.id

ABSTRACT. DNA is widely used to construct heterologously expressed genes. The adaptation of the codons to the host organism is necessary in order to ensure sufficient production of proteins. The GC content, codon identity and the mRNA from the translation site are also important in the design of the gene construct. This study performed a strategy for the design of synthetic gene encoding HPV52 L1 protein and several analyses at the genetic level to optimize its protein expression in the *Escherichia coli* BL21(DE3) host. The determination of the codon optimization was performed by collecting 75 HPV52 L1 protein sequences in the NCBI database. Furthermore, all the sequences were analyzed using multiple global alignments by Clustal Omega web server. Once the model was determined, codon optimization was performed using OPTIMIZER and the web server of the IDT codon optimization tool based on the *E. coli* B. The generated open reading frame (ORF) sequence was analyzed using Restriction mapper web server to choose the restriction site for facilitating the cloning stage, which is adjusted for pJExpress414 expression vector. To maximize the protein expression level, the mRNA secondary structure analysis around the ribosome binding site (rbs) was performed. A slight modification at the 5'-terminal end waa carried out in order to get more accessible rbs and increasing mRNA folding free energy. Finally, the construction of the synthetic gene was confirmed to ensure that no mutation occurs in the protein and to calculate its Codon Adaptation Index (CAI) and GC content. The above strategy, which leads to a good ORF sequence with the value of the free mRNA folding energy around rbs, is -5.5 kcal / mol, CAI = 0.787 and GC content 49.5%. This result is much better than its original gene. This result is much better compared to its native gene. Theoretically it is possible that this synthetic gene construct generates a high level protein expression in *E. coli* BL21 (DE3) under the regulation of the T7 promoter.

Keywords: codon adaptation index; codon optimization; GC-content; heterologous expression; mRNA secondary structure

Article History: Received 09 September 2020; Received in revised form 30 September 2020; Accepted 17 November 2020; Available online 30 December 2020

How to Cite This Article: Sari Dewi KS, Kusharyoto W. 2020. Strategy for designing the synthetic gene encoding *Human papillomavirus* major capsid L1 protein for heterologous expression in *Escherichia coli* system. *Biogenesis: Jurnal Ilmiah Biologi*. vol 8(2): 225-233. doi: https://doi.org/10.24252/bio.v8i2.15805.

INTRODUCTION

Nowadays, synthetic DNA is widely used to construct the gene for heterologous protein expression. Adequate information from genome and metagenome sequencing projects provide virtual DNA, which can be further exploited. The modification of the native DNA sequences is mostly introduced into the synthetic gene to maximize the expression, especially in the heterologous host (Quan *et al*., 2011; Gaspar *et al*., 2012; Luo *et al*., 2016). Several factors should be carefully considered when designing the construct, including codon usage, codon identity, GC-content, mRNA folding free energy, especially around the ribosome binding site (RBS). These play a role in the production of high-level recombinant proteins in *Escherichia coli* (Rosano & Ceccarelli, 2014; Ghavim *et al*., 2017; Dewi & Fuad, 2020).

The final strategy for the design of synthetic gene is to mimic the characteristics of the natural highly-expressed gene of the native host that are relevant for increasing the protein expression. Many researchers have reported that synonymous codons substitution of rare codons in the introduced genes can increase the protein yield dramatically (Chaney & Clark, 2015; Wang *et al*., 2015; Buhr *et al*., 2016). Rare codons are associated with ribosome stalling leading to truncated translation (Handa *et al*., 2011; Guimaraes *et al*., 2020). In addition, synonymous codon changes can influence the protein stability and conformation (Saunders & Deane, 2010; Hunt *et al*., 2014; Mauro & Chappell, 2014). Therefore, codon optimization plays an important role, mainly when the heterologous protein expression was performed.

Human papillomavirus (HPV) is a non-enveloped, double-stranded DNA virus that infects a variety of hosts, with a preference for epithelial cells. High-risk HPV infection is a major cause of cervical cancer development, which was the fourth most common cancer in women worldwide (Yang *et al*., 2017; Arbyn *et al*., 2020). The HPV L1 protein is the major capsid protein with a molecular weight of 55-60 kDa that can be expressed in many expression systems and formed virus-like particles (VLP), which resemble the native HPV morphologically without the assistance of L2 minor capsid protein (Buck *et al*., 2013). VLPs are known to retain the native epitopes of viral particles. Their high immunogenicity can induce the production of neutralization antibodies against HPV infection (Schiller *et al*., 2010; Plummer & Manchester, 2011; Tumban *et al*., 2012). In addition, the VLPs are safe and have no potential cancerogenic to become the primary candidate for HPV vaccines because they do not contain the viral genome (Li *et al*., 2016; Huang *et al*., 2017).

There are many online software packages available for performing codon optimization for designing the synthetic gene. EuGene (Gaspar *et al*., 2012), Parts & pools (Marchisio, 2014), Codon Optimization OnLine (COOL) (Chin *et al*., 2014), etc. are the newest software in the last 10 years. The previous software also still be used for gene design. In this study, the synthetic gene encoding HPV52 L1 protein was design for the development of an *E. coli*-based HPV vaccine. Not all available software was used, though it can be an alternative for designing other synthetic genes. In addition, the codon optimization and mRNA secondary structure formation near the translation initiation region become the primary concern. To the best of our knowledge, this study is the only one that discusses the codon optimization strategy and the construction of the synthetic gene encoding L1 major capsid protein of HPV52 for *E. coli* expression system.

MATERIALS AND METHODS

Determination of the HPV52 L1 protein sequence. About 75 sequences of HPV52 L1 protein from various countries were collected from the National Centre for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/). All the sequences were then subjected to multiple alignment analysis using Clustal Omega web server (https://www.ebi.ac.uk/Tools/msa/clustalo/) (Sievers & Higgins, 2014). The HPV52 L1 protein sequence having 100% identity discovered from this analysis was selected for a template in codon optimization step.

