The EMBO Journal vol.3 no.3 pp.623 – 629, 1984

Mutagenesis of the three bases preceding the start codon of the β-galactosidase mRNA and its effect on translation in *Escherichia coli*

Anna Hui[1], Joel Hayflick[1], Kim Dinkelspiel[2] and Herman A. de Boer[1]*

[1]Molecular Biology Department, and [2]Organic Chemistry Department, Genentech Inc., South San Francisco, CA 94080, USA

*To whom reprint requests should be sent

Communicated by M.Gruber

The effect on the translation efficiency of various mutations in the three bases (the – 1 triplet) that precede the AUG start codon of the β-galactosidase mRNA in *Escherichia coli* was studied. Of the 39 mutants examined, the level of expression varies over a 20-fold range. The most favorable combinations of bases in the – 1 triplet are UAU and CUU. The expression levels in the mutants with UUC, UCA or AGG as the – 1 triplet are 20-fold lower than those with UAU or CUU. In general, a U residue immediately preceding the start codon is more favorable for expression than any other base; furthermore, an A residue at the – 2 position enhances the translation efficiency in most instances. In both cases, however, the degree of enhancement depends on its context, i.e. the neighboring bases. Although the rules derived from this study are complex, the results show that mutations in any of the three bases preceding the start codon can strongly affect the translational efficiency of the β-galactosidase mRNA.

Key words: mutagenesis/start codon/base context/translation initiation

Introduction

In *Escherichia coli*, the initiation of protein synthesis begins at a start codon which in most cases is an AUG. This start codon is at the center of an RNA fragment which is 30 – 40 bases in length and can be isolated from initiation complexes as an RNase-resistant oligonucleotide. By definition, such regions are called ribosome binding sites (reviewed in Steitz, 1979, 1980; Gold *et al.*, 1981; Kozak, 1983). In most ribosome binding sites, the start codon is preceded by a purine-rich region at a distance of 5 – 9 bases. This so-called Shine-Dalgarno sequence (Shine and Dalgarno, 1974) shows a variable degree of complementarity with a region close to the 3′ end of 16S rRNA. Both regions can be co-isolated from initiation complexes as an RNA duplex (Steitz and Jakes, 1975; Steitz and Steege, 1977). The Shine-Dalgarno region (SD-region) is thought to assist the 30S particle in positioning itself at the proper place with respect to the start codon on the mRNA. Variation of the distance between the ATG and the SD-region greatly affects the efficiency of the translation initiation process (Shepard *et al.*, 1982). These three components (the ATG, the SD-region and the length of the spacer in between) are the most well-defined components of functional ribosome binding sites. There is increasing evidence, however, that these three components alone do not define the efficiency of a ribosome binding site (Gold *et al.*, 1981; Kozak, 1983; Scherer *et al.*, 1980). For example, within the coding region of mRNAs, the sequence ATG can be found in and out of the natural reading frame. Although preceded at a proper distance by a Shine-Dalgarno-like sequence, such regions (pseudo-ribosome binding sites) probably are not recognized by ribosomes as protein initiation sites. With this observation in mind, Stormo *et al.* (1982) searched for additional features distinguishing functional ribosome binding sites from non-functional pseudo-ribosome binding sites. After analysis of 124 known ribosome binding site sequences, they concluded that functional ribosome binding sites have a non-random distribution of the bases outside the SD-region and the start codon throughout the entire region protected by ribosomes in initiation complexes. Conversely, pseudo-ribosome binding sites show a random base distribution in this region.

Along with defining the parameters that distinguish functional from non-functional ribosome binding sites lies the task of defining those that determine the efficiency of each ribosome binding site. The difference in the efficiency of various mRNAs can vary as much as 1000-fold (Ray and Pearson, 1974, 1975). The nature of the four bases following the SD-region (de Boer *et al.*, 1983a, 1983b) as well as the distance between the SD-region and the start codon (Shepard *et al.*, 1982) has a considerable effect on the translational efficiency. In a series of experiments where we varied the length of the region on the mRNA that is complementary to the 3′ region of the 16S rRNA, we found that an increase of the base paired region from 4 to 13 bp decreases the translation efficiency by 50% (de Boer *et al.*, 1983b). In addition, some mutations found between the SD-sequence and the start codon (summarized by Gold *et al.*, 1981) affect the expression levels, while other base changes in that region have no effect. Unfortunately, however, no meaningful conclusion can be drawn from these experiments since different systems were used.

