Journal of Bio-Molecule Research and Engineering
Vol. 1 | No. 1 | 1 – 6 | 2022
https://www.e-journal.unair.ac.id/biome

In Silico Design Gene Encoding CYP71AV1 for Expression in *Escherichia coli*

Evi Umayah Ulfa*
Department of Pharmaceutical Biology, Faculty of Pharmacy, Universitas Jember,
Jl Kalimantan I/2, Jember 68121, Indonesia
*Corresponding author: evi.farmasi@unej.ac.id

ARTICLE INFO
Article history
Received 09th Dec 2021
Accepted 12th May 2022
Keywords:
in silico design
CYP71AV1
CAI
GC contents
protein solubility

ABSTRACT
Cytochrome P450 monooxygenase (CYP71AV1) is a crucial enzyme in the artemisinin biosynthesis pathway. This enzyme oxidized Amorpha-4,11-diene to produce artemisinic acid. This study aimed to in silico design high-level expression of CYP71AV1 in the *Escherichia coli* system. In silico techniques are highly suitable for designing protein recombinant production before entering the laboratory. The amino acid sequence of CYP71AV1 was back-translated to the DNA sequence and adapt to *E. coli* codon usage by using Gene Designer. The DNA sequence of optimized CYP71AV1 was analyzed using Rare Codon Analysis to assess the expression efficiency in *E. coli*. The protein solubility prediction was determined using the SoDoPe tool. The optimized CYP71AV1 was determined to have a CAI 0.81, a GC content of 53.08 %, CFD with low frequency, and no negative cis or repeat elements. The result of the probability of solubility of CYP71AV1 was 0.6207 when expressed in *E. coli*. The MBP fusion partner can be used to increase the solubility of CYP71AV1. The in-silico results showed the possibility of high-level protein expression of optimized CYP71AV1 in the *E. coli* system.

1. Introduction

Malaria is a global health problem, especially in a tropical area like Africa and Southeast Asia. World Health Organization (WHO) in 2005 reported that 300 million to 500 million people were infected by *Plasmodium* sp, and more than one million were killed annually. *Plasmodium falciparum* as a malarial agent reported becoming resistant to nearly all antimalarial drugs [[1]]. Multidrug resistance in *P. falciparum* to the commonly used antimalarial agents becomes more widespread.

Artemisinin is a sesquiterpenoid lactone isolated from *Artemisia annua* L. It is known as an alternative antimalarial compound. Artemisinin is effective against both chloroquine-resistant and sensitive strains of *Plasmodium* sp. In 2006, WHO recommended malarial treatment with ACTs (Artemisinin-based combination therapies/ACTs) [[2]]. Artemisinin production in plants followed the terpenoid biosynthesis pathway through the mevalonate pathway. Some enzyme involved in the artemisinin biosynthesis are farnesyl pyrophosphate synthase (ERG20), amorpha-4,11-diene synthase (ADS), cytochrome P450 monooxygenase (CYP71AV1), and cytochrome P450 reductase (CPR) [[3]]. CYP71AV1 oxidized Amorpha 4,11 diene to produce artemisinic acid.

To improve the understanding of artemisinin biosynthesis, molecular cloning of the enzymes involved in the oxidation of amorpha-4,11-diene in heterologous hosts is needed. *Escherichia coli* is the most favorable host for producing recombinant proteins due to its inexpensive, rapid growth rate, ease of manipulation, and high yield of protein expression [[4]].

The expression of a heterologous protein requires a codon between the host and the target gene. Differences in the use of codons can result in lower expression of heterologous proteins due to the inhibition of the translation process [[5],[6]]. Prediction of the success of heterologous protein expression can be made in silico using rare codon analysis software. Codon changes to have a high preference with codons in *E. coli* can be done using the Codon optimization tools software. This study aimed to in silico design the gene encoding CYP71AV1 for expression in *E. coli*. In silico designs were carried out to reduce the risk of failure, saving time, and reduce research costs [[7]].

2. Materials and methods

The bioinformatics tools used in this research include SignalP3.0 software, Gene Designer (DNA 2.0), Rare Codon Analysis (www.genscript.com), and protein solubility prediction (https://tisigner.com/sodope/). The CYP71AV1 amino acid sequences were taken from NCBI (access code: ABI31728.1).

2.1. Prediction of signal peptides

The CYP71AV1 amino acid sequence along with 495 amino acids was obtained from the NCBI gene bank (access code: ABI31728.1). The signal peptide for CYP71AV1 was predicted using SignalP3.0. The method in SignalP3.0 is based on the Hidden Markov Model (HMM) and Neural Network (NN) methods [[8]].

