ENHANCING RECOMBINANT PROTEIN PRODUCTION BY mRNA STRUCTURE OPTIMIZATION: PNEUMOLYSIN AS A CASE STUDY

Filipe Fusco[1,2], Manuella C. Pires[1,2], Alexandre P. Y. Lopes[2], Vítor dos S. Alves[1,2] & Viviane M. Gonçalves[2]*

[1] *Interunits Graduate Program in Biotechnology, University of São Paulo, São Paulo, Brazil.*
[2] *Butantan Institute, São Paulo, Brazil.*
** Corresponding author's email address: viviane.goncalves@butantan.gov.br*

ABSTRACT

Recombinant proteins are of great importance in modern society, mostly as biopharmaceutical products. However, challenging and complex processes with low production yield are major drawbacks. Normally, the optimization to overcome these obstacles is focused on bioreactor and purification processes, and the biomolecular aspects are left behind, seen as less important. Here, we present how 5' mRNA secondary structure region can be relevant for translation and, therefore, protein production. For this, *Escherichia coli* BL21(DE3) clones, producing recombinant detoxified pneumolysin (PdT) with and without N-terminal His-tag, were cultivated in 10 L bioreactors. Target proteins were quantified and *in silico* mRNA analyses were performed using TIsigner and RNAfold. The results showed that the His-tag presence at N-terminus generated a >1.5-fold increase in target protein synthesis, which was explained by the *in silico* mRNA analyses that returned a mRNA secondary structure easier to translate and, therefore, higher protein production than without His-tag. A second version of *pdt* gene with synonymous changes in the 5' end was also analyzed *in silico* and experimentally to clarify the causes of the phenomena and confirmed the hypothesis raised. This work reveals that simple mRNA analyses during heterologous gene design can help to reach high recombinant protein titers.

Keywords: Recombinant *Escherichia coli*. *Streptococcus pneumoniae*. PdT. Translation initiation. Bioreactor.

1 INTRODUCTION

The global market related to recombinant proteins spent USD 1.74 billion only in 2021, with a projected increase in the following years[1]. Therefore, the production and purification process of recombinant proteins, and their optimization, are always of great interest for producers and society. One of the specific challenges faced during protein synthesis is the level of mRNA translation. The mRNA can present different translational rates depending on the genetic sequence and secondary structure[2], consequently influencing the final amount of product recovered. For recombinant genes, the initial portion of the mRNA at the 5'-end can be modified, and the translational rate can be evaluated[3] because the site of the initiation of translation will dictate the ribosome capacity of interaction with the mRNA[4]. This site also determines the energy required for translation initiation, since the mRNA folding should be disrupted for the appropriate interaction with ribosomes[5, 6]. As a result, the insertion of N-terminal tags to promote protein solubility or facilitate protein purification modifies the 5'-end of the recombinant gene sequence and affect the translation initiation[7]. Pneumolysin is a pore-forming toxin secreted by the bacterium *Streptococcus pneumoniae*. This microorganism is leading cause of ill health and death worldwide, responsible for pneumococcal diseases such as pneumonia, meningitis, and sepsis[8]. Pneumolysin and its derivatives are targets for the development of new vaccines[9]. They have long been produced and purified from *S. pneumoniae* and as recombinant proteins from *Escherichia coli*. Genetically detoxified pneumolysin variants as PdT are candidates for new serotype-independent pneumococcal vaccines; thus, it is essential to develop scalable and robust processes that allow high titers of protein production to make it viable for future vaccine production. This work evaluated the production of PdT in *E. coli* BL21(DE3). Three different genetic versions were tested, the protein products were obtained in a 10-L bioreactor, and their production levels were evaluated. We also evaluated the effect of the 5' region of mRNA on protein synthesis by *in silico* analysis focused on the mRNA structure and its impact on heterologous gene expression.