To eliminate this kind of ambiguity, we minimized the number of variables by using systems in which only one component of a ribosome binding site is varied at a given time. In our previous studies of *E. coli* ribosome binding sites, we described the construction of a plasmid with a portable SD-region (de Boer *et al.*, 1983a, 1983b). With the use of small synthetic DNA fragments, we varied the four bases located immediately downstream from the SD-region. The expression levels of these mRNA variants could be compared directly since they contained identical 5′-untranslated regions, SD-regions, start codons and coding sequences. Finally, the relative expression levels of these variants were measured by assaying the same protein. It was shown that A or T residues following the SD-region are favorable for the translation efficiency, whereas G residues and to a lesser extent C residues were inhibitory to translation. Here, we describe the results of an investigation into the role of the three bases immediately preceding the AUG codon (called the ' – 1' triplet) in the initiation of translation. We describe the construction of a series

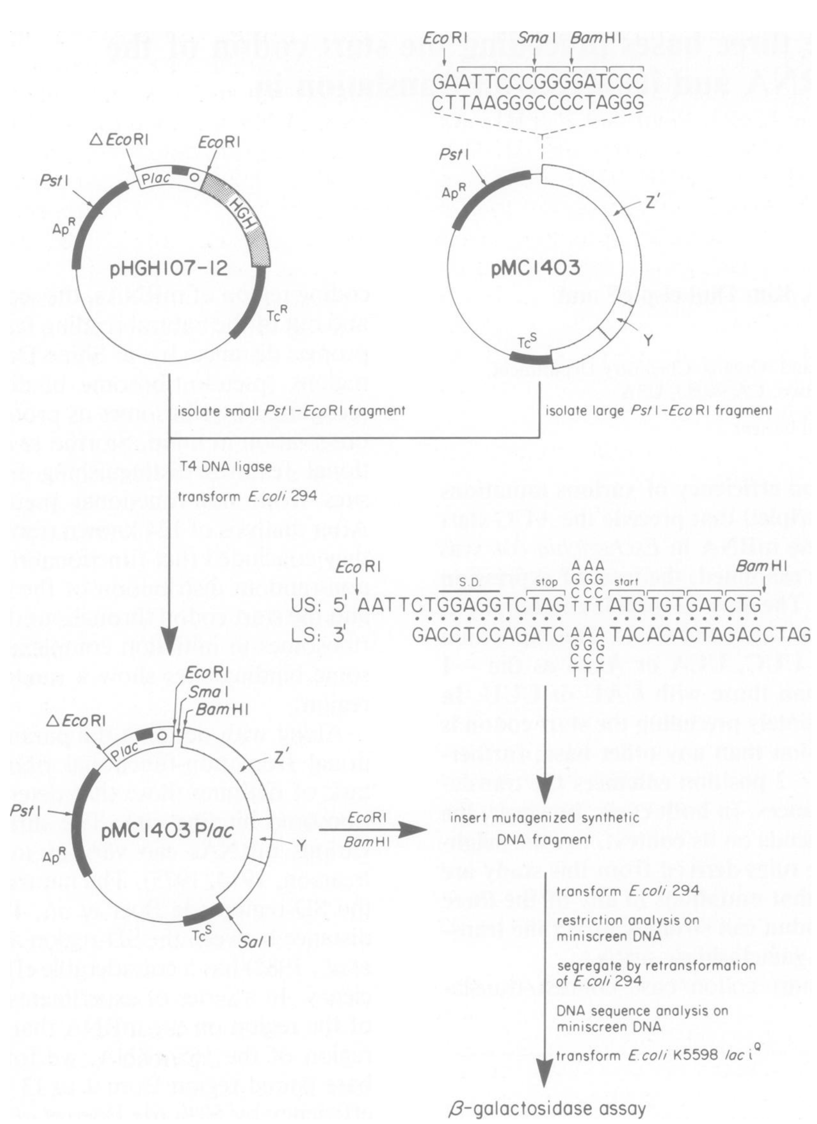
EcoRI
SmaI
BamHI
GAATTCCCGGGGATCCC
CTTAAGGGCCCCTAGGG
ΔEcoRI
EcoRI
PstI
Plac
O
HGH
Ap^R
pHGH107-12
Tc^R
PstI
Ap^R
Z'
pMC1403
Tc^S
Y
isolate small PstI-EcoRI fragment
isolate large PstI-EcoRI fragment
T4 DNA ligase
transform E.coli 294
EcoRI
BamHI
SD
stop
start
US: 5' AATTCTGGAGGTCTAG AAA GGG CCC TTT ATGTGTGATCTG
LS: 3' GACCTCCAGATC AAA GGG CCC TTT TACACACTAGACCTAG
EcoRI
SmaI
BamHI
ΔEcoRI
Plac
O
Z'
PstI
pMC1403 Plac
Ap^R
Y
Tc^S
SalI
EcoRI
BamHI
insert mutagenized synthetic DNA fragment
transform E.coli 294
restriction analysis on miniscreen DNA
segregate by retransformation of E.coli 294
DNA sequence analysis on miniscreen DNA
transform E.coli K5598 lac i^Q
β-galactosidase assay