Codon optimization for *E. coli* **expression**. The HPV52 L1 protein sequence was selected for codon optimization by using OPTIMIZER web server (http://genomes.urv.es/OPTIMIZER/) (Puigbò *et al*., 2007). Codon usage table of *E. coli* B used was used as a reference for accessing Kazusa codon usage database (http://www.kazusa.or.jp/codon/) (Nakamura *et al*., 2000). The “Guided random” method was selected for codon optimization. Generated codon-optimized sequence having a codon adaptation index (CAI) value near 0.8 and GC-content near 51.06 % was chosen for further optimization.

Selection of restriction site for facilitating the cloning stage. In this study, pJExpress414 plasmid with T7 promoter was selected as an expression vector for gene cloning. RestrictionMapper web server (http://www.restrictionmapper.org/) was used to determine the restriction sites located inside the codon-optimized gene. When the desired restriction sites were discovered, they were removed using the feature available in OPTIMIZER software (Puigbò *et al*., 2007). After the digesting enzymes were selected, then the expression cassette of pJExpress414_HPV52L1 recombinant plasmid was arranged.

Determination of transcription start site (TSS). The first nucleotide of transcript (+1) was predicted using Neural Network Promoter Prediction (https://www.fruitfly.org/seq_tools/promoter.h tml). About 115 nt before and 115 nt after the RBS was subjected to this analysis. The possible promoter sequence and +1 were selected based on the highest score.

mRNA secondary structure analysis. The secondary structure formation of mRNA around the ribosome binding site was predicted using RNAstructure web server (https://rna.urmc.rochester.edu/RNAstructure Web/Servers/Predict1/Predict1.html) (Bellaousov *et al*., 2013). The folding free energy of mRNA secondary structure was also observed. Synonymous codon substitution near the start codon was performed to increase the MFE of mRNA secondary structure.

Confirmation of final synthetic gene construct. The final codon-optimized sequence was translated into proteins using ExPASy translate tools (https://web.expasy.org/translate/) (Gasteiger *et al*., 2003). Translated protein was subjected to pairwise sequence alignment analysis by Clustal Omega utilizing a template of the HPV52 L1 protein sequence as a tandem. If no mutation was confirmed, then the CAI value of the codon-optimized sequence was calculated using CAIcal (http://genomes.urv.es/CAIcal/E-CAI/) (Puigbò *et al*., 2008).

RESULTS AND DISCUSSION

Determination of the HPV52 L1 protein sequence. In the determination of HPV52 L1 amino acid sequence, about 75 samples were accessed from NCBI database and collected in Fasta format. Table 1 showed the accession number of HPV52 L1 protein sequence used in this experiment. To find the most promising L1 protein sequence as a template, multiple sequence alignment analysis was performed using a Clustal Omega web server. The percentage identity matrix created by Clustal2.1 demonstrated that all sequences shared >99% identity (data are not shown). However, in this study, only the sequences that shared 100% identity were selected as a template for codon optimization. One is accession number APQ44871.1.

Table 1. The accession number of HPV52 L1 protein sequence accessed from NCBI database.

Country	Accession Number	Country	Accession Number	Country	Accession Number
South Korea	APQ44863.1	USA	AIF71419.1	North-western Guangdong	AMN09991.1
	APQ44881.1		AIF71420.1		AMN09990.1
	APQ44864.1		AIF71421.1	Eastern Guangdong	AFC35275.1
	APQ44865.1		AIF71423.1		AFC35272.1
	APQ44866.1		AIF71418.1		AFC35271.1
	APQ44867.1		AIF71417.1		AFC35256.1
	APQ44868.1	Croatia	AIF71429.1	Southwest China	AML81012.1
	APQ44870.1		AIF71425.1		AML80961.1
	APQ44871.1		AIF71428.1		AML80962.1
	APQ44872.1		AIF71427.1		AML80964.1
Japan	BBD06694.1	Hongkong	AIF71404.1		AML80965.1
	BBD06686.1		AIF71403.1		AML80967.1
	BBD06702.1		AIF71401.1		AML80968.1
	BBD06678.1		AIF71400.1		AML80969.1
	BBA19957.1		AIF71399.1		AML80970.1
	BBA19973.1		AIF71397.1	UK	AIF71434.1
	BBA19949.1	Canada	ABU55796.1		AIF71410.1
	BBA19941.1		ABU55795.1		AIF71409.1
	BBA19933.1		ABU55794.1		AIF71408.1
	BBA19917.1.		ABU55793.1	Zimbabwe	AIF71435.1
	BBA19789.1		ABU55792.1	Iran	QDH43417.1
	BBA19797.1		ABU55791.1	Mexico	AIF71413.1
	BBA19773.1		ABU55790.1	Netherland	CAA52590.1
	BBA19757.1		ABU55789.1		
	BBA19741.1	Italy	AIF71412.1		
			AIF71411.1		

A study conducted by Wei *et al*. (2018) showed that the elimination of several amino acids at the N-terminal of HPV52 L1 proteins improves their solubility in *E. coli* expression system without affecting their immunogenicity. Therefore, the codon optimization lies in amino acids number 15-503, which is counted from the second Methionine. Fig. 1 showed the amino acid sequence with Accession Number APQ44871.1. The amino acid sequence that was subjected to codon optimization was shown in yellow.

```
MVQILFYILVIFYYVAGVNVFHIFLQMSVWRPSEATVYLPPVPVSKVVSTDEYVSRTSIYYYAGSS
RLLTVGHPYFSIKNTSSGNGKKVLVPKVSGLQYRVFRIKLPDPNKFGFPDTSFYNPETQRLVWACT
GLEIGRGQPLGVGISGHPLLNKFDDTETSNKYAGKPGIDNRECLSMDYKQTQLCILGCKPPIGEHW
GKGTPCNNNSGNPGDCPPLQLINSVIQDGDMVDTGFGCMDFNTLQASKSDVPIDICSSVCKYPDYL
QMASEPYGDSLFFFLRREQMFVRHFFNRAGTLGDPVPGDLYIQGSNSGNTATVQSSAFFPTPSGSM
VTSESQLFNKPYWLQRAQGHNNGICWGNQLFVTVVDTTRSTNMTLCAEVKKESTYKNENFKEYLRH
GEEFDLQFIFQLCKITLTADVMTYIHKMDATILEDWQFGLTPPPSASLEDTYRFVTSTAITCQKNT
PPKGKEDPLKDYMFWEVDLKEKFSADLDQFPLGRKFLLQAGLQARPKLKRPASSAPRTSTKKKKVK
R
```

Fig. 1. Amino acids sequence of HPV52 L1 protein from accession number APQ44871.1. The yellow color indicates the codon-optimized region.