2 MATERIAL & METHODS

The codons of pneumolysin gene (Gene ID: 66806991) with the three mutations, C42G, W433F, and D385N described by Berry and colleagues[10] were optimized for *E. coli* codon usage (*pdt* gene version 1). This sequence was also used for insertion of codons of the N-terminal His-tag and the sequence of Tobacco Etch Virus (TEV) protease cleavage site (*his-tev-pdt*). Each gene was transformed in *E. coli* BL21(DE3) and the proteins were separately produced in a 10-L bioreactor using autoinduction medium. For more details see reference[11]. Each protein was produced in two different conditions named C1 and C2. At C1 the temperature was changed from 37 °C to 25 °C after 4 h of exponential growth. At C2 the temperature was changed from 37 °C to 25 °C after glucose exhaustion. In addition, the *pdt* gene without any modification (*pdt* gene version 2) was also produced the same way under C1. All the genetic sequences were also submitted to TIsigner (https://tisigner.com/)[12] for determination of mRNA opening energy (OE, i. e. the energy required for the ribosome to open the mRNA and proceed with the translation process) and expression score (ES), to the RNAfold, ViennaRNA Package version 2.5.1 (http://rna.tbi.univie.ac.at/cgibin/RNAWebSuite/RNAfold.cgi)[13], for determination of minimum free energy (MFE) and mRNA MFE secondary structure, and to ProtParam tool at Expasy (https://web.expasy.org/protparam/)[14] for estimation of half-life.

3 RESULTS & DISCUSSION

The production of His-TEV-PdT was higher than PdT regardless the condition (Fig. 1-A and 1-B). In addition, His-TEV-PdT was also produced in insoluble fraction, which is believed to be related to strong overexpression[15]. A possible explanation for this difference in production can be found on Fig. 1-C and 1-D, that show a mRNA secondary structure with a shorter stem structure for *his-tev-pdt* gene, which makes it easier to be translated than the *pdt* gene version 1. To clarify this hypothesis, PdT from *pdt* gene version 2, (i. e. with a different initial region) was produced under C1 and analyzed by the same *in silico* softwares. A summary of the most important results comparing the three genetic sequences are shown in Table 1.

