Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*

James F Kane

SmithKline Beecham Pharmaceuticals, King of Prussia, USA

Within *Escherichia coli* and other species, a clear codon bias exists among the 61 amino acid codons found within the population of mRNA molecules, and the level of cognate tRNA appears directly proportional to the frequency of codon usage. Given this situation, one would predict translational problems with an abundant mRNA species containing an excess of rare low tRNA codons. Such a situation might arise after the initiation of transcription of a cloned heterologous gene in the *E. coli* host. Recent studies suggest clusters of AGG/AGA, CUA, AUA, CGA or CCC codons can reduce both the quantity and quality of the synthesized protein. In addition, it is likely that an excess of any of these codons, even without clusters, could create translational problems.

Current Opinion in Biotechnology 1995, 6:494–500

Introduction

Escherichia coli remains a popular host for the expression of heterologous proteins. When expression is 'turned on' by the addition of some inducing agent, large quantities of heterologous mRNA are made. As a result, the translational machinery of the *E. coli* cell is more likely encounter this heterologous mRNA than it would encounter a homologous mRNA. The consequences of such an event could range from little or no problem to an enormous problem, depending upon the composition of the mRNA and the heterologous protein.

E. coli, and indeed all cells, uses a specific subset of the 61 available amino acid codons for the production of most mRNA molecules [1,2]. So-called major codons are those that occur in highly expressed genes, whereas the minor or rare codons tend to be in genes expressed at a low level. This division is intriguing from an evolutionary point of view and has been the subject of some recent studies [3•]. From a pragmatic perspective, however, it could create a huge problem when the goal is to make large quantities of high-quality heterologous protein. If the message were to contain a distribution of codons, or the protein a distribution of amino acids that *E. coli* normally encounters, then no problem would probably exist regarding the cells ability to acylate the appropriate tRNA molecules, keep up with translation, and minimize mistranslational events. If, however, the mRNA from the cloned gene were to contain rare codons (see Table 1), or if the amino acid distribution were inordinately skewed relative to typical *E. coli* proteins (see Table 2), then it is likely that translational problems would occur during the production phase, leading to a reduction in either the quantity or quality of the protein synthesized.

Table 1. Codons used by *E. coli* at a frequency of <1%*.

Rare codons	Encoded amino acid	Frequency per 1000 codons	References detailing known problems with codons
AGG/AGA	Arg	1.4/2.1	[5–7,9,10,14,15•]
CGA	Arg	3.1	[20••]
CUA	Leu	3.2	[16••]
AUA	Ile	4.1	[a]
CCC	Pro	4.3	[25•]
CGG	Arg	4.6	–
UGU	Cys	4.7	–
UGC	Cys	6.1	–
ACA	Thr	6.5	–
CCU	Pro	6.6	–
UCA	Ser	6.8	–
GGA	Gly	7.0	[18•]
AGU	Ser	7.2	–
UCG	Ser	8.0	–
CCA	Pro	8.2	–
UCC	Ser	9.4	–
GGG	Gly	9.7	[18•]
CUC	Leu	9.9	–

*Taken from Wada *et al.* [2]. A cut-off of <1% was used to arbitrarily define rare codons. [a] B DelTito, F Watson, G Sathe, M Moyer, J Kane, abstract, FASEB J 1994, 8(suppl 7):A1305.

Abbreviations

BPL—bovine placental lactogen; **CAT**—chloramphenicol acetyltransferase; **COMT**—catechol O-methyltransferase.

The potential for translational problems is not surprising given the cellular environment after 'induction' of transcription of a cloned gene. Translation of the protein of interest could represent 60–70% of the total protein synthesized by the cell following induction. Although the *E. coli* cell has a tremendous capacity to produce high quantities of high-quality protein, there are limits when the composition of the mRNA or protein is not 'typical'.

Table 2. Amino acid composition of a typical *E. coli* protein (taken from Wada *et al.* [2]).

