Published in final edited form as:

Protein Expr Purif. 2010 December ; 74(2): 248–256. doi:10.1016/j.pep.2010.06.005.

Influence of the Second Amino Acid on Recombinant Protein Expression

Louis Bivona, Zhongcheng Zou, Nicole Stutzman, and Peter D. Sun*

Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 12441 Parklawn Drive, Rockville, MD 20852

Abstract

Factors affecting protein expression have been intensely studied to the benefit of recombinant protein production. Through mutational analysis at the +2 amino acid position of recombinant Igα, we examined the effect of all 20 amino acids on protein expression. The results showed that amino acids at the +2 position were affected 10-fold in the recombinant protein expression. Specifically, Ala, Cys, Pro, Ser, Thr, and Lys at the +2 position resulted in significantly higher expression of recombinant Igα than other amino acids, while Met, His and Glu resulted in greatly reduced protein expression. This expression difference depended on the amino acid instead of their codon usages. Consistent with the mutational results, a statistically significant enrichment in Ala and Ser at the +2 position was observed among highly expressed *E. coli* genes. This work suggests a general approach to enhance protein expression by incorporating an Ala or Ser after the initiation codon.

Keywords

Recombinant expression; 2nd codon position; Igα

Introduction

The efficiency of translation initiation plays a key role in determining the amount of protein expressed in all organisms. Initiation in *Escherichia coli* has been intensively studied and involves multiple *cis*- and *trans*-acting factors, each contributing to the overall efficiency of translation. Several important *cis*-acting factors involved in translation initiation include the Shine-Dalgarno sequence and its spacing relative to the start codon [1-3], the initiation codon itself (AUG, GUG, rarely AUU or CUG) [4,5], and the non-random distribution of bases located both upstream and downstream of the start codon [6,7].

The downstream region immediately to the 3' side of the initiation codon has been the focus of many studies related to protein expression. It has been previously shown that changes at the second amino acid position can lead to a 15-fold difference in expression [8]. This work was expanded by Stenstrom et al, in which it was shown that the AAA triplet is the most prevalent codon at the second amino acid position [9]. In addition, they showed that the AAA codon resulted in higher gene expression at the +2 position. However, it has also been reported that

*Address correspondence to: Sun P.D., Ph.D. 12441 Parklawn Drive, Rockville, Maryland 20852. Tel.: 301-496-3230. FAX: 301-402-0284; psun@nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

tandems of AGA or AGG codons promote favorable translational efficiency [10]. In highly expressed *E. coli* genes, it was found that guanosine is most frequently represented at the first codon position [11]. In addition to finding codons that positively affect protein expression, codons consisting of the sequence NGG at the second amino acid position showed marked decreases in expression [12]. Others have found that identical pairs of all four of the CGN triplets result in drastically inefficient translation [10]. In general, early studies favor nucleotide specific determinants for protein expression, namely codon bias.

Combining a mutational analysis on recombinant protein expression and statistical analysis on highly expressed *E. coli* genes, we show that individual amino acids at the +2 position can greatly affect their protein expression, and that the preference appears to be in amino acid rather than codon usage. The results suggest a general approach to increase protein expression level by a simple insertion of serine or alanine at the second position.

Materials and Methods

Mutation of second amino acid

The pET30 vector containing the extracellular domain of Ig α was kindly provided by Pavel Tolar. Mutations were made to the second codon of Ig α by using Stratagene's QuikChange® II Site-Directed Mutagenesis Kit. Primers were designed by using Stratagene's QuickChange® Primer Design Program (Supplemental Table 1), and synthesized by Integrated DNA Technologies (IDT). After transformation of the mutated plasmid into XL1-Blue Supercompetent *E. coli* cells, the cells were spread on Luria-Bertani (LB) agar plates supplemented with 25 μ g/ml kanamycin overnight at 37 °C. Colonies from the transformation were inoculated at 37 °C in 20 ml of LB broth supplemented with 25 μ g/ml kanamycin, and plasmid DNA were extracted using Qiagen's MiniPrep kit. The mutations were confirmed by DNA sequencing (ACTG Inc.). The above procedure was also performed for the mutagenesis of CXCL10.