2.2. Design and optimization of the gene encoding of CYPAV171

Codon optimization of the gene encoding of CYP71AV1 was carried out using Gene Designer (DNA 2.0). The amino acid sequence of CYP71AV1 was back-translated according to the reference codon in *E. coli*. The adjusted parameters were codon usage bias, GC content, and enzyme restriction side that might interfere with the cloning. The nucleotide sequence of non-optimized CYP71AV1 and optimized CYP71AV1 were analyzed using Rare Codon Analysis (https://www.genscript.com/cgi-bin/tools/rare_codon_analysis) to assess the expression efficiency of CYP71AV1 in *E. coli*. The parameters assessed were codon adaptation index (CAI), GC content, and Codon Frequency Distribution (CFD).

2.3. Protein solubility prediction

The solubility of CYP71AV1 protein was performed by the SoDoPe tool (https://tisigner.com/sodope).

3. Results and discussion

3.1. Signal peptide prediction

CYP71AV1 was an extracellular protein that accumulated in the trichome. Protein sequence encoding of CYP71AV1 was obtained from NCBI GenBank (accession number ABI31728.1) from *A. annua*. To obtain the mature protein, the signal peptide of CYP71AV1 should be removed. The signal peptide of CYP71AV1 was confirmed by signal P3.0 software using the Hidden Markov Model and Neural Network methods.

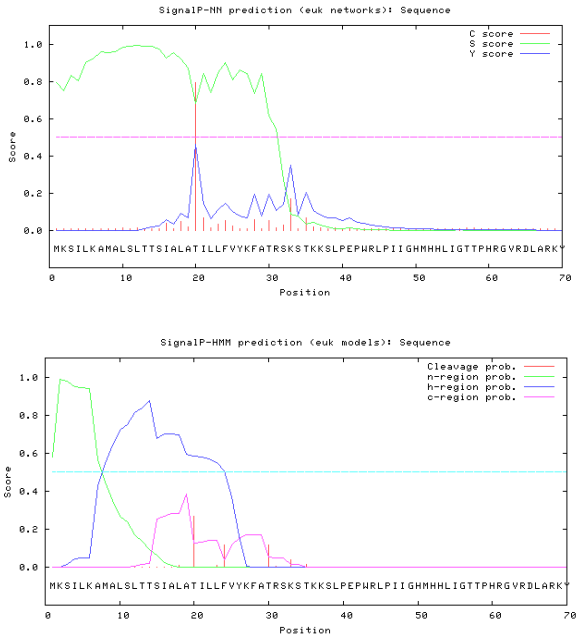

Figure 1. The results of signal peptide prediction analysis using SignalP 3.0. (A) Hidden Markov Model Method. (B). Neural Network Method.

Signal peptide confirmation from both methods showed that the first 19 amino acids from the N terminal of CYP71AV1 protein were signal peptide. Cleavage sites occur between A19 and T20 (Figure 1). Based on Figure 2, the mature protein of CYP71AV1 is composed of 476 amino acids starting with Threonine (T20) to Phenylalanine (F495).

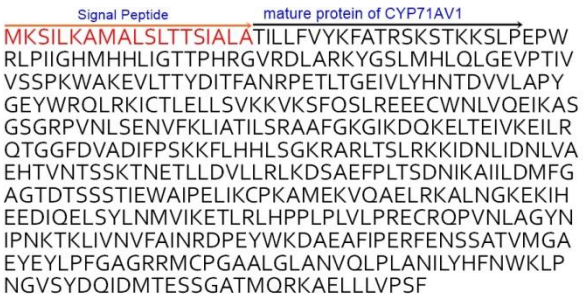

Figure 2. The CYP71AV1 amino acid sequence. Signal peptides (red color), mature protein of CYP71AV1 (black).

Signal peptides are peptides with a length of 16–30 amino acids present at the N terminal of nascent proteins. Signal peptides or leader sequences are co-translationally-translocated nascent proteins into the Endoplasmic Reticulum (ER) lumen. Secreted proteins are delivered to the cell surface via vesicular carriers, budding from TGN or AG directly [[9]].