Amino acid	Occurrence in protein
Cysteine	1.08%
Tryptophan	1.28%
Histidine	2.23%
Methionine	2.65%
Tyrosine	2.88%
Phenylalanine	3.74%
Asparagine	4.02%
Proline	4.29%
Glutamine	4.33%
Lysine	4.85%
Threonine	5.37%
Aspartic acid	5.41%
Serine	5.70%
Arginine	5.74%
Isoleucine	5.78%
Glutamic acid	6.26%
Valine	7.12%
Glycine	7.45%
Alanine	9.46%
Leucine	10.03%

In this review, I focus solely on the effect of rare codons on expression in *E. coli.* For a review addressing the impact of amino acid distribution on protein quality, the reader is referred elsewhere [4].

Clusters of the rare arginine codons AGG/AGA

In *E. coli* mRNA, the specific rare arginine codons AGG and AGA occur at a frequency of ~0.14% and 0.21%, respectively. They were the first rare codons demonstrated to have a detrimental effect on protein expression [5–7]. AGG/AGA remain the focus of most of the attention on rare codons. Two distinct tRNA molecules [8], one of which has been cloned on a pACYC184 vector (i.e. the *argU* gene encoding $tRNA_{UCU}$), recognize these two codons. A mutation in the *argU* gene was first described as a temperature-sensitive DNA mutation and given the designation of *dnaY* [6,9]. This suggests a role for this rare tRNA species in some physiological function(s) other than just protein synthesis and may explain some of the different effects seen with these codons during high-level expression.

The effects on expression of AGG/AGA codons can be assessed in two ways: using model systems, or studying high-level expression of heterologous proteins. In general, the objectives are different for each. In the former case, test proteins contain several tandem AGG (or other) codons inserted at various positions in the mRNA. Shake-flask cultures of the host-vector are induced at low optical densities (>1.0 at 600 nm) for relatively short periods of time (10 min to 2 h). These systems provide a snap-shot of translation during the period of measurement, but may not show the cumulative effects of severe physiological stress. In the latter case, the objectives are to isolate enough heterologous protein either to do some structural and biological studies or to meet some commercial targets. These systems generally use fermentation vessels, with induction of the desired protein occurring at high optical densities (>5.0 at 600 nm) for longer periods of time (2–10 h). It is likely that large-scale fermentation would either magnify problems present in the model systems at very low levels or generate a different set of problems associated with the large-scale process. The effects of these codons, however, can be measured directly by comparing protein expression in a host with or without the plasmid containing the *argU* ($tRNA_{UCU}$) gene.

Rosenberg *et al.* [10] developed a codon test system that measures the effects of rare codon clusters on the expression of contiguous test and control genes from gene 9 of phage T7. This system is very useful in that the test gene and the control gene are adjacent on the vector and are controlled by the same T7 promoter, with a chromosomal copy of the T7 polymerase in the host *E. coli* B strain BL21(DE3). The test does not depend on enzymatic activity, but measures product as labeled protein and is designed for easy insertion of various codons of interest for testing. Although some of the disadvantages of model systems are enumerated above, it is nevertheless a very useful system to look for problematic codons. Specifically, these investigators studied the effects of two to five tandem AGG codons within a message encoding a 312 amino acid protein. They found that the magnitude of the effects of these tandem insertions depended upon their number and position within the mRNA. AGG clusters from two to five had the most significant effects on expression of the test protein when placed after amino acid 13, with expression increasing slightly when inserted after amino acid 223. As the number of AGG codons increased, expression of the test protein decreased. As the AGG clusters approached the carboxyl terminus, effects on the test protein decreased. Even so, overall expression from both the test and the control genes dropped as the number of AGG codons increased from two to five. Presumably, these codons affect expression, even at the carboxyl terminus, by frameshifting and impacting the expression of the downstream control gene.

The expression of a 306 amino acid hemagglutinin fragment from an influenza virus (LA Wysocki, DP

Myers, JF Kane, abstract, FASEB J 1994, 8(suppl 7):A1305) provides another example of the 'dastardly deeds' wrought by tandem AGG codons. In this system, a pL promoter controls transcription of the heterologous product in an *E. coli* K-12 host containing a chromosomal copy of λ *cI857* gene. A temperature shift upwards induced product expression. Growth and expression occurred in 10 liter fermenters. Western analysis indicated a full-length protein together with one major and two minor molecular weight species. The major truncated species coincided with a +1 frameshift occurring at the second codon of an AGG tandem located at positions 208–209. Co-expression of the *argU* gene completely eliminated this prematurely terminated species, but did not affect the two minor molecular weight species. Interestingly, the culture medium also affected these values. In a rich medium, the full-length protein and frameshift fragment represented ~5% and 2% of total cell protein, respectively. In defined medium, total expression decreased to <1%, and production of the frameshift fragment equaled that of the full-length protein. Addition of arginine to the minimal medium did not restore expression levels to those seen in the rich medium, suggesting the supply of arginine alone was not responsible for low expression.

