Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 347 (2005) 11–19
GENE
SECTION
FUNCTIONAL GENOMICS
www.elsevier.com/locate/gene

In silico diagnosis of inherently inhibited gene expression focusing on initial codon combinations

Yoshiaki Ohashi[a], Akiko Yamashiro[b], Takanori Washio[a,c], Nobuyoshi Ishii[a], Hideyuki Ohshima[a], Tetsuko Michishita[a], Masaru Tomita[a], Mitsuhiro Itaya[a,d,*]

[a]*Institute for Advanced Biosciences, Keio University, 403-1 Nipponkoku, Daihoji, Tsuruoka, Yamagata 997-0017, Japan*
[b]*Faculty of Nursing and Medical Care, Keio University, 4411 Endo, Fujisawa, Kanagawa 252-8530, Japan*
[c]*Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan*
[d]*Mitsubishi Kagaku Institute of Life Sciences, 11 Minamiooya, Machida, Tokyo 194-8501, Japan*

Received 24 May 2004; received in revised form 19 October 2004; accepted 17 November 2004
Available online 4 January 2005
Received by D.L. Court

Abstract

The translation start site, immediately downstream from the start codon, is a dominant factor for gene expression in *Escherichia coli*. At present, no method exists to improve the expression level of cloned genes, since it remains difficult to find the best codon combination within the region. We determined the expression parameters that correspond to all sense codons within the first four codons using *GFPuv* which encodes a derivative of green fluorescent protein. Using a genetic algorithm (GA)-based computer program, these parameters were incorporated in a simple, static model for the prediction of translation efficiency, and optimized to the expression level for 137 randomly isolated *GFPuv* genes. The calculated initial translation index (*ITI*), also proven for the *DsRed2* gene that encodes a red fluorescent protein, should provide a solution to overcome the gene expression problem in cloned genes whose expression is often inherently blocked at the translation process. The proposed method facilitates heterologous protein production in *E. coli*, the most commonly used host in biological and industrial fields.

Keywords: DsRed; Genetic algorithm; Green fluorescent protein; Translation; *Escherichia coli*

1. Introduction

The technology of heterologous gene expression in the commonly used bacterial host *Escherichia coli* has facilitated advances not only in industrial and pharmaceutical fields but also in biology. In the course of gene expression, initiation of the translation process is a critical step (Makrides, 1996; Jonasson et al., 2002). When synonymous codons were used, differences between the host genome and heterologous cloned genes resulted in a substantial delay on the elongation rate (Rosenberg et al., 1993; Kane, 1995).

The absence of correlation between codon bias and gene expression under specific circumstances (Young et al., 1983; Ernst and Kawashima, 1988; Lee et al., 1992) indicates that the translation rate of genes is simultaneously modulated not only by codon bias but also by other unknown factor(s). Several reports indicate that codons immediately following the initiation codon affect the expression level of reporter genes in *E. coli*. Looman et al. (1987) reported that the second codon (+2 codon) influences the expression level of the *lacZ* gene in *E. coli*. Stenström et al. (2001), who demonstrated the same effect using not only *lacZ* but also IgG-binding protein A from *Staphylococcus aureus*, concluded that the effect of the +2 codon associated with the

Abbreviations: DB, downstream box; DE, downstream element; DR, downstream region; DsRed, *Discosoma* sp. red fluorescent protein; GA, genetic algorithm; GFP, green fluorescent protein; *ITI*, initial translation index.

* Corresponding author. Institute for Advanced Biosciences, Keio University, 403-1 Nipponkoku, Daihoji, Tsuruoka, Yamagata 997-0017, Japan. Tel.: +81 235 29 0526; fax: +81 235 29 0529.
E-mail address: mita2001@sfc.keio.ac.jp (M. Itaya).

0378-1119/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.gene.2004.11.027

amount of recognizing tRNA and the high adenine content of the +2 codon. The codons within the sequence following the initiation codon (downstream region) represent a rate-determining factor of translation. Secretion of the alkaline phosphatase gene attached heat-stable enterotoxin II signal sequence is influenced by the codon combinations in the downstream region (Simmons and Yansura, 1996). Stenström and Isaksson (2002) also found that the codons following the +2 codon slightly affect the expression level. Griswold et al. (2003) reported that optimization of the first seven codons of the *Fusarium solani* cutinase gene reduces the expression level in *E. coli*. These observations are supported by bioinformatics data that suggest that codon usage within the downstream region of genes is biased in the *E. coli* and *Bacillus subtilis* genome (Bulmer, 1988; Rocha et al., 1999; Ohno et al., 2001; Sato et al., 2001). The stability of mRNA also influences the gene expression levels. Because RNase E may recognize the AU-rich consensus sequence of RNA and may initiate the decay of mRNA (Ehretsmann et al., 1992; Steege, 2000), these effects on gene expression cannot be excluded.

