Protein Expression and Purification 149 (2018) 23–30

Contents lists available at ScienceDirect

Protein Expression and Purification

journal homepage: www.elsevier.com/locate/yprep

Effect of rare codons in C-terminal of green fluorescent protein on protein production in *Escherichia coli*

Yaru Yan[a,b,1], Xiaoqing Liu[a,1], Qingbin Li[c,a], Xiaoyu Chu[a], Jian Tian[a,*], Ningfeng Wu[a]

[a] Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing, 100081, China
[b] College of Food Science and Technology, Agricultural University of Hebei, Baoding, HeBei Province, 071001, China
[c] College of Biological Sciences, China Agricultural University, Beijing, 100094, China

ARTICLE INFO

Keywords:
Escherichia coli
Rare codons
eGFP
Fluorescence intensity
C-terminal region

ABSTRACT

In the previous study, the results on two interesting *egfp* genes indicated that the expressed eGFP production of *egfp-codon* containing multiple rare codons was 2.3-fold than that of *egfp-genscript* with mainly high-frequency-usage codons. Therefore, the rare codons also play important roles for the functional expression of genes and it is interesting to know which rare codons in the *egfp* affect the functional expression of eGFP. In this study, the structure-guided SCHEMA recombination method and site-specific mutagenesis were proposed to detect the contribution of the rare codons on the functional expression of eGFP. The 12 chimeric *egfps* were generated from *egfp-codon* and *egfp-genscript* by the software SCHEMA. The results indicated that it was the rare codons in the C-terminal coding region (residues from 147 to 239) of eGFP resulting in the higher expression levels in *Escherichia coli*. The single and multiple point mutations also indicated that the presence of rare codons in 3′ coding regions of *egfp* could enhance the functional expression of eGFP in *E. coli*. Therefore, the gene sequence on the C-terminal could also affect its functional expression and the strategy of substituting rare codons into coding sequences might be an effective method for increasing heterologous proteins in the host.

1. Introduction

Most amino acids are encoded by two to six synonymous codons in all genomes, however these synonymous codons occur with different frequencies [1–4]. This is a phenomenon known as codon bias [1–5], which is a universal feature of eukaryotic and prokaryotic genomes that is thought to have evolved as a mechanism to regulate gene expression and protein folding [6–10]. These synonymous codons can be classified as either frequent codons with frequent codons recognized by more abundant tRNAs [11,12], or rare codons, whilst rare codons partner with less abundant tRNAs.

Escherichia coli is the most commonly used expression system for heterologous proteins [13], as it allows for overexpression of recombinant proteins under the regulatory control of strong promoters (e.g. the T7 promoter) [14]. However, there are many heterologous proteins that are not expressed efficiently in *E. coli*, with their gene sequences known to be a major factor in determining whether functioning soluble proteins are obtained. The expression levels of heterologous proteins in *E. coli* are also known to be affected by codon bias [6,15–17], with strategies involving codon optimization [15,16], use of codon bias-adjusted strains [18], co-expression of protein-folding chaperone proteins [19], having been developed to improve protein expression levels. However, based on those method mentioned above, there are still lots of proteins that could not be highly expressed in the expression host.

An increasing number of studies have shown that synonymous substitution can have a significant impact on the rate of mRNA translation, which, in turn, can affect protein expression levels downstream [8,20,21]. Both the rare codons and the high-frequency-usage codons could affect the functional expression of the target gene [6,22,23]. It is known that substitution of synonymous codons into genes at the 5′-ends of mRNA can affect their structure and stability, leading to changes in their rate of translation, initiation and elongation [24–27]. For example, introduction of low-frequency codons into sequences in the 5′ coding region of the *src* gene of *Streptococcus equisimilis* was shown to enhance soluble protein expression levels in *E. coli* [28]. Similarly, introduction of a single synonymous rare codon into the early coding region of the beta-strand of sucrose phosphorylase (SP) resulted in enhanced soluble expression levels when compared to expression levels of constructs containing rare codon substitutions coding for other regions of its protein [29].