Protein expression in *E. coli*

Mutant Ig α or CXCL10 plasmids were transformed into *Escherichia coli* BL21 (DE3) cells (Novagen) for protein expression. Individual colonies from the transformation were inoculated in 20 ml of LB broth and grown at 37 °C for overnight. 50 μ l of the overnight cultures were inoculated in 5 ml of antibiotic free LB media, and induced at an OD₆₀₀ between 0.6-0.8 with 1 mM isopropyl β -D-1-thiogalactopyranoside (IPTG) for 3 hours. The cells were then centrifuged and the pellets were resuspended in B-PER® (Bacterial Protein Extraction Reagent) (Pierce) lysis buffer supplemented with 40 μ g/ml DNase for at room temperature 10 min (4 ml of BPER per 1 g of wet cell pellet). The lysates were centrifuged at 13000 rpm for 5 min. As Ig α forms inclusion bodies in *E. coli*, the supernatants were removed and the pellets were resuspended in 350 μ l of water. Ig α expression was evaluated by SDS-gel electrophoresis using Laemmli Buffer (Sigma) and 4-12% NuPAGE Bis-Tris Gels (Invitrogen) for 1 $\times$, 2 $\times$, and 4 $\times$ dilutions of the insoluble cell lysate fractions to avoid saturation of Ig α band (Figure 2). The SDS-gels were stained with Coomassie blue, destained, and scanned for intensity analysis using the ImageJ densitometry program. The intensity of Ig α was normalized against a 40 kDa endogenous bacterial that was used as a gel loading control band to take account the SDS-gel sample loading affect.

Statistical analysis of second amino acid position

The 200 highest expressing *E. coli* genes was listed by Ishihama et al [13], in which they used a quantitative mass spectrometry approach to determine cytosolic protein concentrations. For comparison, 200 genes were selected at random using the online *E. coli* genome database,

EcoCyc [14]. The amino acids at the second position were counted and graphed. These calculations were also performed for the 3rd, 4th, 5th, and 40th amino acids.

RNA structure prediction

RNA structures were predicted between the Shine-Dalgarno sequence and +24 nucleotides from the initiation codon using the RNAfold webserver (<http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>).

Results

Serine mutation at the second amino acid position increased Igα expression

pET plasmids remain the most widely used series of prokaryotic vectors for recombinant protein expression (Novagen, Inc), partly because the presence of various multiple cloning sites with different conveniently built-in affinity tags. In the existing pET vectors, the +2 amino acid immediately after the initiation methionine varies among Gly, Ala, Ser, Asp, Asn, Pro, His, Lys, and Arg. While the choice of a specific pET vector is primarily based on the availability of unique cloning sites and desirable affinity tags, we have occasionally cloned our target genes into two or more pET vectors. We observed in several cases that significant variations in recombinant protein expression existed among different pET vectors. Specifically, the pET vectors with +2 amino acids being Ala and Ser appeared to result in 2-5 fold higher expression of the same recombinant protein. When the mature sequence of a human B cell co-receptor, Igα, was cloned into a pET-30a vector using the 5'-NdeI site, which resulted in a simple addition of the initiation Met to the N-terminal leucine of the mature Igα sequence, this wildtype Igα was expressed at low levels in BL21(DE3) strain of *E. coli* (Fig. 1). When the wildtype Igα construct was transformed into a Rosetta strain of BL21 cells, it resulted in a 1.7-fold increase in Igα expression, consistent with observed improvements in expression using the Rosetta strain of bacteria [15]. A subsequent mutation of Leu to Ser at the +2 position further increased the Igα expression by 2-fold over the wildtype in BL21 Rosetta cells (Fig. 1). Thus, the combination of Rosetta cells and the Ser mutation resulted in almost a 4-fold increase in Igα expression.