Fig. 1. Plasmid pHGH107-12 has been derived from pHGH107-11 by deleting the upstream *Eco*RI site through a partial digestion with *Eco*RI followed by a nuclease S1 treatment and reclosure of the plasmid with T4 DNA ligase. The plasmid is resistant to ampicillin and tetracycline. Plasmid pMC1403 has been described by Casadaban *et al.* (1980). It contains the entire *lac*Z gene except for its first eight codons, which are replaced by sequences specifying the indicated restriction sites. Both plasmids were combined as shown, using standard procedures (Goeddel *et al.*, 1980; de Boer *et al.*, 1983a). Thus the plasmid pMC1403P*lac* was obtained. In this plasmid the *lac*-UV5 promoter transcribes the β-galactosidase gene. A translational start signal was provided by insertion of a synthetic DNA fragment whose design and properties are described in the text. The upper strand sequence of the 5′ region of hybrid *lac*Z mRNA including its mutated region is shown below. The three underlined sequences represent the *lac* operator (a), the synthetic insert (b), and the binding site for the sequencing primer (c), respectively. The sequence 'NNN' refers to the mutagenized −1 triplet.

```
                                      EcoRI                          BamHI
                                             SD     start stop
            •         •         •         •                   •         •         •
5′ AATTGTGAGCGGATAACAATTTCACACAGGAAACAGAATTCTGGAGGTCTAGNNNATGTGTGATCTGGATCCCGTCGTTT
            (a)                                       (b)

                                                                         PvuII
            •         •         •         •         •         •         •         •
   TACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGA
                                  (c)
```

of plasmids that are identical except for the sequence at the −1 triplet. Thus any difference in expression level is caused solely by the nature of this sequence. From our observation of a 20-fold range in expression levels depending upon the nucleotides in these positions, we conclude that the −1 triplet is an important element in determining the efficiency of the translation initiation process.

Results

Insertion of the lac-UV5 promoter into the β-galactosidase fusion vector

We have used short double-stranded DNA segments encoding a start codon which can be inserted in a simple ligation reaction between the *lac*-UV5 promoter and the 9th codon of the β-galactosidase (*lacZ*) gene. The plasmid pMC1403, described by Casadaban *et al.* (1980), contains the entire β-galactosidase gene with the first eight authentic codons replaced by sequences specifying unique *Eco*RI, *Sma*I and *Bam*HI sites. The *lac*-UV5 promoter was obtained from pHGH107-12 (Figure 1) which we derived from pHGH107-11 (de Boer *et al.*, 1983c) by destroying the *Eco*RI site upstream of the *lac*-UV5 promoter/operator. The *lac*-UV5 promoter from pHGH107-12 was joined to the *lac*Z gene on pMC1403 using the *Eco*RI site, resulting in pMC1403 P*lac* (Figure 1). To obtain ribosome binding site variants, pMC1403P*lac* was opened with *Eco*RI and *Bam*HI and a 31 bp long synthetic DNA fragment was inserted containing random mutations in the three bases preceding the start codon.