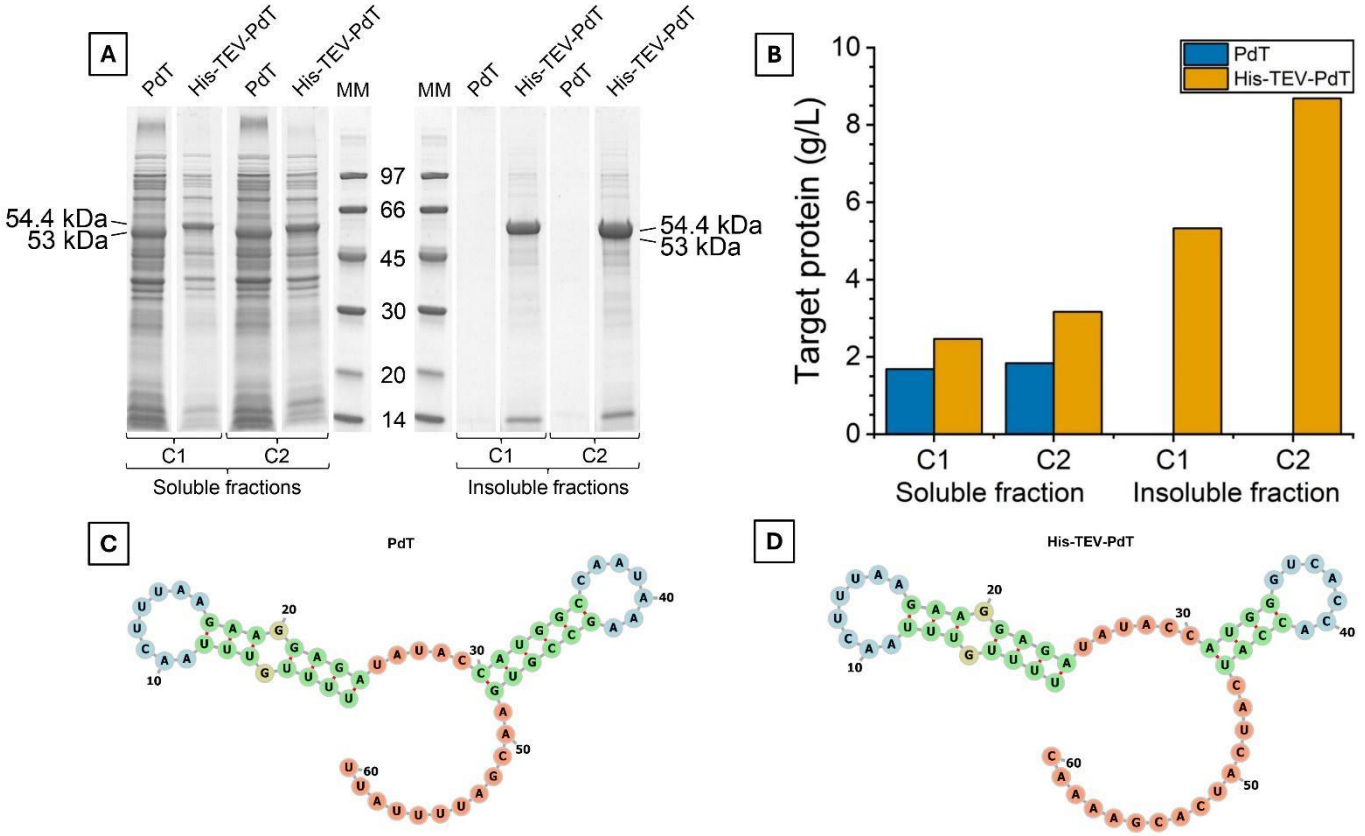

Figure 1 - A: SDS-PAGE to compare PdT (from gene version 1) and His–TEV–PdT production in the biomass processed in parallel. MM – molecular marker. B: Estimated concentration of PdT and His–TEV–PdT. C and D: −30:30 region of the mRNA MFE secondary structure predicted by RNAfold for the translation of *pdt* (version 1) and *his-tev-pdt* genes. The numbers indicate the ribonucleotide position from the beginning of the sequence used as input. Start codons are located from 31 to 33, considering the figure numbers. The colors represent the different structures observed. Green - stems, yellow - interior loops, blue - hairpin loops, and orange - 5′ and 3′ unpaired regions. Source: Fusco et al. 2024[11].

Table 1 - Summary of *in silico* analysis and production results under C1.

Gene	Soluble protein (g/L)	Open Enrgy (kcal/mol)	Expression Score	Minimum Free Energy (kcal/mol)	Half-life (h)
pdt gene version 1	1.7	14.19	33.60	-9.30	>10
pdt gene version 2	2.0	9.80	83.27	-5.20	>10
his-tev-pdt	2.5	8.70	90.72	-4.30	>10