Effect of other amino acids at +2 position on the expression of Igα protein

To further evaluate whether other amino acids at the +2 position also influence the Igα expression in *E. coli*, we systematically replaced the +2 position Leu of Igα with all 19 amino acids using QuikChange® II Site-Directed Mutagenesis and analyzed the effect of mutations to the protein expression. After mutagenesis, the plasmids were transformed into *E. coli* BL21 (DE3) cells. Individual colonies were cultured in LB broth and induced with 1 mM IPTG at OD₆₀₀ between 0.6-0.8. In all cases, individual colonies of the same mutants resulted in consistent levels of protein expression with no significant clone variations. For example, three clones from each of the Glu, His and Met mutations as well as two clones from the Ser and Val mutations showed little variation in their Igα expression (Fig. 2A & 2B). However, substantial variations in Igα expression were observed among different mutants (Fig. 2C & 2D, Table 1). To quantify the differences in mutant Igα expression, serial dilutions of the wildtype and mutant expression samples were analyzed by SDS-gel electrophoresis and the intensities of Igα bands were quantified using the ImageJ densitometry program. The overall difference in Igα expression between the highest and the lowest amino acid substitutions at this position exceeded 10-fold. Specifically, Ala, Cys, Pro, Ser, Thr, and Lys, at the +2 position had the highest expressions of Igα that were approximately 2-fold higher than wild type Leu. Ile, Asp, Phe, Gln, Gly, Val, Asn, Arg, Trp, and Tyr mutations expressed the next highest levels. Glu, His and Met mutants expressed 2 to 5-fold lower than the wild type.

The second position amino acid preference among highly expressed proteins in *E. coli*

While individual amino acids at the +2 position of Igα affected its expression, it is not clear if the observation has a broader implication. Equally uncertain is whether the observed correlation between amino acids and their protein expressions is gene specific. Namely, the amino acids that contributed most to Igα expression may have no effect or even negatively impact other gene expressions. To address the generality of the current observation, we compared the enrichment of individual amino acids at the +2 position between groups of highly expressed and randomly selected proteins in *E. coli*. A wealth of genomic and proteomic data have been accumulated on *E. coli*, including their protein expression profiles as determined by mass spectrometry. Using an exponentially modified protein abundance index (emPAI) [13], we ranked the level of *E. coli* gene expression from their mass spectrometry profiles, and selected the top 200 genes as a representation of highly expressed genes in *E. coli*. As a comparison, an additional list of 200 genes was chosen randomly from the database EcoCyc. The enrichment of an individual amino acid at a given position is calculated as the percentage of that amino acid appearing at the given position among the 200 selected genes. The overall average enrichment of an individual amino acid is calculated as the percentage of that amino acid occurring at any position among the selected genes. Of the 200 highly expressed genes, the most abundant amino acids at the +2 position are Ala, Ser and Lys occurring at 21%, 19.5% and 15%, respectively (Fig 3). Together, the three amino acids appeared at the +2 position in over 50% of the highly expressed proteins. When compared with their overall enrichment in these proteins, Ala, Ser and Lys are the top three amino acids that appear more frequently at the +2 position than their overall enrichment. Specifically, they occur 2× (Ala), 4× (Ser) and 3× (Lys) more frequent at the second position than overall. The frequencies of Ala, Ser and Lys found in the 200 randomly chosen proteins are 13.5%, 14% and 15.5%, respectively. While all three occur more frequently at the +2 position than other amino acids even in the control group, there is a clear enrichment of Ala and Ser among the more abundantly expressed genes, suggesting their potential role in enhancing gene expression. In contrast, the five lowest expressing Igα constructs containing residues Met, His, Glu, Leu or Arg showed lower representations at the +2 position than their average frequencies.

To further investigate if significant deviations in amino acid enrichment from their averages occur at other positions, the statistical amino acid enrichment analysis was also carried out for the 3rd, 4th, 5th and 40th positions between the highly expressed and the control groups of genes (Fig 3). While Lys remained enriched at the 3rd, 4th and 5th position in the over expressed gene group compared to the control group, there were no differences in the preference of Ala and Ser between the two groups at these positions. At the 40th position, the frequencies of individual amino acids from the highly expressed gene group agreed well with those from the control group and with their overall average frequencies, namely both groups converged to their overall average frequencies. All together, these results show that Ala, Ser and Lys are significantly enriched while His, Met, Leu and Tyr are poorly represented when compared to their average occurrence at the +2 position among the highly expressed genes and these differences disappear as the analysis is further extended downstream from the translational initiation site. The correlation between the high frequency occurrence of Ala, Ser and Lys at the second position among the highly expressed genes and their increased expression in Igα mutants, and between the low frequency occurrence of His and Met and their lower expressions in Igα mutants suggest a significant role for amino acids at +2 position to gene expression in general.