In our previous work, we generated two reporter genes encoding

* Corresponding author.
E-mail address: tianjian@caas.cn (J. Tian).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.pep.2018.04.011
Received 16 February 2018; Received in revised form 2 April 2018; Accepted 16 April 2018
Available online 17 April 2018

enhanced green fluorescent protein (eGFP) that coded to the same amino acid sequence and were expressed in *E. coli* [30]. The first gene was *egfp-codon* that contained a combination of rare and frequent codons, whose sequence was designed using the Presyncodon method [30]. The second gene was *egfp-genscript*, which mainly contained high-frequency-usage codons in *E. coli* and was designed using GenScript software. Although these two genes were designed to encode the same amino acid sequence, the expressed fluorescence intensity for eGFP encoded by *egfp-codon* was found to be 2.3 fold greater than for eGFP encoded by *egfp-genscript*. This clearly demonstrated that the presence of rare codons could be used to improve the expression of eGFP reporter genes in *E. coli*. However, the mechanism that these rare codons employ to improve eGFP expression was still unknown.

In this study, we have employed a series of fragment replacements in *egfp-codon* and *egfp-genscript* genes to determine the effect of rare codon in different regions of these genes on recombinant eGFP production in *E. coli*. The code for the C-terminal region of eGFP was shown to be important to the protein folding and greater levels of functional eGFP. To the best of our knowledge, this is the first report describing that the presence of rare codons coded for the C-terminal of a protein can affect the soluble expression of a gene in the host. These results provide further evidence that the introduction of rare codons into some selected gene sequences may be used to enhance the expression of heterologous proteins in *E. coli*.

2. Results

2.1. Differences in codon usage between egfp-codon and egfp-genscript

Differences in codon usage between *egfp-codon* and *egfp-genscript* were determined by calculating their Codon Adaptation Index (CAI) scores, with *egfp-codon* having a lower CAI score than *egfp-genscript* for most codons, except for codons in the 175–185 region, where *egfp-codon* had a higher CAI score (Fig. S1). Analysis of the codon composition of both genes revealed that *egfp-codon* contained a much larger number of rare codons than *egfp-genscipt* (Table S1). For example, there are 21 codons that can potentially encode for leucine of eGFP, with *egfp-genscript* only containing a single frequently used CUG codon for leucine, whilst the *egfp-codon* contained 8 CUG codons and 5 other types of codon coding for leucine. Additionally, 8 out of the 10 codons encoding for proline that were present in the *egfp-codon* gene had codon usage frequencies below 10‰ in *E. coli* genes. The levels of soluble eGFP expressed by *egfp-codon* were found to be approximately 2.3-fold higher than for *egfp-genscript* [30]. Therefore, the rare codons in the genes can have a positive effect on the expression levels of eGFP in *E. coli*.

2.2. Rare codons present in egfp-codon have no influence on the structure and character of eGFP

Synonymous codon substitutions have been reported to change the structure and character of proteins [10,24,31–33]. Consequently, we next investigated whether the greater absorption found in the fluorescent assay used to measure levels of eGFP encoded by egfp-codon might be a result of structural changes, rather than increased levels of protein. To confirm this, we purified the two eGFP proteins encoded by egfp-codon and egfp-genscript (Fig. 1A), and analyzed their secondary structure using circular dichroism (CD) spectroscopic analysis in the far-ultraviolet region (Fig. 1B). Comparison of their CD spectra, although the spectrum appeared to be a bit of a difference and the minimum at about 218 had shifted, the secondary structure of the two proteins were essentially identical. Furthermore, Differential Scanning Calorimetry (DSC) measurements of temperature on the conformational stability of eGFP encoded by *egfp-codon* and *egfp-genscript* showed identical T_m values of 87.2 °C for both proteins (Fig. 1*C*). This data suggests that the presence of rare codons in the *egfp-codon* gene did not result in any significant changes to the structure and fluorescent character of expressed eGFP proteins.