The human RNA polymerase II associating protein RAP74 is another example where arginine (and other rare) codons affect expression [11]. RAP74 was expressed using the *E. coli* B strain BL21(DE3) T7 polymerase system. The mRNA for this protein contains 10 AGG/AGA codons out of 512, or ~2% of the total. One tandem AGG occurs at codons 166 and 167. Wang *et al.* [11] synthesized a 375 bp segment (from codons 135 to 259) using optimal *E. coli* codons. Both the tandem and one other AGG codon was within this segment. The authors reported that a "...76 kDa protein was strongly induced in cells containing the altered vector ... but not in the original vector." An examination of the western blots indicates that although the 76 kDa protein did increase, so did a number of other bands. As a side note, it appeared that expression in this T7 system was not tight because the uninduced culture contained a similar number of immunoreactive bands. Such a 'leaky' expression system could confound the results, particularly if the protein is 'toxic' to the cell.

In a very unusual case, Gursky and Beabealashvilli [12] have reported a stimulation of expression of a chloramphenicol acetyltransferase (CAT) protein when the terminal three codons are AGG-AGG-UGA, rather than CGU-CGU-UGA. In their system, the CAT gene was linked to an out-of-frame *lacZα* fragment and was separated by a translation stop codon. When the mRNA contained CGU-CGU-UGA, the bulk of the protein was 215 amino acids, and *lacZα* (encoding 270 amino acids) was translated following a +1 frameshift. Although no band intensities were given, I estimate a 3 : 1 ratio of the 215 amino acid protein to the 270 amino acid protein (on the basis of a comparison of lanes 4 and 5 in Fig. 2 from Gursky and Beabealashvilli [12]). When AGG codons replaced CGU codons, the amount of the 215 amino acid protein increased ~6–10-fold, and there was evidence of a –1 frameshift protein at a level that I estimate to be 25% of the intensity of the 215 amino acid protein. The authors proposed that a –1 frameshift caused an increased mRNA stability, allowing more protein to be made. This is a very puzzling proposal because the level of the –1 frameshift product, which presumably occurred because of the low level of $tRNA_{UCU}$, was much less than the product that was translated by the very same rare $tRNA_{UCU}$. If enough $tRNA_{UCU}$ were present to translate the tandem AGG codons, then why would the ribosome pause and frameshift at the AGG pair?

An alternative explanation for the above data seems to me to be an example of a fortuitous attenuation [13]; that is, an unexpected control of transcription by translation of an upstream leader-like sequence in the mRNA. Attenuation depends upon the ribosome correctly reading a sequence in the mRNA in such a way that a loop forms in the message which feeds forward to stop the RNA polymerase from continuing with transcription. Stalling of the ribosome at some hungry codons, in this case by inserting AGG codons with and without a UGA codon, prevents the formation of the loops in the mRNA that are necessary for transcription termination. As a result, an enhancement in the level of transcript occurs as a result of increased synthesis of mRNA, and not as a result of an increased stability of mRNA arising from ribosomal frameshifting. This hypothesis would also explain why adding the $tRNA_{UCU}$ or changing the AGG to CGU codons reduced the amount of the 215 residue protein, removed the –1 frameshift protein and restored the +1 frameshift protein. Both conditions would allow the ribosome to continue translation and form the necessary secondary structures to stop transcription.