Codon optimization for *E. coli* expression and selection of restriction site for facilitating the cloning stage. The codon optimization was carried out with the OPTIMIZER web server. Using the "Guided Random" method, codons were selected at random with probabilities obtained from the codon usage table of *E. coli* B. Generated codon-optimized sequence using this software gave the highest CAI value of 0.739 and GC-content of 49.2% (Table 2).

In this experiment, the pJExpress414 plasmid was used as an expression vector. The *Sal*I and *Nco*I restriction sites were selected at 5' and 3' end to facilitate the cloning stage, respectively. However, by using the RestrictionMapper web server, it was reported that those restriction sites were inside the codon-optimized gene (data not shown). In addition, there were also repeated bases of A/T (Table 2). Therefore, both restriction sites and A/T repeated bases were removed using the feature discovered in OPTIMIZER web server.

To increase the CAI and GC-content, the DNA sequence was further optimized manually using IDT codon optimization tool (https://sg.idtdna.com/CodonOpt). Table 2 showed that the modified codon-optimized sequence has a higher CAI and GC-content, without *Sal*I and *Nco*I restriction sites inside the gene.

In this study, the codon optimization of HPV52 L1 open reading frame (ORF) was carried out using OPTIMIZER web server according to *E. coli* B codon usage table. Furthermore, the codon usage bias in a gene towards frequent codons is often measured by Codon Adaptation Index (CAI). CAI reflected how well our optimized-codon sequences will adapt to the new host protein expression machinery. The CAI value ranges from 0 to 1 based on how close the foreign gene sequences against the reference gene sequences. However, the CAI value = 1 means "one amino acid – one codon" approach has several disadvantages, including translational error, tRNA depletion, frame-shift, hard to avoid repetitive elements, and hard to introduce or eliminate the restriction sites inside the ORF. Therefore, in this experiment wa "Guided random" method was used to optimize the gene instead of the "Most frequent" method provided by Puigbò *et al*. (2007), Welch *et al.* (2009), and Villalobos *et al.* (2016). The "Guided random" flexibility makes it easier to remove the desired restriction site and A/T repeated bases located inside the gene (Table 2).

Table 2. The sequences of the codon-optimized gene using OPTIMIZER tool and its modification afterward. Green color indicates the *Sal*I and *Nco*I restriction enzymes. Yellow color indicates the repeated A/T ≥ 6 bases.