2.3. Single rare codon substitution does not result in improved expression of eGFP

We next attempted to determine which rare codons in *egfp-codon* were contributing the most important role in improving the functional expression of eGFP. Therefore, site-directed mutagenesis techniques were used to replace 13 frequent codons in the *egfp-genscript* gene with rare codons present in the analogous regions of the *egfp-codon* gene (Table 1). The fluorescent intensity of the eGFP proteins that were expressed by these 13 *egfp-genscript* single mutants was found to be essentially unchanged when compared to those obtained from the *egfp-genscript* gene (Fig. 2). Only *egfp-genscript*-14 and *egfp-genscript*-212 were seen to express slightly higher levels of eGFP protein, with slightly lower levels of protein expression observed for some of the other single mutants. Therefore, these results clearly indicate that single rare codon substitution in the *egfp-genscript* gene is ineffective in improving overall production levels of eGFP protein.

2.4. Introduction of rare codons into the C-terminal coding region of egfp-codon results in increased production of eGFP

We next attempted to determine what effect that substitution of rare codons into different regions of the *egfp-codon* gene would have on eGFP expression levels. A structure-guided SCHEMA recombination method was used to generate 12 chimeric *egfps* [34], using a structural template provided by the tertiary structure of the eGFP protein for guidance. Firstly, the three-dimensional (3D) structure of eGFP (Fig. S2) was used to divide the coding sequence of *egfp-codon* and *egfp-genscript* into three sections (I, II and III). These sections corresponded to sequences coding for the first 1–81 amino acid (aa) residues, the middle 82–146 aa residues and the terminal 147–239 residues of eGFP, respectively. The first, second and third sections of these transcripts code for three (1–3), three (4–6) and five (7–11) beta sheets of the eGFP beta-barrel structure, respectively. As shown in Fig. 3B, three chimeric *egfp* (*egfp-c1*, *egfp-c2* and *egfp-c3*) were constructed, by individually replacing three sections of *egfp-genscript* with analogous regions of *egfp-codon*. The expression levels of *egfp-c3* for eGFP production were found to be increased by 161% when compared to *egfp-genscript,* and to 93% when compared to *egfp-codon*. However, there was little change in the amount of eGFP produced by *egfp-c1*, *egfp-c2* when compared to *egfp-genscript* (Fig. 4A). Therefore, rare codon substitutions in the *egfp-genscript* region that code for the C-terminal region (147–239) of eGFP are responsible for producing higher levels of protein expression.

We then decided to try and determine which rare codons in which regions of *egfp-codon* were responsible for the improved expression levels of eGFP. Section III of the gene was divided into three subsections [III-1 (codons for 147–173 aa), III-2 (codons for 174–195 aa) and III-3 (codons for 196–239 aa)], corresponding to the regions coding for the 7, 8–9 and 10–11 beta sheets of eGFP, respectively. These three subsections in *egfp-genscript* were replaced with analogous subsections of *egfp-codon* respectively to construct *egfp-c3-1*, *egfp-c3-2* and *egfp-c3-3* (Fig. 3*B*). Compared with *egfp-genscript*, the expression levels of these three constructs were not increased significantly, with the expression levels of *egfp-c3-1* unexpectedly decreased by 50% (Fig. 4*B*). Since replacement of single fragments in region III of *egfp-codon* in *egfp-genscript* had not improved the expression levels of *egfp-genscript*, it was decided to replace two of subsections III-1, III-2 and III-3 of *egfp-genscript* with their analogous regions from *egfp-codon,* to yield *egfp-co-1* (combination of III-2 + III-3), *egfp-co-2* (combination of III-1 + III-3) and *egfp-co-3* (combination of III-1 + III-2) (Fig. 3*B*). The expression levels of *egfp-co-1* and *egfp-co-2* was found to be increased by 11% and 65%, respectively, while the expression level of *egfp-co-3* were essentially the same as *egfp-genscript* (Fig. 4*C*). Therefore, it was concluded that all three