There is no difference on half-life of the proteins (Table 1), so the difference in protein production cannot be due to differences in protein stability. Additionally, the higher the expression score, the higher is the protein production. The reverse logic is valid for MFE, where the lower the MFE, the lower is the protein production, which makes sense since a more stable structure is more difficult to unfold during translation. Finally, the higher the mRNA opening energy, the lower is the protein production, since it means a secondary structure more resistant to open. These results show that there is a correlation between mRNA secondary structure and protein production. According to the literature, the protein production is inversely proportional to mRNA opening energy[16]. The OE of *pdt* gene version 1 is 1.6-fold higher than the *his-tev-pdt* (Table 1) and the production of His-TEV-PdT was 1.5-fold higher than PdT (gene version 1) production under C1. In addition, Tisigner showed a 1.4-fold higher OE for the *pdt* gene version 1 compared to the *pdt* version 2, which is in accordance with the experimental results as the PdT production was 1.2-fold greater with the gene version 2 than with the gene version 1. Comparing the 5' mRNA region of the *pdt* version 2 to the *his-tev-pdt*, a 1.1-fold higher OE was calculated. This result is also in accordance with the experimental data, which returned a 1.2-fold increase in target protein production for His-TEV-PdT when compared to PdT obtained with the gene version 2. According to the literature, changes in the first nine codons can achieve nearly optimum accessibility when compared to full-length modifications[16]. Therefore, codon differences in regions other than the 5'-end of the *pdt* gene version 2 are not likely to promote a significant reduction in opening energy and, consequently, should not increase protein production. It is also interesting to highlight that in this work, codon usage optimization worsened target protein production, showing that the mRNA initial region may play a much more important role in protein synthesis than generally supposed.

In this study, the difference between PdT (gene version 1) and His-TEV-PdT was only the presence of the His-tag and TEV cleavage site sequences. As it can be seen in Figure 1-C and 1-D, the ribonucleotides related to the His-tag sequence seemed to have major impact on this result, since only the two last codons of TEV cleavage site were considered in RNAfold analysis. The region from -30 to +30 was considered because this region was employed to calculate MFE in previous studies[16]. The prediction of the entire mRNA structure would generate a different structure that would be much complex to analyze, and we do not know how much time the complete mRNA remains stable until degradation starts. We also do not know how relevant is this time interval for production in comparison with the translation that occurs in parallel with transcription. Besides, other studies have reported that the presence of N-terminal His-tag improved the target protein synthesis, confirming that the mRNA structures that contain the His-tag codons at 5' end region can be easier and faster translated than without these additional codons, increasing the amount of target protein obtained[17, 18, 19, 20]. Nonetheless, it is important to highlight that the phenomenon reported here was not provoked by the inclusion of the N-terminal His-tag itself. If the mRNA secondary structure without this tag has higher MFE and lower opening energy, as consequence, a more favorable translational structure resulting in higher expression score than the counterpart with His-tag, the His-tag insertion would worsen the target protein production.

4 CONCLUSION

The literature shows that TIsigner presented about 70% accuracy when predicting successes or failures of 11,430 expression experiments from a data bank and allowed 4 times higher production of green fluorescent protein (GFP) and 1.5 times higher production of luciferase after 5' end region optimization[16]. This tool also correctly predicted our experimental production differences, which strengths that it should be widely used for optimization of the recombinant protein production to save money and time in research and industry. We showed that instead of a classical process optimization, that would take much longer time, the titer can be increased optimizing the mRNA fold by changing the codons of the first amino acids. In comparison, the application of TIsigner software could provide an optimized protein production in much shorter time, possibly achieving even better results than using the traditional process optimization strategies. Finally, it is important to highlight that this strategy is rarely applied neither to explain the results nor to improve production, and the case study presented here can contribute to disseminate this knowledge.

REFERENCES

1 Polaris, Market Research. Recombinant Proteins Market Share, Size, Trends, Industry Analysis Report, By Product & Services (Product, Production Services); By Application; By End-Use; By Region; Segment Forecast, 2022 – 2030. 2022.

2 Hoernes, T. P., Hüttenhofer, A., Erlacher, M. D. (2016). mRNA modifications: Dynamic regulators of gene expression?. RNA Biol. 13(9). 760–765.

3 Ma, J., Campbell, A., Karlin, S. (2002). Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J. Bacteriol. 184(20). 5733–5745.

4 Vellanoweth, R. L., Rabinowitz, J. C. (1992). The influence of ribosome-binding-site elements on translational efficiency in Bacillus subtilis and *Escherichia coli* in vivo. Mol. Microbiol. 6(9). 1105–1114.