Isocodon effect on expression

To further evaluate codon variation to protein expression at the second position, we generated isocodon mutations for the Ala, Ser and Lys mutants of Igα. All four Ala codons (GCU, GCC, GCA, GCG) resulted in comparable expression of Igα (Fig. 4A). Similarly, the protein expressions were consistent among the six serine (AGU, AGC, UCU, UCA, UCC, UCG) and

two lysine (AAA and AAG) codons (Fig. 4B & 4C). The data indicate that while subtle variations may occur associated with different codons, their overall effect on protein expression is marginal. Thus, these data demonstrated a non-distinctive codon usage for Ala, Ser and Lys at the second position. Further analysis between isocodons at the second position of highly expressed and randomly selected genes has provided more evidence for the important effects that the amino acid has on translation (Fig. 4D). For Ser and Ala, there seems to be no real clear bias towards any particular codon. In fact, isocodons appear at relatively the same frequency in the highly expressed and randomly selected groups. Lysine, however, is biased toward the AAA codon among highly expressed *E. coli* genes. Specifically, nearly all the +2 Lys residues from the highly expressed group use AAA codon compared to about two-third of them from the random group (Fig 4D).

Nucleotide composition of the second codon correlates little with protein expression level

Previous reports suggested that changes in gene expression can be affected by specific nucleotide at the second position [10-12]. For example, efficient expression of gene variants was seen when the second position contained the AGA (Arg) codon [10], guanine (G) at the first codon position correlated with high expression of *E. coli* genes [11], and NGG codons appeared to reduce gene expression [12]. To further investigate if the expression of Igα mutants correlated with a potential nucleotide or codon preference, the compositions of the three nucleotides for the second codon of mutant Igα were displayed against their expression level (Fig. 5). The result showed little correlation between mutant Igα expression and their first and second nucleotide compositions. High expressing mutants, Ala, Ser, Thr, Pro, and Lys contain codons that begin with each of the four (A, C, U, G) nucleotide base (Fig. 5A). The lowest expressed His, Met and Glu mutants have three of the four nucleotides (C, A, G) at the first base. Codons that began with A or had two A's did not perform exceptionally well compared to other codons. NGG codons, encoding Arg, Gly, and Trp, expressed wildtype like level of Igα. When grouped by their second base, nucleotides A, U, and G resulted in similar expressions, whereas C resulted in a higher expression (Fig. 5B). It should be noted that codons with C as the second nucleotide represent the highest expressing amino acids A, P, S, and T of Igα. Furthermore, Ser has two codons with G as the second base, each isocodon expressing the same amount of protein as its other isocodons (Fig 4B). Overall, the average expressions for the individual nucleotide at both the first and second position of the +2 codon do not deviate significantly with their compositions. In all, our experiments showed no clear codon bias at the second amino acid position suggesting that protein expression maybe more determined by the type of amino acid rather than their nucleotide composition at the second position.

CXCL10 expression is increased by mutating +2 Val to Ala and Ser

To evaluate the broader implication of the +2 position amino acid to protein expression, we mutated the wildtype Val at the +2 position of CXCL10, an interferon-γ inducible chemokine, to Ala and Ser. Similar to the Igα mutations, both Ala and Ser mutations resulted in mild increases in the protein expressions, 1.2 and 1.5-fold for Ala and Ser mutations, respectively (Fig 2F), indicating that the +2 position Ala or Ser can be used to improve the expression of proteins besides Igα.

Effect of local RNA structure and stability on protein expression

It is possible that by creating mutations so close to the initiation complex, the mRNA structure is altered, which affects translational efficiency. Zhang et al have found from huIL-10 that when the initiation codon is not involved in complementary base pairing, 10-fold higher protein expression can be achieved [16]. Upon analyzing catechol-*O*-methyltransferase, Nackley et al found that in terms of protein expression, more abundant levels were found when the mRNA strand was least stable [17]. We analyzed the effect of mutations at the +2 site to RNA structure

and stability using the RNAfold webserver. Although the initiation AUG codon is located in an unpaired loop conformation in many mutant structures, there are significant variations among the mutants and their predicted RNA structures lack a conserved motif (Fig. 6A, 6B, supplemental Figure S1). Nevertheless, the calculated RNA stabilities for some +2 residues, such as Ala, Pro, Thr, and His correlated with their expression levels. Namely, the mutants with more stable RNA structure resulted in less protein expression (Fig 6C, Table 1). In general, the calculated RNA stabilities did not correlate with their expression. For instance, the well expressed Ser mutation adopted one of the most stable RNA structures minimum free energy of -7.40 kcal/mol, well below that of the wildtype Leu (-5.7 kcal/mol). Similarly, the glutamate mutant resulted in poor expression in Igα, yet is predicted to have less stable RNA structure with minimum free energy of -3.14 kcal/mol than the wildtype (supplemental Fig S1).