Type	Sequences	CAI	%GC
Optimized	ATGCCGGTACCGGTTTCTAAAGTGGTCTCAACAGATGAATATGTTAGTCGC ACCAGCATCTATTATTATGCGGGAAGCTCGAGGCTCCTGACCGTCGGCCAC CCGTATTTCAGCATCAAAAATACCAGCTCGGGCAATGGTAAAAAGTTCTG GTGCCGAAAGTGTCGGGTTTACAGTATAGAGTGTTCCGCATTAAATTACCT GACCCGAATAAGTTCGGTTTTCCAGATACCTCGTTTTATAATCCCGAAACC CAGCGCCTGGTGTGGGCCTGCACAGGTTTGGAAATTGGCCGTGGCCAACCG CTGGGTGTGGGTATTAGCGGCCACCCATTGTTAAATAAATTCGATGATACC GAAACTAGTAACAAGTACGCCGGGAAACCGGGTATCGACAATCGCGAGTGT CTCTCCATGGATTATAAACAAACTCAGCTGTGCATCCTGGGTTGCAAGCCG CCGATCGGAGAACATTGGGGAAAAGGCACCCCGTGCAATAATAATTCGGGG AATCCGGGGGATTGCCCGCCCCTCCAACTGATCAATAGTGTCATTCAAGAT GGTGACATGGTGGATACCGGCTTTGGTTGCATGGATTTTAATACGCTGCAA GCATCGAAAAGTGACGTGCCGATTGATATTTGCTCCTCGGTGTGTAAATAC CCAGATTATCTGCAGATGGCATCAGAACCGTATGGTGATAGTCTGTTTTTT TTTTTACGCCGTGAACAGATGTTTGTGCGCCATTTTTTTAATCGCGCAGGC ACCCTGGGTGACCCGGTCCCGGGCGATTTATATATTCAAGGCAGTAATAGC GGCAACACCGCGACGGTCCAGTCTTCGGCATTTTTCCCCACCCCGAGCGGC AGTATGGTGACCTCTGAATCCCAACTCTTTAACAAACCGTATTGGCTTCAA CGCGCCCAGGGTCATAACAACGGCATCTGCTGGGGCAACCAGCTTTTTGTG ACCGTCGTGGACACCACCCGTTCCACAAACATGACCTTGTGTGCGGAAGTA AAAAAGGAAAGCACCTACAAAAACGAAAATTTTAAGGAATATCTTCGCCAT GGTGAAGAGTTTGATTTGCAATTTATCTTTCAACTGTGCAAAATCACGTTA ACCGCCGACGTTATGACCTATATTCATAAAATGGACGCCACGATACTTGAA GATTGGCAGTTTGGGCTTACGCCCCCTCCGAGCGCGTCGCTGGAAGACACG TATCGCTTTGTAACTAGCACCGCGATCACCTGCCAGAAAAACACGCCGCCG AAAGGCAAGGAAGATCCGCTGAAAGACTATATGTTTTGGGAAGTCGACCTG AAGGAGAAATTTAGTGCAGATTTGGATCAATTTCCTCTGGGACGCAAATTC CTGCTGCAGGCAGGTCTCCAGGCACGCCCTAAACTGAAGCGCCCGGCGTCC AGCGCACCGCGTACGAGCACCAAGAAAAAAAAAGTGAAACGC	0.739	49.2
Modified	ATGCCAGTGCCGGTGTCTAAAGTTGTTTCCACGGATGAATATGTGAGCCGT ACATCGATATACTACTATGCCGGCTCGTCGCGTCTGTTGACCGTGGGTCAT CCGTATTTCTCTATCAAAAATACCTCGAGCGGCAATGGCAAAAAGGTGTTG GTGCCGAAGGTGTCGGGTTTGCAGTACCGTGTGTTTCGCATTAAACTCCCG GATCCGAATAAGTTTGGTTTCCCGGATACCAGCTTTTATAACCCGGAGACG CAGCGCCTGGTGTGGGCGTGCACGGGCTTGGAGATTGGCCGGGGTCAACCG CTGGGCGTTGGCATTAGCGGGCACCCGTTGCTGAACAAATTTGATGACACC GAGACCTCGAATAAATACGCGGGTAAACCAGGTATTGACAACCGTGAATGC CTGAGCATGGACTATAAACAGACCCAACTGTGCATTTTGGGCTGCAAACCG CCGATTGGCGAGCATTGGGGTAAAGGTACCCCGTGTAATAATAACAGTGGC AATCCAGGCGATTGTCCGCCATTACAGCTCATCAACTCCGTCATCCAAGAT GGTGACATGGTGGACACGGGTTTCGGCTGTATGGATTTTAATACTTTGCAG GCCTCGAAAAGTGACGTTCCTATTGACATTTGCAGTTCGGTGTGTAAATAT CCAGATTATCTGCAGATGGCGAGCGAACCTTACGGCGATTCGTTATTCTTT TTCCTGCGTCGCGAGCAGATGTTTGTGCGTCATTTCTTTAACCGTGCAGGG ACGTTGGGGGACCCGGTACCGGGCGATTTGTATATTCAAGGTAGTAACAGC GGCAATACCGCCACGGTGCAGTCGAGCGCGTTCTTTCCAACCCCGTCGGGC AGTATGGTTACCTCCGAAAGTCAATTATTCAACAAACCGTATTGGCTGCAA CGTGCGCAAGGCCATAATAATGGTATTTGCTGGGGTAATCAACTGTTTGTG ACCGTGGTGGACACGACGCGCAGTACGAATATGACCCTGTGTGCCGAGGTT AAGAAGGAATCGACGTATAAAAACGAAAATTTCAAGGAATATTTACGTCAT GGTGAGGAGTTCGACCTGCAGTTTATTTTCCAGTTGTGCAAAATCACGCTG ACCGCAGATGTGATGACCTATATTCATAAAATGGATGCAACTATCCTGGAG GATTGGCAGTTCGGATTGACCCCACCGCCTAGCGCAAGTCTGGAAGATACG TATCGCTTTGTGACCAGCACCGCAATTACTTGCCAGAAAAATACGCCACCG AAGGGCAAAGAAGATCCATTAAAAGACTACATGTTCTGGGAGGTTGACTTG AAGGAAAAATTTAGTGCCGATTTGGATCAGTTCCCTCTGGGCAGAAAGTTC TTACTGCAAGCAGGGTTACAGGCGCGTCCGAAATTGAAACGTCCAGCAAGC TCAGCCCCGCGCACCTCGACTAAGAAAAAGAAAGTGAAACGC	0.787	49.4

Determination of transcription start site (TSS) and mRNA secondary structure analysis around the ribosome binding site. The pJExpress414 plasmid was used as an expression vector. The expression cassette of pJExpress414_HPV52L1 recombinant plasmid was arranged as seen in Fig. 2. The TSS was predicted using Neural Network Promoter Prediction web server. After the first nucleotide of transcript (+1) was determined, the mRNA sequence was subsequently decided, which will be subjected to secondary structure analysis around the area of RBS.

[…TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTT
TAACTTT**AGGAGGTAAAAAAAGTCGACATGCCAGTGCCGGTGTCTAAAGTTGTTTCCACGGATGAA**
TATGTGAGCCGTACATCGATATACTACTATGCCGGCTCGTCGCGTCTGTTGACCGTGGGTC…. TAA
CCATGGGTTTTTTG..]

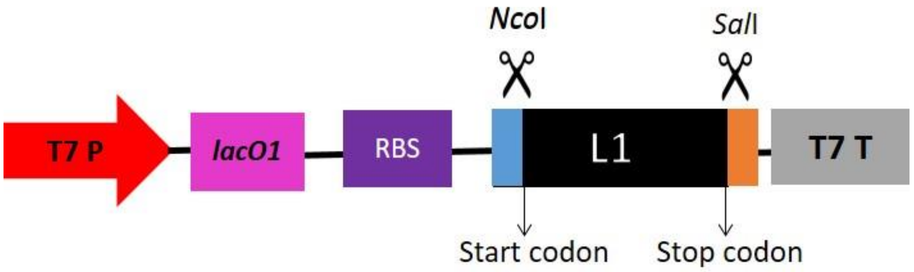


Fig. 2. The expression cassette of pJExpress414_HPV52L1 recombinant plasmid. Red colour = T7 Promoter; magenta = lac operator; purple = ribosome binding site; blue = *Nco*I restriction site; black = L1 gene; orange = *Sal*I restriction site; grey: T7 terminator.

The sequence used for the analysis of the mRNA secondary structure started after codon start from +1 to seven amino acids (Table 3). The result showed that the original sequence has low mRNA folding free energy. In addition, the RBS sequence was shown in pairs with the downstream nucleotides (Fig. 3a). Therefore, the synonymous codon substitution was performed in order to increase the mRNA folding free energy and get unpaired RBS sequences. Fig. 3b showed that the mRNA folding free energy was increased from -7.3 to -5.5 kcal/mol and the RBS sequence was exposed after synonymous codon substitution was performed. Our previous experiment showed that mRNA folding free energy of -7.3 kcal/mol in the area around the RBS was enough to produce high protein expression.