Even single AGG/AGA rare codons cause problems

It is not only clusters of AGG codons that cause translational problems. The mRNA for the protein bovine placental lactogen (BPL) contains nine single rare AGG/AGA codons out of 200 amino acids (i.e. 4.5%). An *E. coli* K-12 strain expressing BPL was grown in 10 liter fermentation vessels. Although BPL accumulated to <10% of total cell protein, there were physiological problems similar to those previously reported [6]. In addition, an unexpected translational phenomenon was observed [14]. One species of BPL isolated during purification had different physical properties from the full-length protein. It had a slightly smaller molecular weight, lost two tryptic peptides, and gained a new one. Normally, amino acid residues 74–109 generated two peptides when trypsin cleaved at Arg86 (three-letter amino acid code). From this unusual species of

BPL, however, peptides 74–86 and peptide 87–109 disappeared, and a new peptide appeared. This new peptide had an amino acid sequence consistent with residues 74–85 linked to 88–109. In other words, residue 86, arginine, and residue 87, leucine, were missing. The absence of arginine explained why this sequence was no longer trypsin-sensitive. The question was how did an in-frame deletion of two amino acids occur? Examination of the mRNA sequence for BPL provided a hypothesis to explain this phenomenon. Codons 85–87 were UUG-AGG-UUG, specifying LeuArgLeu. The hypothesis was that the ribosome had to pause at codon 86 awaiting arginyl $tRNA_{UCU}$ because the large number of AGG/AGA codons in the mRNA depleted the intracellular pool of this acylated species. The extended pause time allowed the peptidyl-tRNA in the P site (UUG) to hop over codon 86, and land on the identical UUG codon at position 87 and continue in-frame to complete protein synthesis. The result was a protein shortened by two amino acids, arginine and leucine. This event occurred at an astonishingly high rate of 2% or, ~100-fold more frequent than might have been predicted on the basis of frequencies of translational mistakes measured during normal growth.

In another study, the *argU* gene has been reported to impact expression in an unusual way. Hua *et al.* [15•] have reported that a plasmid encoding the *argU* gene has a significant effect on expression, even though only one AGG codon was present in the mRNA being expressed. It was not clear whether this effect was reproducible because only one experiment was shown, or whether the physiology of host strain, JA221, was enhanced in some non-specific manner by the expression of $tRNA_{UCU}$.

Impact of other rare codon clusters on expression

The AGG/AGA codon pair represent rare codons that can negatively impact expression of heterologous proteins. Do other rare codons have similar effects? CUA and AUA are the next rarest codon pair in *E. coli*, occurring only one and a half to twice as frequently as AGG/AGA [2,3•]. That makes them candidates for causing translational problems during expression of heterologous proteins.

Recently, Goldman *et al.* [16••] have examined the effects of nine consecutive CUA codons on the translation of test and control genes in a model system using the T7 polymerase promoter. In this case, however, addition of T7 phage initiated induction of the test and control proteins. These authors concluded that CUA codons affect expression only when they occur in the amino-terminal portion (codon 13) of the mRNA. No effects were seen when these nine consecutive codons appeared at positions 223 and 307. Additionally, no evidence was found for frameshifting at the CUA cluster. The hypothesis is that translational complexes are most likely to dissociate early in the message, rather than late in the message. Thus, pausing created by rare codons has a more drastic effect when it occurs early in the mRNA sequence, rather than later. In the model system, there are two notes of caution. First, it is not clear if the active T7 infection skews the results in this system. It would be helpful to study the expression of heterologous proteins whose mRNA contains a high number or tandem CUA codons to know if position effects are associated with these codons. Second, the study was carried out using *E. coli* B, and it would be helpful to know if *E. coli* K-12 would behave in an identical manner. It is not unreasonable to expect different strain backgrounds to contain different levels of these rare tRNAs.

Recent reports indicate that the AUA codons influence the expression or accumulation of at least two heterologous genes expressed in *E. coli.* One of these proteins was a fusion between a non-structural influenza viral protein and a portion of the hemagglutinin from an influenza virus (B DelTito, F Watson, G Sathe, M Moyer, J Kane, abstract, FASEB J 1994, 8(suppl 7):A1305). In this 304 amino acid protein, isoleucine residues occur at positions 188 and 189 and are encoded by AUA codons. The protein accumulated to <5% in defined medium, and no evidence was found of any frameshifts associated with the AUA codon arranged in tandem. The addition to the host of a plasmid containing a cloned *ileX* gene, which encodes the tRNA that recognizes this codon [17], allowed the product to reach levels of 25–30%, indicating that this acylated tRNA limited expression. Similar results were observed with a mupiricin-resistant isoleucyl-tRNA synthetase isolated from *Staphylococcus aureus* (J Ward, J Hodgson, H Edwards, C Gershater, personal communication). This mRNA contains 33 AUA codons out of 1024 codons, or 3.3% of the total. Tandem AUA codons are at positions 258–259 and 834–835. Expression is essentially undetectable in *E. coli* unless the *ileX* gene is co-expressed, at which point the isoleucyl-tRNA synthetase reaches 7–9% of the total cellular protein.