Discussion

Multiple factors contribute to the efficiency of gene translation, including the secondary structure of mRNA, the Shine-Dalgarno sequence, the drop-off rates of peptidyl-tRNAs, and the sequence of the mRNA proximal to the initiation site. In general, early studies favor nucleotide specific determinants for protein expression, namely codon bias. For example, early studies concluded that the AAA triplet leads to higher protein expression and that proteins with adenine-containing codons close to the initiation site express more efficiently [8,9]. Others have shown that efficient expression can be obtained when the second codon is AGA or when codons have G at the first position [10,11]. The second codon may also affect protein expression negatively, as has been found with the NGG triplets [12]. Through mutational analysis, we systematically investigated the influence of individual amino acids at the +2 position to the expression of recombinant Igα. The results showed that amino acids at the +2 position affected significantly recombinant Igα expression and variations in the recombinant protein expression were observed to exceed 10-fold depending on the amino acid. Small amino acids, such as Ala, Ser, Cys, Pro, Thr, and Lys at the +2 position resulted in higher expression of Igα than bulky ones (Met, His, Trp, Glu, Leu, Arg). Met, His and Glu at this position produced the lowest amount of the protein. While codon-bias is often found to be important to gene expression, the variation in recombinant Igα expression at the +2 position appears to be determined by the type of amino acid rather than their nucleotide usage as no base-pair preference was observed among the high expression Igα mutants and the isocodon mutations of serine and alanine resulted in similar expression levels of Igα. Furthermore, when isocodon usage of Ser, Ala, and Lys were examined, Ser and Ala showed no significant codon preference between the high expressing and randomly chosen groups (Fig. 4D). The AAA codon of Lysine appears preferred over the AAG codon among the highly expressed genes compared to the control group. Thus, while codon preference may influence gene expression, there are clearly amino acid related factors important for expression.

The statistical survey of amino acid composition in highly expressed *E. coli* genes supports the preference of alanine and serine at the second position, correlating well with their observed high level of mutant protein expressions (Fig. 3). Similarly, the lowest expressing mutants of Igα, His and Met, are found much less frequently at the +2 position than their overall occurrence. Further, this observed disparity in amino acid preference is unique to the +2 position as the 3rd, 4th, 5th and 40th amino acid positions showed progressively less deviation in their frequency from their mean in both the 200 highly expressed and randomly picked genes. This correlation between the expression of Igα mutants and the frequency of the amino acid occurring at the +2 position suggest that amino acids at this position generally influence their gene expressions. While we did not observe low expressing mutations occurring with high frequency at the +2 position, namely all mutations of high frequency amino acids at +2 position expressed well, we did observe some well expressed mutations, such as Cys and Pro, that are not enriched at the +2 position among the well expressed genes. This suggests other adversary

factors may disfavor their incorporation at the +2 position. For example, Cys and Pro can impose structural restrictions due to their properties of forming disulfides and disrupting regular peptide conformations. As to the contribution of AAA codon (lysine) to gene expression, our results showed that Lys at the +2 position yielded an above average level of Igα expression. Lys was the most common amino acid at the second position in the randomly chosen group of *E. coli* genes, consistent with the earlier results by Looman et al. and Stenstrom et al. [8,9]. However, Lys was not enriched when comparing between the highly expressed and randomly selected *E. coli* genes. Ser and Ala, instead, were the most frequent +2 position residues among highly expressed genes, in agreement with Tats et al. [18].

To investigate if the variation in protein expression at +2 amino acid is determined by their mRNA structure and stability, we calculated the local RNA structures at the initiation site using RNAfold. Overall, there are significant structural variations among the mutant RNAs and their stability did not correlate with their expressions, suggesting factors other than mRNA structure may be important for gene expression. Charges do not appear to be a critical factor as very different expressions of Igα were observed between Asp and Glu mutations, and between Lys and His mutations.