3.2. Design and optimization of the gene encoding of CYPAV171

The efficiency of heterologous protein expression may be affected by codon sequence, mRNA stability, host strain, and environment [[10],[11]]. The difference in codon preference used in the original host and target host can reduce protein expression efficiency due to low mRNA translation and the cessation of the polypeptide elongation process during translation [[12],[13]]. Codon

optimization is only performed on mature proteins. The amino acids that make up the signal peptide were removed after in silico analysis. The signal peptide was not codon optimized because the signal peptide was not expressed. CYP71AV1 expression is targeted to the cytoplasm.

According to the rare codon analysis of non-optimized CYP71AV1, the CAI (codon adaptation index) value was calculated to be 0.58 outside the ideal value for expression in *E. coli*. The GC content is 43.06%, and the optimal codon frequency (CFD) value is 30% which is also outside the ideal value (**Table 1**). This result showed that non-optimized CYP71AV1 contains a rare codon for *E. coli*, thus, ineffective translation might occur [[14],[6]].

Table 1. Rare codon analysis result

Sequence	CAI	GC	CFD
Non-optimized CYP71AV1	0.58	43.06%	16%
Optimized CYP71AV1	0.81	53.08%	2%

The strategy to increase the expression of a heterologous protein is by replacing rarely used codons with high-frequency codons in the desired host [[15]]. The sequence of optimized codons was then chemically synthesized for the cloning process. This method is known to be much more effective than gene isolation. It can also reduce the risk of pathogenic microbial infection [[16]]. In this research, the optimized codon of CYP71AV1 was done by an online tool from DNA 2.0. The sequence of nucleotides from codon optimization can be seen in **Figure 3**.

Optimized CYP71AV1 →

```
ATGACCATCCTGCTGTTCGTCTACAAATTCGCAACTCGCTCAAAGAGCACCAAGAAGAGCCTGCCGGAG
CCGTGGCGCCTGCCGATCATTGGCCACATGCATCACCTGATTGGTACCACCCCGCACCGTGGTGTGCGT
GACCTGGCGCGTAAGTACGGTTCTCTGATGCACCTGCAGCTGGGCGAGGTCCCAACCATCGTGGTCAG
CAGCCCGAAATGGGCAAAGGAAGTTCTGACGACGTACGACATCACCTTTGCTAATCGTCCGGAAACCTT
GACCGGCGAGATCGTATTGTATCATAATACCGATGTTGTCCTGGCCCCGTACGGCGAATACTGGCGTCA
ACTGCGTAAGATTTGTACCCTGGAACTGCTGTCCGTGAAAAAGGTGAAATCTTTTCAAAGCCTGCGCGA
AGAAGAGTGCTGGAATCTGGTGCAGGAGATTAAGGCGAGCGGTAGCGGTCGCCCGGTTAATCTGTCCG
AAAATGTGTTCAAACTGATTGCGACTATCCTGTCCCGTGCGGCGTTCGGTAAAGGCATTAAAGACCAAA
AAGAACTGACGGAGATTGTCAAAGAGATCCTGCGCCAAACGGGCGGTTTTGATGTTGCGGACATCTTC
CCTTCTAAAAAGTTCCTGCATCATCTGAGCGGCAAACGTGCCCGTCTGACCAGCCTGCGCAAGAAAATT
GACAACTTGATCGATAACCTGGTTGCAGAGCATACCGTTAATACGAGCTCGAAAACGAACGAAACCTTG
CTGGACGTGCTGTTGCGCTTGAAGGATAGCGCGGAGTTCCCGCTGACTAGCGACAACATCAAGGCGAT
TATTCTGGACATGTTTGGTGCGGGTACCGATACGAGCAGCAGCACCATTGAGTGGGCCATTCCGGAGCT
GATCAAATGCCCGAAAGCAATGGAAAAAGTTCAGGCTGAGCTGCGTAAGGCGCTGAACGGTAAGGAA
AAGATTCACGAAGAGGATATTCAGGAACTGAGCTACCTGAATATGGTTATCAAGGAAACGTTGCGTCTG
CACCCTCCGTTGCCGCTGGTTCTGCCGCGTGAGTGTCGTCAGCCGGTGAACTTGGCCGGCTACAATATC
CCGAACAAGACCAAACTGATCGTCAACGTATTTGCAATCAATCGCGATCCGGAGTATTGGAAAGACGCT
GAGGCTTTTATCCCAGAGCGTTTTGAAAATTCCAGCGCAACGGTGATGGGCGCAGAGTATGAGTATCTG
CCGTTTGGTGCCGGTCGTCGTATGTGCCCAGGTGCGGCCCTGGGTCTGGCGAACGTCCAGCTGCCGTT
GGCGAACATTCTGTATCACTTCAACTGGAAGCTGCCGAACGGCGTCAGCTACGATCAGATTGACATGAC
CGAAAGCAGCGGTGCAACGATGCAACGCAAAGCGGAGCTGCTGCTGGTTCCGAGCTTCTAA
```