The mRNA for the RAP74 protein (which has 10 AGG/AGA codons in its mRNA) contains numerous other rare codons, such as 17 CGG/CGA and 15 CCC codons ([11]; Table 1). As indicated above, the expression level improved when these investigators synthesized a segment of the mRNA (codons 135–259) that included AGG codons in tandem at positions 166–167, and CGG codon in tandem at positions 180–181. Even so, this fragment only reduced the numbers of these problem codons from 10 AGG/AGA to six, 17 CGG/CGA to 14, and 15 CCC to 13. Because the CGG/CGA and CCC codons occur at a frequency similar to that of AUA [2], it would not be unexpected for translational difficulties to be observed. Apparently, these opportunities for translational mishaps did occur and were evident in the western blots presented in [11].

Martin *et al.* [18•] have recently synthesized a 2.210 kb gene for human tropoelastin because expression of the cDNA was reportedly poor [19]. The protein has both an unusual amino acid composition as well as a large number of codons rarely used by *E. coli*, namely, CCC (which occurs at a level of 3.1% and encodes proline) and GGG/GGA (which occur at levels of 5.7% and 10.8%, respectively, and encode glycine). The tropoelastin protein reached in excess of 20% of total cellular protein when expressed from the synthetic gene. Although not conclusive, this suggests that either, or both, of these codons could cause translational problems if their numbers reach a certain level.

Arginine has more than its share of poor codons, accounting for four of the top six rare codons. A recent paper by Curran [20••] indicates that the CGA codon would be a real problem for *E. coli* because it is decoded by tRNA_{ICG} and A·I pairing is very inefficient [20••]. A string of CGA codons reduced expression of β-galactosidase and increased frameshifting around these clusters. These results are very reminiscent of the AGG/AGA pair and should be considered a problem for high-level expression of heterologous proteins. One wonders whether changing the anticodon of tRNA_{UCU} to UCG would make a more efficient tRNA for decoding CGA.

Physiologically creating 'hungry' codons in the major codon set

It is important to consider the physiological demands on the cell during synthesis of heterologous proteins. With this in mind, it is not impossible for abundant codons to become limiting if the demand were great enough [21]. For example, lysine occurs at a frequency of ~5% in *E. coli* proteins (Table 2). Lysine hydroxamate inhibits the formation of lysyl-tRNA_{UUU}, which decodes the two lysine codons, AAA and AAG. Addition of this analog creates these 'hungry' codons, at which point the ribosome pauses awaiting the lysyl-tRNA_{UUU}, and frameshifting occurs at this site in either the +1 or –1 direction [22].

Expression of the heterologous protein RAP74 may also represent a situation where a major codon could be limiting [11]. Lysine residues comprise 12% of RAP74, or almost threefold the normal level found in *E. coli* proteins. The T7 polymerase, a very strong promoter, directs the transcription of the *rap74* gene in *E. coli* B strain BL21(DE3). Although Wang *et al.* [11] gave no details on the medium used in their study, it is safe to say that high-level expression of a protein that is 12% lysine could be a problem. Such a high demand for lysyl-tRNA could create hungry AAA and AAG codons. Numerous 'fragments' were obvious in the gels, even after re-coding a segment within the gene that contained the tandem AGG and CGG codons. It is likely that the numerous fragments represent frameshift products, although it is difficult to say whether the frameshifts are associated with these numerous lysine codons or with the other rare codons found in the mRNA. With respect to the lysine problem, it would be interesting to see whether the amount of lysine or the level of tRNA were limiting expression. Nevertheless, it is a testament to the resilience of *E. coli* that Wang *et al.* [11] were able to isolate 30 mg of protein from this system. Because mistranslations occur under much less severe situations [23], this 30 mg of protein should have been checked to ensure the quality remained high.