Interestingly, the second amino acid was found to be crucial in determining a protein's half life stability in vivo [19,20]. In particular, the so called N-end rule stipulates that proteins with Phe, Leu, Trp, and Tyr at +2 position tend to have less in vivo stability, whereas those with Ser, Ala, and Thr at +2 position have longer half lives [20]. Our current findings have a striking resemblance to the N-end rule effect. For example, Leu, Trp, and Tyr were found to be in the bottom half of the overall expression ranking, while Ser, Ala, and Thr were among the highest ranked mutations. Since both Igα and CXCL10 were expressed in a non-functional, aggregated inclusion body form in *E. coli* that is relatively resistant to proteolytic degradation, their expression levels are unlikely related to their functional stability. On the other hand, a measured half-life of an intracellular protein may be affected by its expression level. It is tempting to speculate that the increased expression related to serine and alanine at the +2 position may contribute to their long half lives in vivo. Thus, it is conceivable that 1) the mechanism of N-end rule may be directly related to the +2 amino acid translational efficiency, and 2) our current conclusion regarding the contribution of +2 amino acid on recombinant protein expression also applies to soluble proteins.

In summary, our experimental data together with statistical analysis demonstrate the advantage to have serine and alanine inserted as the second amino acid for recombinant protein production.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. M. Gordon Joyce for insightful discussions of the experimental results. The work is supported by the intramural research funding of National Institute of Allergy and Infectious Diseases, National Institutes of Health.

References

1. Shine J, Dalgarno L. Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. *Eur J Biochem* 1975;57:221–30. [PubMed: 809282]
2. Kozak M. Initiation of translation in prokaryotes and eukaryotes. *Gene* 1999;234:187–208. [PubMed: 10395892]

3. Chen H, Bjerknes M, Kumar R, Jay E. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. *Nucleic Acids Res* 1994;22:4953–4957. [PubMed: 7528374]
4. Gren EJ. Recognition of messenger RNA during translational initiation in *Escherichia coli*. *Biochimie* 1984;66:1–29. [PubMed: 6370317]
5. O'Donnell SM, Janssen GR. The initiation codon affects ribosome binding and translational efficiency in *Escherichia coli* of *cl* mRNA with or without the 5' untranslated leader. *J Bacteriol* 2001;183:1277–1283. [PubMed: 11157940]
6. Stenström CM, Holmgren E, Isaksson LA. Cooperative effects by the initiation codon and its flanking regions on translation initiation. *Gene* 2001;273:259–265. [PubMed: 11595172]
7. Stormo GD, Schneider TD, Gold LM. Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Res* 1982;10:2971–2996. [PubMed: 7048258]
8. Looman AC, Bodlaender J, Comstock LJ, Eaton D, Jhurani P, de Boer HA, van Knippenberg PH. Influence of the codon following the AUG initiation codon on the expression of a modified *lacZ* gene in *Escherichia coli*. *EMBO J* 1987;6:2489–2492. [PubMed: 3311730]
9. Stenström CM, Jin H, Major LL, Tate WP, Isaksson LA. Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene* 2001;263:273–284. [PubMed: 11223267]
10. Zamora-Romo E, Cruz-Vera LR, Vivanco-Domínguez S, Magos-Castro MA, Guarneros G. Efficient expression of gene variants that harbour AGA codons next to the initiation codon. *Nucleic Acids Res* 2007;35:5966–5974. [PubMed: 17726048]
11. Gutiérrez G, Márquez L, Marín A. Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency. *Nucleic Acids Res* 1996;24:2525–2527. [PubMed: 8692691]
12. Gonzalez de Valdivia EI, Isaksson LA. A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*. *Nucleic Acids Res* 2004;32:5198–5205. [PubMed: 15459289]
13. Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, Frishman D. Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics* 2008;9
14. Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT, Peralta-Gil M, Santos-Zavaleta A, Shearer AG, Karp PD. EcoCyc: A comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res* 2009;37:464–470.
15. Tegel H, Tourle S, Ottonson J, Persson A. Increased levels of recombinant human proteins with the *Escherichia coli* strain Rosetta(DE3). *Protein Expr Purif* 2010;69:159–167. [PubMed: 19733669]
16. Zhang W, Xiao W, Wei H, Zhang J, Tian Z. mRNA secondary structure at start AUG codon is a key limiting factor for human protein expression in *Escherichia coli*. *Biochem Biophys Res Commun* 2006;349:69–78. [PubMed: 16930549]
17. Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L. Human catechol-*O*-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* 2006;314:1930–1933. [PubMed: 17185601]
18. Tats A, Remm M, Tenson T. Highly expressed proteins have an increased frequency of alanine in the second amino acid position. *BMC Genomics* 2006;7
19. Bachmair A, Finley D, Varshavsky A. In vivo half life of a protein is a function of its amino terminal residue. *Science* 1986;234:179–186. [PubMed: 3018930]
20. Varshavsky A. The N-end rule: functions, mysteries, uses. *Proc Natl Acad Sci* 1996;93:12142–12149. [PubMed: 8901547]

Fig. 1.