Figure 3. Nucleotide sequence of optimized CYP71AV1

The optimized CYP71AV1 was determined to have a CAI of 0.81, a GC content of 53.08%, CFD with low frequency and no negative cis, or repeat elements using rare codon analysis. The codon optimization tool has improved the CAI value of optimized CYP71AV1 from 0.58 to 0.8. The CAI value > 0.8 indicates good genes expressed in the desired host. The average value of the GC content of the optimized CYP71AV1 was under the excellent value, namely 53.08%, and there was no peak outside of 30%–70%. High GC content values (> 70%) can reduce efficiency and inhibit translation, while low GC content values (<30%) can result in a delayed transcription elongation process [[6]]. All parameters of optimized CYP71AV1 are in ideal range compared with non-optimized CYP71AV1.

Research on increasing the level of heterologous protein expression in *E. coli* through codon optimization has been carried out, including interleukin-2, which increased 16 times. Interleukin 18

increased five times; Troponin T increased 40 times, and glutathione transferase increased 140 times [[6]].

3.3. Protein Solubility Prediction

Protein solubility of CYP71AV1 was predicted by Soluble Domain Protein (SoDoPe) tools. The amino acid sequence of CYP71AV1 has probability of solubility of 0.6207 when expressed in *E. coli* (Table 2). SoDoPe online tool calculates the protein solubility based on Solubility-Weighted Index (SWI) [[17]]. Solubility score above 0.5 indicates CYP71AV1 is a soluble expression in *E. coli*.

Table 2. SoDoPe Analysis Results

Protein	Probability of solubility	Fexibility	GRAVY
CYP71AV1	0.6207	1.001	-0.2019

Obtaining recombinant protein, which is expressed at a high level and soluble, is the main requirement of recombinant protein production [[18]]. However, almost half of the heterologous proteins fail to be expressed and half of them are successfully expressed in inclusion bodies (http://targetdb.rcsb.org/metrics/). Solubility of protein is influenced by physicochemical characteristics and protein structure such as molecular weight, hydrophobicity, aromaticity, isoelectric point, and the polarity of residue as well as by extrinsic factors including ionic strength, pH, and temperature [[19]-[22]].

SoDoPE online tool also provides options for solubility prediction in the presence of fusion partner proteins. Fusion partner proteins such as thioredoxin (TRX), maltose-binding protein (MBP), small ubiquitin-related modifier (SUMO), and glutathione S-transferase (GST) can increase protein solubility [[18]]. Prediction of the effect of different fusion partner proteins on the solubility of CYP71AV1 can be seen in **Figure 4**.

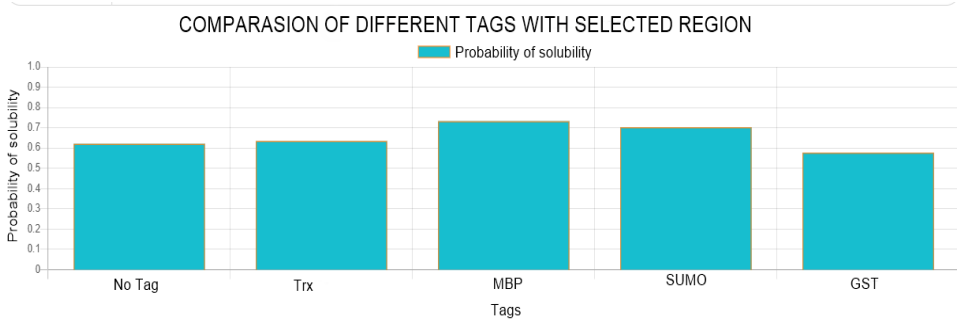

Figure 4. The effect of different tags on solubility of CYP71AV1

Based on **Figure 4**, the fusion partner protein that is likely to provide the highest increase in CYP71AV1 solubility was MBP. MBP is known to be one of the most effective solubilizing enhancers for increasing solubility, improving yield, and aiding the proper folding of its fusion partners. Fusion partner MBP can be placed at the C-terminus (MBP-CYP71AV1), after the CYP71AV1, or at the N-terminus (CYP71AV1-MBP), before the CYP71AV1. Both fusion proteins have the same probability of solubility of 0.736. These results indicate that SoDoPe tool cannot distinguish the effect of the position of the fusion protein. In fact, several publications have shown that MBP at the N terminus of the target protein is more efficient than MBP at the C-terminus [[24]]. Currently, various commercial

expression vectors containing MBP tags are available for production in *E. coli*, including the pMAL (New England Biolabs), MBP tag vector (Sigma) and pIVEX (Roche) series. The exact mechanism of MBP to increase solubility is unknown, but it is thought that MBP serves primarily as a "holdase" that will mediate proper folding [[23],[24]].