5 Wolfsheimer, S., & Hartmann, A. K. (2010). Minimum-free-energy distribution of RNA secondary structures: Entropic and thermodynamic properties of rare events. Physical review. E, Statistical, nonlinear, and soft matter physics. 82(2 Pt 1). 021902.

6 Achar, A., Sætrom, P. (2015). RNA motif discovery: a computational overview. Biol. Direct. 10. 61.

7 Malhotra A. (2009). Tagging for protein expression. Methods Enzymol. 463, 239–258.

8 Weiser, J. N., Ferreira, D. M., Paton, J. C. (2018). *Streptococcus pneumoniae*: transmission, colonization and invasion. Nat. Rev. Microbiol. 16(6). 355–367.

9 Pichichero M. E. (2017). Pneumococcal whole-cell and protein-based vaccines: changing the paradigm. Expert Rev. Vaccines. 16(12). 1181–1190.

10 Berry, A. M., Alexander, J. E., Mitchell, T. J., Andrew, P. W., Hansman, D., Paton, J. C. (1995). Effect of defined point mutations in the pneumolysin gene on the virulence of *Streptococcus pneumoniae*. Infect. Immun. 63(5). 1969–1974.

11 Fusco, F., Pires, M. C., Lopes, A. P. Y., Alves, V. d. S., Gonçalves, V. M. (2024). Influence of the mRNA initial region on protein production: a case study using recombinant detoxified pneumolysin as a model. Front. Bioeng. Biotechnol. 11.

12 Bhandari, B. K., Lim, C. S., and Gardner, P. P. (2021). TISIGNER.com: web services for improving recombinant protein production. Nucleic Acids Res. 49. W654–W66.

13 Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., and Hofacker, I. L. (2008). The Vienna RNA websuite. Nucleic Acids Res. 36.

14 Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005). "Protein identification and analysis tools on the ExPASy Server," in The proteomics protocols handbook, ed. J. M. Walker (Springer Protocols Handbooks. Humana Press). 571–607.

15 Beygmoradi, A., Homaei, A., Hemmati, R., and Fernandes, P. (2023). Recombinant protein expression: challenges in production and folding related matters. Int. J. Biol. Macromol. 233.

16 Bhandari, B. K., Lim, C. S., Remus, D. M., Chen, A., van Dolleweerd, C., and Gardner, P. P. (2021). Analysis of 11,430 recombinant protein production experiments reveals that protein yield is tunable by synonymous codon changes of translation initiation sites. PLoS Comput. Biol. 17.

17 Cèbe, R., and Geiser, M. (2006). Rapid and easy thermodynamic optimization of the 5′-end of mRNA dramatically increases the level of wild type protein expression in *Escherichia coli*. Protein Expr. Purif. 45. 374–380.

18 Doray, B., Chen, C. Di, and Kemper, B. (2001). N-terminal deletions and His-tag fusions dramatically affect expression of cytochrome P450 2C2 in bacteria. Arch. Biochem. Biophys. 393. 143–153.

19 Park, W.-J., You, S.-H., Choi, H.-A., Chu, Y.-J., and Kim, G.-J. (2015). Over-expression of recombinant proteins with N-terminal His-tag via subcellular uneven distribution in *Escherichia coli*. Acta Biochim Biophys Sin (Shanghai). 47. 488–495.

20 Wang, L., Watzlawick, H., Fridjonsson, O., Hreggvidsson, G., and Altenbuchner, J. (2013). Improved soluble expression of the gene encoding amylolytic enzyme Amo45 by fusion with the mobile-loop-region of co-chaperonin GroES in *Escherichia coli*. Biocatal. Biotransformation. 31. 335–342.

ACKNOWLEDGEMENTS

Financial support: Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP grant numbers 2021/02930-1, 2018/13469-0, 2017/24832-6 and 2016/50413-8. Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq grant numberS 131307/2021-5 and 310973/2022-8. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES grant number 88887.512025/2020-00, and Fundação Butantan.