Predicting the secondary structure of mRNA secondary is usually completed by finding the lowest folding free energy, which is the most likely folding ensemble structure (Reuter & Mathews, 2010). The increased folding free energy makes a more unstable mRNA secondary structure, which is preferred in *E. coli* expression. In addition, the open RBS makes it accessible for the ribosome to start the protein synthesis, which is expected to improve the expression of recombinant proteins in the *E. coli* system.

Table 3. The mRNA sequences around the ribosome binding site used for mRNA used for the analysis of the secondary structure using RNAstructure web server. Transcription start (+1) is shown in a larger font, and blue color indicates the substituted codons.

Type	mRNA sequences	mRNA folding free energy (kcal/mol)
Before synonymous codon substitution	TGTTTAACTTTAGGAGGTAAAAAAAGTCGAC ATG CCA GTG CCG GTG TCT AAA GTT	-7.3
After synonymous codon substitution	TGTTTAACTTTAGGAGGTAAAAAAAGTCGAC ATG CCA GTG CCG GTC AGC AAA GTG	-5.5

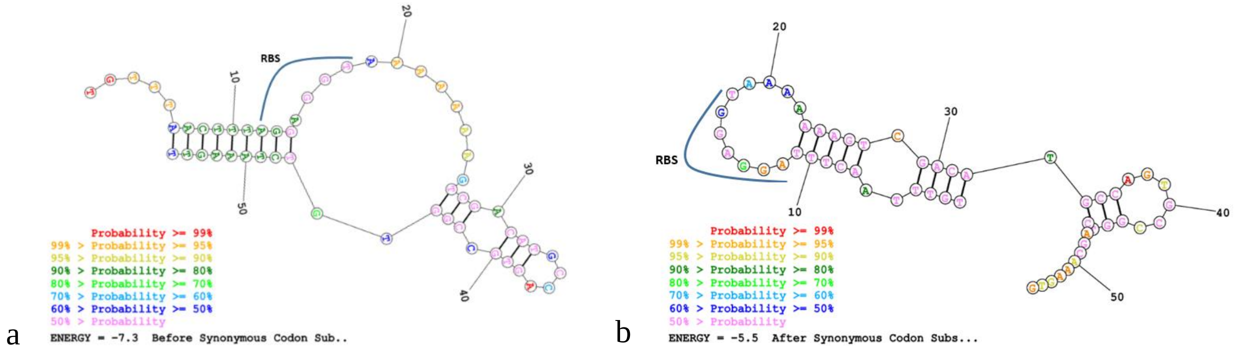
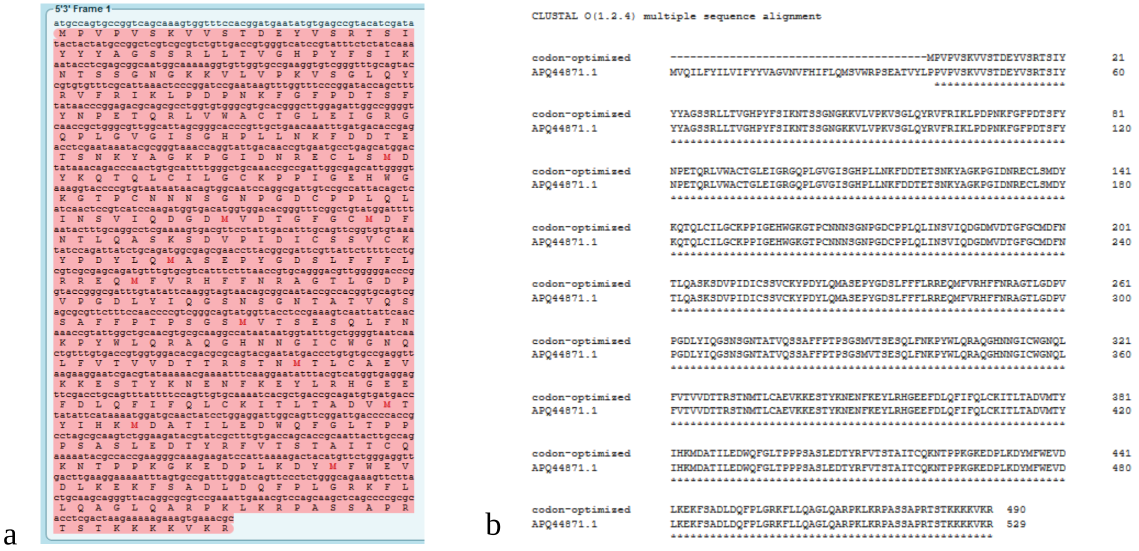


Fig. 3. The mRNA secondary structure: a. before analysis; b. after analyzed by N-terminal synonymous codon substitution using RNAstructure web server.