Starvation for lysyl-tRNA may also explain some of the difficulties associated with the expression of chicken linker histone H5 from the T7 promoter in an *E. coli* B strain [24••]. In this protein, five amino acids, lysine (23%), alanine (15%), serine (14%), arginine (12%), and proline (7%), constitute 71% of the protein. In the carboxy-terminal portion, which is particularly problematical, this distribution is even more skewed, and lysine (37%), alanine (18%), arginine (14%), serine (12)%, and proline (10%) constitute 91% of the protein. Couple these numbers with the presence of some rare codons, such as CCC for proline, and translational problems may be inevitable. Another histone called linker H1 has a similar distribution: lysine (29%), alanine (26%), proline (9%) serine (7%), and arginine (2%). In this case, Gerchman *et al.* [24••] indicate that the protein is expressed better than H5, but not as well as constructs lacking the highly positively charged carboxy-terminal region.

Ribosomal mutations and rare codons

It is appropriate to discuss the host in combination with the effect of rare codons on expression. As mentioned above, Spanjaard and Van Duin [5] were surprised that they saw a 50% frameshift at tandem AGG codons, particularly because their host contained a mutation in the *rpsL* locus. The *rpsL* gene encodes the S12 protein for the small ribosomal subunit, and mutations at this locus have been shown to increase the fidelity of translation. At first, it appeared that the mutation was having the opposite effect. Subsequently, Sipley and Goldman [21] reported that, contrary to their expectations as well, such *rpsL* mutations actually increased a programmed translational frameshift for release factor 2 in *E. coli*. These authors concluded that "Ribosome pause time at the frameshift site determines the frequency of the shift; this pause time can be affected by cognate tRNA availability; the longer the pause, the greater the shift frequency." It is, therefore, essential to understand the background of the host cell in order to properly interpret the effects of rare codons on expression. Put another way, some rare codons may only affect expression when they are in a *rpsL* host.

The above is possibly the explanation for some observations made by Vilbois *et al.* [25•] when they expressed catechol O-methyltransferase (COMT) in *E. coli*. Two forms of the protein were produced following induction with isopropyl-β-D-thiogalactopyranoside. Both were purified and subjected to electrospray mass spectrometry. These forms differed by ~1.423 kDa in molecular weight. Further analysis revealed that the sequence of amino acids was consistent with a frameshift occurring at the final amino acid codon, CCC. The CCC codon occurs in *E. coli* at a frequency of 0.43%, about the same as the AUA codon. The authors proposed that a +1 frameshift occurred at this final codon, allowing continued translation with the addition of 11 amino acids. The molecular weight expected from such a shift coincided with the observed molecular weight from the electrospray analysis. Analysis of the mRNA coding sequence for COMT revealed five CCC codons, which represented ~2.3% of the total number. This figure is lower than that proposed by Brinkmann *et al.* [6] to be a problem, and no clusters of CCC codons occur in the mRNA. Even so, these investigators used an *E. coli* K-12 strain SG13009 that contained an *rpsL* mutation [26]. Perhaps, in the presence of the *rpsL* mutation, this level of CCC codons can cause problems. It is not unprecedented that a single codon could be responsible for translational problems [14].

In support of the hypothesis proposed by Sipley and Goldman [21], the *rpsL* mutation markedly increased the frameshifting at the tandem AGG codons (positions 207–208) in an mRNA encoding an influenza hemagglutinin (LA Wysocki, DP Myers, JF Kane, abstract, FASEB J 1994, 8(suppl 7):A1305). In the prototrophic strain, the ratio of full-length protein to frameshift fragment was 2.3 : 1, whereas in an *rpsL* background, the ratio was only 0.5 : 1. These results clearly demonstrate the effect of the *rpsL* mutation on ribosomal frameshifting.

The degree of pausing alone is not sufficient to cause frameshifting, however. Tandem AUA codons in a hemagglutinin protein from a different strain of influenza are poorly translated and reduce overall expression, but no evidence for frameshifting was found associated with these codons. In the *rpsL* mutant, the expression levels dropped, but the qualitative profile of the protein produced did not change. That is, no frameshift products were observed, despite a prolonged pause at these AUA codons.