Although testing in-silico predictions were good, they could not be ascertained when tested in vitro and in vivo, which is directly proportional to the results of in-silico due to various factors that can influence the final results of research that is expected.

4. Conclusions

The codon optimization results indicate the optimized CYP71AV1 had a high probability of being expressed in *E. coli* compared to the non-optimized CYP71AV1. In silico design can be considered as a suitable method to predict efficient protein expression before entering the laboratory. However, experimental studies are required to verify this construct.

References

[1] E.L. Korenromp, B.G. Williams, E. Gouws, C. Dye and R.W. Snow, Lancet Infect. Dis 3, 349–358, 2003.
[2] D. Rathore, T.F. Mc Cutchan, M. Sullivan, and S. Kumar, Expert Opin. Investig. Drugs 14, 871–883, 2005.
[3] V.J.J. Martin, D.J. Pitera, S.T. Withers, J.D. Newman, J.D. Keasling J.D, Nature Biotechnology, Nature Publishing Group 21, 796-802, 2003.
[4] L.D. Cabrita, D. Weiwen and P.B. Stephen. BMC biotechnology 6, 12, 2006.
[5] H.P. Sorensen, K.K. Mortensen, Journal of Biotechnology 115,113–128, 2005.
[6] C. Gustafsson, S. Govindarajan, J. Minshull. TRENDS in Biotechnology.22 (7), 346-353, 2004.
[7] M.A. Ullah, B. Sarkar, and S.S. Islam. Immunobiology, 225 (3), 1–80, 2020.
[8] H. Nielsen, J. Engelbrecht, S. Brunak, G.V. Heijne, International Journal of Neural System 8,581–599, 1997.
[9] X. Wang, X. Li, Z. Zhang, X. Shen and F. Zhong, Protein Expression and Purification 72, 101-106, 2010.
[10] E. U. Ulfa, E. Munadziroh, H. Hermansyah, N.N.T. Puspaningsih, J. Chem. Technol. Metall. 55 (6), 1999-2008, 2020.
[11] D. Kernain, M.A. Samad, S. Shamsuddin, Internal Medicine Journal 24,451–454, 2017.
[12] G. Hanson, J. Coller.Nat Rev Mol Cell Biol.19(1), 20-30, 2018.
[13] A. Evelina Biotechnol J. 6 (6), 650–659, 2011.
[14] A. Ghasemi, R. Ranjbar, J. Amani. Iran J Basic Med Sci 17 (3), 172-180, 2014.
[15] B. Zheng, X. Ma, N. Wang, et al. Nat Commun 9, 3616, 2018.
[16] C. Elena, P. Ravasi, ME. Castelli, S. Peirú, H.G. Menzella, Front Microbiol 5:21, 2014.
[17] B. K. Bhandari, P.P. Gardner, C.S. Bioinformatics. 36(18):4691-4698, 2020.
[18] D. Esposito, D.K. Chatterjee Curr. Opin. Biotechnol 17, 353–358, 2006.
[19] F. Chiti. M. Stefani,N. Taddei, G. Ramponi, and C.M. Dobson, Nature, 424, 805–808, 2003.
[20] A.A.Diaz, E. Tomba, R. Lennarson, R. Richard, M.J. Bagajewicz, R.G. Harrison. Biotechnol. Bioeng.105, 374–383, 2010
[21] G. G. Tartaglia, A.Cavalli, R. Pellarin, A. Caflisc. Protein Sci. 13, 1939–1941, 2004.
[22] D. L. Wilkinson, R.G Harrison Biotechnology, 9, 443–448, 1991.
[23] S. Raran-Kurussi and D.S. Waugh. PLoS One 17, 2012.
[24] S. Raran-Kurussi, K. Keefe, D.S. Waugh. Protein Expr Purif. 110, 159-164, 2015.