Initial results indicating that substituting a serine at the second amino acid position results in higher protein expression in *E. coli*. The lanes represent the un-induced (-) and induced (+) samples of Igα, from left to right, in BL21 (DE3), BL21 Rosetta cells and Leu to Ser mutation in BL21 Rosetta cells. The low molecular weight standard is shown on the left.

Fig. 2.

Coomassie blue stained SDS gels to examine the expression of all amino acid substitutions. (A, B) Multiple clones of individual mutations (Ser, Val, Glu, His and Met) were examined by SDS-gel for their clonal variation. The bands for the 40 kDa endogenous bacterial protein and Igα are indicated. Un-induced (-) and IPTG induced (+) lanes are indicated. (C,D) Expression of mutant Igα at the +2 position with their mutations indicated by single letter amino acid code. (E) The normalized expressions of each Igα mutation. Intensities were derived from three separate experiments. (F) Expressions of CXCL10 with Val (wildtype), Ala and Ser as +2 residues.

Fig. 3.

Amino acid enrichment at the 2nd -5th and 40th positions for 200 highly expressed and 200 randomly chosen proteins from *E. coli*. The frequency of an amino acid appearing at a given position is defined as the ratio between the number of times that amino acid occurred at that position and the total number of amino acids.

Fig. 4.

Isocodon effect on protein expression. (A) Isocodons for alanine mutation are shown in Lane 1-4 for GCG, GCA, GCC, GCU, respectively. Lane 5 is the wild type protein with leucine at the second position. (B) The expression using six serine isocodons are shown in Lane 1-6 for UCG, UCC, UCA, UCU, AGC, AGU respectively. (C) Expression of lysine using both isocodons. Lane 1 is the wild type leucine, Lanes 2 and 3 are AAA and AAG, respectively. (D) Ala, Ser, and Lys codon usage at the +2 position was compared between high expressing genes and randomly selected genes.

Fig. 5. Mutations grouped by their nucleotide usage for the first (A) and second (B) bases of at the +2 amino acid.

Fig. 6. RNA structure prediction of the wild type (A) and alanine (B) mutant. RNA structures were predicted using the RNAfold webserver. The dashed red oval indicates the Shine-Dalgarno sequence, while the AUG initiation codon is located in the red box. (C) The minimum free energy obtained from the RNAfold webserver for each mutation was plotted against normalized expression.

Table 1

List of Igα mutant codons, their expressions and calculated RNA Minimum Free Energies (MFE).

Amino Acid	Codons	Normalized Expression	MFE (kJ/mol)
P	CCG	2.20 ± 0.43	-4.50
A	GCG	2.14 ± 0.04	-4.64
C	UGC	2.11 ± 0.21	-9.50
T	ACC	2.11 ± 0.35	-3.14
S	UCG	2.08 ± 0.42	-7.40
K	AAA	2.0 [†]	-5.70
	AAG	1.70 ± 0.17	-3.14
I	AUA	1.91 ± 0.05	-2.52
D	GAU	1.90 ± 0.13	-3.60
F	UUC	1.81 ± 0.16	-5.20
Q	CAG	1.80 ± 0.36	-7.20
G	GGG	1.74 ± 0.13	-4.90
V	GUG	1.64 ± 0.23	-6.20
Y	UAU	1.60 ± 0.16	-6.20
N	AAU	1.25 ± 0.06	-3.60
W	UGG	1.18 ± 0.15	-6.30
L	CUG*	1.00	-5.70
R	CGG	0.92 ± 0.08	-6.30
E	GAG	0.50 ± 0.05	-3.14
M	AUG	0.30 ± 0.02	-4.70
H	CAU	0.17 ± 0.07	-6.60

[†] The AAA codon was not analyzed in triplicate.