Conclusions

I think we are now in a position to answer some of the questions surrounding the impact of rare codons on expression. A subset of codons, namely AGG/AGA, AUA, CUA, CGA (most likely CGG), and CCC, appear to cause problems from a translational point of view. The results suggest that clusters of these codons create most translational errors, although simply the presence of a large number of these codons could introduce translational errors as well. We can readily detect the large errors, that is, the frameshifts and decreases in expression. More subtle changes, such as mistranslations, are probably occurring, although these are perhaps more difficult to sort out. The recommendation would be to examine the sequence of the heterologous gene for these codons and to either use a host containing a plasmid with the appropriate tRNA or synthesize the gene to remove the codons. An analysis of the amino acid usage is also recommended so that an appropriate growth and expression medium can be used. This latter analysis could help anticipate potential problems, even with the major codons. It should be kept in mind that the magnitude of the problem can depend on the genetic background of the host, at least with respect to the *rpsL* mutation, and the medium used to grow the host vector.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Zhang S, Zubay G, Goldman E: **Low-usage codons in *Escherichia coli*, yeast, fruit fly, and primates.** *Gene* 1991, **105**:61–72.

2. Wada K, Wada Y, Ishibashi F, Gojobori T, Ikemura T: **Codon usage tabulated from the genebank genetic sequence data.** *Nucleic Acid Res* 1992, **20**:2111–2118.

3. • Bagnoli F, Lio P: **Selection, mutations and codon usage in a bacterial model.** *J Theor Biol* 1995, **173**:271–281.
These authors use a mathematical approach to explain the occurrence of rare codons and their physiological significance.

4. Jakubowski H, Goldman E: **Editing of errors in selection of amino acids for protein synthesis.** *Microbiol Rev* 1992, **56**:412–429.

5. Spanjaard RA, Van Duin J: **Translation of the sequence AGG-AGG yields 50% ribosomal frameshift.** *Proc Natl Acad Sci USA* 1988, **85**:7967–7971.

6. Brinkmann U, Mattes RE, Buckel P: **High-level expression of recombinant genes in *Escherichia coli* is dependent on the availability of the *dnaY* gene product.** *Gene* 1989, **85**:109–114.

7. Spanjaard RA, Chen K, Walker JR, Van Duin J: **Frameshift suppression at tandem AGA and AGG codons by cloned tRNA genes: assigning a codon to argU tRNA and T4 tRNAarg.** *Nucleic Acids Res* 1990, **18**:5031–5036.

8. Komine Y, Adachi T, Inokuchi H, Ozeki H: **Genomic organization and physical mapping of the transfer RNA genes in *Escherichia coli* K12.** *J Mol Biol* 1990, **212**:579–598.

9. Garcia GM, Mar PK, Mullin DA, Walker JR, Prather NE: **The *E. coli dnaY* gene encodes an arginine transfer RNA.** *Cell* 1986, **45**:453–459.

10. Rosenberg AH, Goldman E, Dunn JJ, Studier FW, Zubay G: **Effects of consecutive AGG codons on translation in *Escherichia coli*, demonstrated with a versatile codon test system.** *J Bacteriol* 1993, **175**:716–722.

11. Wang BQ, Lei L, Burton ZF: **Importance of codon preference for production of human RAP74 and reconstitution of the RAP30/74 complex.** *Protein Eng Purif* 1994, **5**:476–485.

12. Gursky YG, Beabealashvilli RS: **The increase in gene expression induced by introduction of rare codons into the C terminus of the template.** *Gene* 1994, **148**:15–21.

13. Landick R, Yanofsky C: **Transcription attenuation.** In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology.* Edited by Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE. Washington, DC: American Society for Microbiology; 1987:1276–1301.

14. Kane JF, Violand BN, Curran DF, Staten NR, Duffin KL, Bogosian G: **Novel in-frame two codon translational hop during synthesis of bovine placental lactogen in a recombinant strain of *Escherichia coli*.** *Nucleic Acids Res* 1993, **20**:6707–6712.

15. • Hua Z, Wang H, Chen D, Chen Y, Zhu D: **Enhancement of expression of human granulocyte-macrophage colony stimulating factor by *argU* gene product in *Escherichia coli*.** *Biochem Mol Biol Int* 1994, **32**:537–543.

A somewhat unusual observation. One rare AGG codon in human granulocyte-macrophage colony stimulating factor is shown to have a deleterious effect on expression.