The expression of heterologous protein in *E. coli* was limited, especially at the initiation of translation. A strong mRNA downstream of the translation start site is known and can interfere with expression regardless of the general CAI or GC content (Gu *et al*., 2010; Gingold & Pilpel, 2011; Ghavim *et al*., 2017). Our previous study confirmed that the synonymous codon substitution immediately after the initial codon can significantly improve the expression of human granulocyte colony-stimulating factor (hG-CSF) recombinant proteins in *E. coli* system (Dewi & Fuad, 2020).



```
atgccagtgccggtcagcaaagtggtttccacggatgaatatgtgagccgtacatcgata
 M  P  V  P  V  S  K  V  V  S  T  D  E  Y  V  S  R  T  S  I
tactactatgccggctcgtcgcgtctgttgaccgtgggtcatccgtatttctctatcaaa
 Y  Y  Y  A  G  S  S  R  L  L  T  V  G  H  P  Y  F  S  I  K
aatacctcgagcggcaatggcaaaaaggtgttggtgccgaaggtgtcgggtttgcagtac
 N  T  S  S  G  N  G  K  K  V  L  V  P  K  V  S  G  L  Q  Y
cgtgtgtttcgcattaaactcccggatccgaataagtttggtttcccggataccagcttt
 R  V  F  R  I  K  L  P  D  P  N  K  F  G  F  P  D  T  S  F
tataacccggagacgcagcgcctggtgtgggcgtgcacgggcttggagattggccggggt
 Y  N  P  E  T  Q  R  L  V  W  A  C  T  G  L  E  I  G  R  G
caaccgctgggcgttggcattagcgggcacccgttgctgaacaaatttgatgacaccgag
 Q  P  L  G  V  G  I  S  G  H  P  L  L  N  K  F  D  D  T  E
acctcgaataaatacgcgggtaaaccaggtattgacaaccgtgaatgcctgagcatggac
 T  S  N  K  Y  A  G  K  P  G  I  D  N  R  E  C  L  S  M  D
tataaacagacccaactgtgcatttgggctgcaaaccgccgattggcgagcattggggt
 Y  K  Q  T  Q  L  C  I  L  G  C  K  P  P  I  G  E  H  W  G
aaaggtaccccgtgtaataataacagtggcaatccaggcgattgtccgccattacagctc
 K  G  T  P  C  N  N  N  S  G  N  P  G  D  C  P  P  L  Q  L
atcaactccgtcatccaagatggtgacatggtggacacgggtttcggctgtatggatttt
 I  N  S  V  I  Q  D  G  D  M  V  D  T  G  F  G  C  M  D  F
aatactttgcaggcctcgaaaagtgacgttcctattgacatttgcagttcggtgtgtaaa
 N  T  L  Q  A  S  K  S  D  V  P  I  D  I  C  S  S  V  C  K
tatccagattatctgcagatggcgagcgaaccttacggcgattcgttattctttttcctg
 Y  P  D  Y  L  Q  M  A  S  E  P  Y  G  D  S  L  F  F  F  L
cgtcgcgagcagatgtttgtgcgtcatttctttaaccgtgcagggacgttgggggacccg
 R  R  E  Q  M  F  V  R  H  F  F  N  R  A  G  T  L  G  D  P
gtaccgggcgatttgtatattcaaggtagtaacagcggcaataccgccacggtgcagtcg
 V  P  G  D  L  Y  I  Q  G  S  N  S  G  N  T  A  T  V  Q  S
agcgcgttctttccaaccccgtcgggcagtatggttacctccgaaagtcaattattcaac
 S  A  F  F  P  T  P  S  G  S  M  V  T  S  E  S  Q  L  F  N
aaaccgtattggctgcaacgtgcgcaaggccataataatggtatttgctggggtaatcaa
 K  P  Y  W  L  Q  R  A  Q  G  H  N  N  G  I  C  W  G  N  Q
ctgtttgtgaccgtggtggacacgacgcgcagtacgaatatgaccctgtgtgccgaggtt
 L  F  V  T  V  V  D  T  T  R  S  T  N  M  T  L  C  A  E  V
aagaaggaatcgacgtataaaaacgaaaatttcaaggaatatttacgtcatggtgaggag
 K  K  E  S  T  Y  K  N  E  N  F  K  E  Y  L  R  H  G  E  E
ttcgacctgcagtttatttccagttgtgcaaaatcacgctgaccgcagatgtgatgacc
 F  D  L  Q  F  I  F  Q  L  C  K  I  T  L  T  A  D  V  M  T
tatattcataaaatggatgcaactatcctggaggattggcagttcggattgaccccaccg
 Y  I  H  K  M  D  A  T  I  L  E  D  W  Q  F  G  L  T  P  P
cctagcgcaagtctggaagatacgtatcgctttgtgaccagcaccgcaattacttgccag
 P  S  A  S  L  E  D  T  Y  R  F  V  T  S  T  A  I  T  C  Q
aaaaatacgccaccgaagggcaaagaagatccattaaaagactacatgttctgggaggtt
 K  N  T  P  P  K  G  K  E  D  P  L  K  D  Y  M  F  W  E  V
gacttgaaggaaaaatttagtgccgatttggatcagttccctctgggcagaaagttctta
 D  L  K  E  K  F  S  A  D  L  D  Q  F  P  L  G  R  K  F  L
ctgcaagcagggttacaggcgcgtccgaaattgaaacgtccagcaagctcagccccgcgc
 L  Q  A  G  L  Q  A  R  P  K  L  K  R  P  A  S  S  A  P  R
acctcgactaagaaaaagaaagtgaaacgc
 T  S  T  K  K  K  K  V  K  R
```

```
CLUSTAL O(1.2.4) multiple sequence alignment


codon-optimized      -------------------------------------MPVPVSKVVSTDEYVSRTSIY      21
APQ44871.1           MVQILFYILVIFYYVAGVNVFHIFLQMSVWRPSEATVYLPPVPVSKVVSTDEYVSRTSIY      60
                                                            ********************

codon-optimized      YYAGSSRLLTVGHPYFSIKNTSSGNGKKVLVPKVSGLQYRVFRIKLPDPNKFGFPDTSFY      81
APQ44871.1           YYAGSSRLLTVGHPYFSIKNTSSGNGKKVLVPKVSGLQYRVFRIKLPDPNKFGFPDTSFY      120
                     ************************************************************

codon-optimized      NPETQRLVWACTGLEIGRGQPLGVGISGHPLLNKFDDTETSNKYAGKPGIDNRECLSMDY      141
APQ44871.1           NPETQRLVWACTGLEIGRGQPLGVGISGHPLLNKFDDTETSNKYAGKPGIDNRECLSMDY      180
                     ************************************************************

codon-optimized      KQTQLCILGCKPPIGEHWGKGTPCNNNSGNPGDCPPLQLINSVIQDGDMVDTGFGCMDFN      201
APQ44871.1           KQTQLCILGCKPPIGEHWGKGTPCNNNSGNPGDCPPLQLINSVIQDGDMVDTGFGCMDFN      240
                     ************************************************************

codon-optimized      TLQASKSDVPIDICSSVCKYPDYLQMASEPYGDSLFFFLRREQMFVRHFFNRAGTLGDPV      261
APQ44871.1           TLQASKSDVPIDICSSVCKYPDYLQMASEPYGDSLFFFLRREQMFVRHFFNRAGTLGDPV      300
                     ************************************************************

codon-optimized      PGDLYIQGSNSGNTATVQSSAFFPTPSGSMVTSESQLFNKPYWLQRAQGHNNGICWGNQL      321
APQ44871.1           PGDLYIQGSNSGNTATVQSSAFFPTPSGSMVTSESQLFNKPYWLQRAQGHNNGICWGNQL      360
                     ************************************************************

codon-optimized      FVTVVDTTRSTNMTLCAEVKKESTYKNENFKEYLRHGEEFDLQFIFQLCKITLTADVMTY      381
APQ44871.1           FVTVVDTTRSTNMTLCAEVKKESTYKNENFKEYLRHGEEFDLQFIFQLCKITLTADVMTY      420
                     ************************************************************

codon-optimized      IHKMDATILEDWQFGLTPPPSASLEDTYRFVTSTAITCQKNTPPKGKEDPLKDYMFWEVD      441
APQ44871.1           IHKMDATILEDWQFGLTPPPSASLEDTYRFVTSTAITCQKNTPPKGKEDPLKDYMFWEVD      480
                     ************************************************************

codon-optimized      LKEKFSADLDQFPLGRKFLLQAGLQARPKLKRPASSAPRTSTKKKKVKR  490
APQ44871.1           LKEKFSADLDQFPLGRKFLLQAGLQARPKLKRPASSAPRTSTKKKKVKR  529
                     *************************************************
```