Current hypotheses regarding the role of the downstream region are confusing. The downstream box (DB), located immediately downstream from the start codon, was originally defined as a sequence complementary to the anti-DB found in 16S rRNA (Sprengart et al., 1996). As direct interaction between the DB and the anti-DB has been ruled out by several independent studies (O'Connor et al., 1999; Moll et al., 2001), the mechanism(s) underlying the enhancing effect by the DB remains unclear. Etchegaray and Inouye (1999) and Sato et al. (2001) suggested that the DB is involved in the enhancement of translation initiation in concert with the ribosome-binding site (RBS).

Using a genetic algorithm (GA), we determined the contribution of the initial codons to gene expression. Our approach yielded a method for the prediction of gene expression based on the first four codons.

2. Materials and methods

2.1. Plasmid construction

The TA-cloning of PCR products was performed as previously described (Ohashi et al., 2003). The PCR products of *GFPuv* (BD Biosciences Clontech, Palo Alto, CA, USA) variants were amplified using the forward primer 5′-ATG-12nt-AGTAAAGGAGAAGAACTTTTC-3′, where 12 nt stands for the individual codon quartets, and reverse primer GFPUVR 5′-TTATTTGTAGAGCTCATCCATG-3′. To construct random *GFPuv* variants, the 12-nt sequence in the forward primer was designed as an N dodecamer. The *GFPuv* variants thus amplified were ligated to the T-vector of pHASH103. For DB, the forward primer DB1 (5′-ATGA-ATCACAAAGTGAGTAAAGGAGAAGAACTTTTC-3′) was used. The *DsRed2* derivatives were amplified by PCR using pDsRed2 plasmid (BD Biosciences Clontech) as a template. The sequences of primers used are DsRed2N12-F, 5′-ATG-12N-ACCATGATTACGCCAAGC-3′ and DsRed2-R, 5′-CTACAGGAACAGGTGGT-3′. pHASH121 was used to clone and express the *GFPuv* variants in *E. coli* (Ohashi et al., 2003). The W3110 *E. coli* strain was used to determine the expression level of *GFPuv* and its variants.

2.2. Growth conditions and fluorescence measurement

LB medium was selected for *GFPuv* production because it is widely used to express cloned genes in *E. coli* (Maniatis et al., 1982). *E. coli* cells were pre-grown in medium containing 50 μg/ml ampicillin (Ap) at 37 °C for 12 h. Culture aliquots 10 μl were inoculated into 1 ml of fresh LB medium containing 50 μg/ml Ap and this was followed by 37 °C incubation with shaking. When the cells reached the late exponential phase, they were harvested by brief centrifugation, cell pellets were washed once with Milli-Q water, and then resuspended in 500 μl of Milli-Q water. Finally, a 10-μl portion of the cell suspension was diluted 20-fold with Milli-Q water and fluorescence intensity (excitation at 395 nm and emission at 509 nm) and optical density (OD) at 600 nm were measured simultaneously using a fluorescence Spectra Max GeminiXS spectrometer and a Spectra Max Plus384 spectrophotometer (Molecular Devices, Sunnyvale, CA, USA), respectively.

2.3. Computation of Gibbs' free energy

To analyze the mRNA secondary structure, we used RNA secondary structure prediction software (RNAstructure Ver.3.71;http://128.151.176.70/RNAstructure.html). Nucleotides located between the 20th base pair (bp) downstream of the start codon and the 15th bp of the downstream region in the coding region were used to calculate changes in free energy values upon the formation of specific mRNA secondary structures.

2.4. Modeling and parameter tuning using GA

Parameters for each codon were obtained with the *GFPuv* variants containing sense codon quartets as the downstream elements. Optimization of the parameters to the expression data experimentally obtained (see Table 1) was performed with a real coded GA (RCGA) (Herrera et al., 1998). The RCGA was obtained with the following steps:

(i) Generation of an initial population. Random numbers (±1.0) were assigned and summed to each parameter for 20 individuals. Random numbers were generated using a linear congruence method.
(ii) Evaluation. Evaluation was done to fit the calculated *ITI* and the expression values for the 137 *GFPuv* variants to the individuals using the square of the correlation coefficients (fitness value).