16. •• Goldman E, Rosenberg AH, Zubay G, Studier FW: **Consecutive low-usage leucine codons block translation only when near the 5′ end of the message in *Escherichia coli*.** *J Mol Biol* 1995, **245**:467–473.

An elegant model (described by these authors in an earlier paper [10]) is used to investigate the effects of rare leucine codon clusters on gene expression. In this system, CUA codons had major effects when located near to the 5′ end of the mRNA. These strong positional effects are different from the authors' observations using the rare arginine codons AGG/AGA.

17. Muramatsu, T, Nidhikawa, K, Nemoto, F, Kuchino,Y, Nishimura, S, Miyazawa, T, Yokoyama, S: **Codon and amino-acid specificities of a transfer RNA are both converted by a single post-transcriptional modification.** *Nature* 1988, **336**:179–181.

18. • Martin SL, Vrhovski B, Weiss AS: **Total synthesis and expression in *Escherichia coli* of a gene encoding human tropoelastin.** *Gene* 1995, **154**:159–166.

High levels of glycine codons GGA/GGG, which are the rarest of the four glycine codons, appear to affect expression levels in *E. coli*. These authors constructed a synthetic tropoelastin gene, adjusting codon content to better suit the *E. coli* polypeptide biosynthetic machinery, and obtained significantly improved expression compared with the wild-type gene.

19. Indik Z, Abrams WR, Kucich U, Gibson CW, Mecham RP, Rosenbloom J: **Production of recombinant human tropoelastin: characterization and demonstration of immunologic and chemotactic activity.** *Arch Biochem Biophys* 1990, **280**:80–86.

20. •• Curran JF: **Decoding with the A-I wobble pair is inefficient.** *Nucleic Acids Res* 1995, **23**:683–688.

This paper predicts the problems that would be encountered if the gene of interest had a high content of CGA codons. All in all, four out of six arginine codons could present translational difficulties.

21. Sipley J, Goldman E: **Increased ribosomal accuracy increases a programmed translational frameshift in *Escherichia coli*.** *Proc Natl Acad Sci USA* 1993, **90**:2315–2319.

22. Lindsley D, Gallant J: **On the directional specificity of ribosome frameshifting at a 'hungry' codon.** *Proc Natl Acad Sci USA* 1993, **90**:5469–5473.

23. Bogosian G, Violand BV, Jung PE, Kane JF: **The effect of protein overexpression on mistranslation in *Escherichia coli*.** In *The Ribosome: Structure, Function, and Evolution.* Edited by Hill WE, Dahlberg A, Garrett RA, Moore PB, Schlessinger D, Warner JR. Washington, DC: American Society for Microbiology; 1990:546–558.

24. •• Gerchman SE, Graziano V, Ramakrishnan V: **Expression of chicken linker histones in *E. coli*: sources of problems and methods for overcoming some of the difficulties.** *Protein Express Purif* 1994, **5**:242–251.

Chicken linker histone contains very high levels of lysine, and high expression may result in a physiological situation similar to the effects of starvation on the translation of lysine codons described in [22]. It is easy to imagine that a high physiological demand for lysine would have the same deleterious effects as the addition of lysine hydroxamate.

25. • Vilbois F, Caspers P, Da Pradad M, Lang G, Karrer C, Lahm H, Cesura AM: **Mass spectrometric analysis of human soluble catechol o-methyltransferase expressed in *Escherichia coli*: identification of a product of ribosomal frameshifting and of reactive cysteines involved in S-adenosyl-L-methionine binding.** *Eur J Biochem* 1994, **222**:377–386.

In a similar fashion to [14], this paper shows the power of electrospray mass spectrometry in helping to sort out translational problems. This is an essential technique in examining the quality of proteins synthesized in hosts, and in this specific instance, supports the concept that a single rare proline codon, CCC, is associated with frameshifting.

26. Gottesman S: **Role of sulA and sulB in filamentation by Lon mutants of *Escherichia coli* K-12.** *J Bacteriol* 1995, **145**:265–273.

JF Kane, Biological Process Sciences, Biopharmaceutical R&D, SmithKline Beecham Pharmaceuticals, King of Prussia, Pennsylvania 19406-0939, USA.
E-mail: James_F_Kane%notes@sb.com@INET