Fig. 4. The final synthetic gene construction: a. translation result of the final synthetic gene construction using ExPASy translate tool; b. sequences alignment using Clustal Omega tool showed that no mutation has occurred in protein level except for the intentional N-terminal truncation.

Confirmation of final synthetic gene construct. The final construction of the synthetic gene needs to be confirmed to ensure that no mutations occur at the amino acid level**.** The ExPASy translation tool was used to convert DNA sequences into protein sequences (Fig. 4a). The sequence alignment was performed using by Clustal omega using sequence with Accession Number APQ44871.1 as a reference. Fig. 4b showed that no mutation occurred in the protein sequences except for the intentional N-terminal truncation. In the end, the final CAI was calculated using CAIcal web server, resulting in CAI = 0.787 with GC-content 49.5%. This GC-content value is close to the required *E. coli* host (51.06%). The natural characteristic of the native HPV52 L1 gene is low GC-content. The CAI of native HPV52 L1 gene from Accession Number KY077853 was calculated using CAIcal web server with *E. coli* B codon usage table as a reference, and resulted in CAI = 0.616 with GC-content 39.7%. Therefore, the synthetic gene construct showed good improvement of both CAI and GC-content compared to its native gene sequence.

CONCLUSION

The design of the synthetic gene encoding HPV52 L1 protein using the strategy above showed a good result with mRNA folding free energy around rbs of -5.5 kcal/mol, CAI = 0.787 and GC-content 49.5%. This result is much better compared to its native HPV52 L1 gene. Theoretically, thr obtained synthetic gene is prospective to produce high-level protein expression in *E. coli* BL21(DE3) under the regulation of T7 promoter.

ACKNOWLEDGEMENT

This research was supported by the LPDP research grant from Indonesian Ministry of Finance, under the program of development of research cell bank (RCB) for HPV recombinant vaccine material during 2020-2021 fiscal years.

REFERENCES

Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, Bray F. 2020. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. *The Lancet of Global Health*. vol 8(2): 191–203. doi: https://doi.org/10.1016/S2214-109X(19)30482-6.

Bellaousov S, Reuter JS, Seetin MG, Mathews DH. 2013. RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Research*. vol 41: 471–474. doi: https://doi.org/10.1093/nar/gkt290.

Buck CB, Day PM, Trus BL. 2013. The papillomavirus major capsid protein L1. *Virology*. vol 445(1–2): 169–174. doi: https://doi.org/10.1016/j.virol.2013.05.038.

Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, Schwalbe H, Rodnina MV, Komar AA. 2016. Synonymous codons direct cotranslational folding toward different protein conformations. *Molecular Cell*. vol 61(3): 341–351. doi: https://doi.org/10.1016/j.molcel.2016.01.008.

Chaney JL, Clark PL. 2015. Roles for synonymous codon usage in protein biogenesis. *Annual Review of Biophysics*. vol 44: 143 – 166. doi: https://doi.org/10.1146/annurev-biophys-060414-034333.

Chin JX, Chung BKS, Lee DY. 2014. Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. *Bioinformatics*. vol 30(15): 2210–2212. doi: https://doi.org/10.1093/bioinformatics/btu192.

Dewi KS, Fuad AM. 2020. Improving the expression of human granulocyte colony stimulating factor in *Escherichia coli* by reducing the GC-content and increasing mRNA free folding energy at 5'-terminal end. *Advanced Pharmaceutical Bulletin*. vol 10(4): 610–616. doi: https://dx.doi.org/10.34172%2Fapb.2020.073.

Gaspar P, Oliveira JL, Frommlet J, Santos MA, Moura G. 2012. EuGene: maximizing synthetic gene design for heterologous expression. *Bioinformatics*. vol 28(20): 2683–2684. doi: https://doi.org/10.1093/bioinformatics/bts465.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*. vol 31(13): 3784–3788. doi: https://doi.org/10.1093/nar/gkg563.

Ghavim M, Abnous K, Arasteh F, Taghavi S, Nabavinia MS, Alibolandi M, Ramezani M. 2017. High level expression of recombinant human growth hormone in Escherichia coli: crucial role of translation initiation region. *Research in Pharmaceutical Sciences*. vol 12(2): 168–175. doi: https://dx.doi.org/10.4103%2F1735-5362.202462.

Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Molecular Systems Biology*. vol 7(1): 1–13. doi: https://doi.org/10.1038/msb.2011.14

Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Computational Biology*. vol 6(2): 1–8. doi: https://doi.org/10.1371/journal.pcbi.1000664.

Guimaraes JC, Mittal N, Gnann A, Jedlinski D, Riba A, Buczak K. Schmidt A, Zavolan M. 2020. A rare codon-based translational program of cell proliferation. *Genome Biology*. vol 21(1): 1–20. doi: https://doi.org/10.1186/s13059-020-1943-5.

Handa Y, Inaho N, & Nameki N. 2011. YaeJ is a novel ribosome-associated protein in Escherichia coli that can hydrolyze peptidyl–tRNA on stalled ribosomes. *Nucleic Acids Research*. vol 39(5): 1739–1748. doi: https://doi.org/10.1093/nar/gkq1097.

Huang X, Wang X, Zhang J, Xia N, Zhao Q. 2017. Escherichia coli-derived virus-like particles in vaccine development. *npj Vaccines*. vol 2(1): 1–9. doi: https://doi.org/10.1038/s41541-017-0006-8.

Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. 2014. Exposing synonymous mutations. *Trends in Genetics*. vol 30(7): 308–321. doi: https://doi.org/10.1016/j.tig.2014.04.006.

Li S, Mo X, Wei M, Pan H, Zhang J, Xia N. 2016. Truncated L1 protein of human papillomavirus type 52. (US Patent No. 9,499,591). U.S. Patent and Trademark Office. https://patents.google.com/patent/US9499591B2/en#patentCitations.

Luo Y, Enghiad B, Zhao H. 2016. New tools for reconstruction and heterologous expression of natural product biosynthetic gene clusters. *Natural Product Reports*. vol 33(2): 174–182. doi: https://doi.org/10.1039/C5NP00085H.

Marchisio MA. 2014. Parts & pools: a framework for modular design of synthetic gene circuits. *Frontiers in Bioengineering and Biotechnology*. vol 2: 1–10. doi: https://doi.org/10.3389/fbioe.2014.00042.

Mauro VP, Chappell SA. 2014. A critical analysis of codon optimization in human therapeutics. *Trends in Molecular Medicine*. vol 20(11): 604–613. doi: https://doi.org/10.1016/j.molmed.2014.09.003.

Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research*. vol 28(1): 292. doi: https://doi.org/10.1093/nar/28.1.292.

Puigbò P, Bravo IG, Garcia-Vallve S. 2008. CAIcal: a combined set of tools to assess codon usage adaptation. *Biology Direct*. vol 3: 1–8. doi: https://doi.org/10.1186/1745-6150-3-38.

Plummer EM, Manchester M. 2011. Viral nanoparticles and virus-like particles: platforms for contemporary vaccine design. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*. vol 3(2): 174–196. doi: https://doi.org/10.1002/wnan.119.

Quan J, Saaem I, Tang N, Ma S, Negre N, Gong H, White KP, Tian J. 2011. Parallel on-chip gene synthesis and application to optimization of protein expression. *Nature Biotechnology*. vol 29(5): 449–452. doi: https://doi.org/10.1038/nbt.1847.

Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. vol 11(1): 1–9. doi: https://doi.org/10.1186/1471-2105-11-129.

Rosano GL, Ceccarelli EA. 2014. Recombinant protein expression in *Escherichia coli*: advances and challenges. *Frontiers in Microbiology*. vol 5: 1–17. doi: https://doi.org/10.3389/fmicb.2014.00172.

Saunders R, Deane CM. 2010. Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Research*. vol 38(19): 6719–6728. doi: https://doi.org/10.1093/nar/gkq495.

Schiller JT, Day PM, Kines RC. 2010. Current understanding of the mechanism of HPV infection. *Gynecologic Oncology*. vol 118(1): 12–17. doi: https://doi.org/10.1016/j.ygyno.2010.04.004.

Sievers F, Higgins DG. 2014. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Molecular Biology*. vol 1079: 105–116. doi: https://doi.org/10.1007/978-1-62703-646-7_6.

Tumban E, Peabody J, Tyler M, Peabody DS, Chackerian B. 2012. VLPs displaying a single L2 epitope induce broadly cross-neutralizing antibodies against human papillomavirus. *PloS One*. vol 7(11): 1–11. doi: https://doi.org/10.1371/journal.pone.0049751.

Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S. 2016. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics*. vol 7: 1–8. doi: https://doi.org/10.1186/1471-2105-7-285.

Wang JR, Li YY, Liu DN, Liu JS, Li P, Chen LZ, Xu SD. 2015. Codon optimization significantly improves the expression level of α-amylase gene from *Bacillus licheniformis* in *Pichia pastoris*. *BioMed Research International*. vol 2015: 1–9. doi: https://doi.org/10.1155/2015/248680.

Wei M, Wang D, Li Z, Song S, Kong X, Mo X, Yang Y, He M, Li Z, Huang B, Lin Z, Pan H, Zheng Q, Yu H, Gu Y, Zhan J, Li S, Xia N. 2018. N-terminal truncations on L1 proteins of human papillomaviruses promote their soluble expression in *Escherichia coli* and self-assembly in vitro. *Emerging Microbes and Infection*. vol 7(1): 1–12. doi: https://doi.org/10.1038/s41426-018-0158-2.

Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, Gustafsson C. 2009. Design parameters to control synthetic gene expression in *Escherichia coli*. *PloS One*. vol 4(9): 1–10. doi: https://doi.org/10.1371/journal.pone.0007002.

Yang X, Cheng Y, Li C. 2017. The role of TLRs in cervical cancer with HPV infection: a review. *Signal Transduction and Targeted Therapy*. vol 2(1): 1–10. doi: https://doi.org/10.1038/sigtrans.2017.55.