Design and synthesis of a DNA fragment with a randomized sequence preceding the start codon

A short synthetic DNA fragment was designed with an *Eco*RI and a *Bam*HI end suitable for insertion between the *lac*-UV5 promoter and the *lac*Z gene at the corresponding sites on pMC1403P*lac* (Figure 1). The sequence of the upper strand consists of all the features that are characteristic for a ribosome binding site. It has a Shine-Dalgarno sequence (5′-G-G-A-G-G) with a 5-bp homology with a region close to the 3′ end of 16S rRNA. The fragment contains an ATG start codon in phase with the β-galactosidase reading frame and located at a proper distance (eight bases) from the preceding SD-region. A stop codon (TAG) in phase with the reading frame of the *lac*Z messenger was introduced before the ATG to abort any translation that might come from upstream regions. The two bases (5′-TC) between the Shine-Dalgarno sequence and the stop codon provide for the optimal spacer length of eight bases (Shepard *et al.*, 1982).

To obtain as many mutations as possible in the three bases preceding the start codon, the sequences between the stop and the start codon were fully randomized. This was done by carrying out three consecutive coupling cycles with all four activated deoxynucleotides present in the reaction mixture. The monomer additions were done by the procedures described (Crea, 1980) for the trimer additions using the phosphotriester method (de Rooij *et al.*, 1979). The three triplets following the start codon are similar to those found in the human leukocyte interferon mRNA (Goeddel *et al.*, 1980). These triplets were chosen since we are interested in studying the expression of interferons. The lower strand is complementary to the upper strand, except for the three bases in the center which were also fully randomized. The upper and the lower strand were mixed and the 5′ termini were phosphorylated using T4 polynucleotide kinase and inserted into the *Eco*RI-*Bam*HI opened vector derived from pMC1403P*lac* using T4 DNA ligase. This DNA mixture was subsequently used to transform *E. coli* 294. The plasmid DNA in ampicillin-resistant colonies was screened for the presence of the synthetic insert and for the absence of the *Sma*I site. Plasmid DNA of positive candidates was used to re-transform *E. coli* 294 to accomplish segregation of the two types of mutants expected to be present in the primary transformants (Figure 1). Plasmid DNA from a single colony of the segregated transformants was isolated, and the DNA sequence of the insert and its surroundings was determined directly by the dideoxy-chain termination reaction on denatured plasmid DNA (E.Chen and P.Seeburg, unpublished results) using a synthetic primer (5′-GTTGGGTAACGCC) that binds to a region 40 bp downstream of the *Bam*HI site on pMC1403P*lac* (see legend to Figure 1).

After establishing the sequence of each clone containing a unique mutated ribosome binding site, part of each mini-screen DNA was used to transform *E. coli* K5598 (*laci*^{q}) and four transformants were grown in M9 minimal medium (Miller, 1972) and their β-galactosidase levels were determined according to the procedures of Putnam and Koch (1975). The host used in this study (*E. coli* K5598) contains excess *lac* repressor (*laci*^{q}) which, in the absence of inducer, greatly reduces transcription from the *lac* promoter (Miller, 1972; de Boer *et al.*, 1983b). Under these conditions only a small amount of *lac* message is produced, avoiding the accumulation of high β-galactosidase levels which undoubtedly would distort the measurement of relative translational efficiencies in the various mutants. We also measured the relative β-lactamase levels, reflecting the relative plasmid copy number in four different clones having high or low β-galactosidase levels, using the nitrocefin assay (see Materials and methods). Since no differences in plasmid copy numbers were found under the conditions used here, no correction for the relative β-galactosidase levels was necessary.

Using this approach, we isolated and characterized 39 different clones with a distinct mutation in one of the three bases preceding the start codon. Cells of *E. coli* K5598 containing a mutated plasmid were assayed for β-galactosidase activity and the results are shown in Table I. The mutants CAA and UUU were assayed each time and used as internal standards for the assay of all other mutants. The β-galactosidase levels relative to those found in the mutant CAA are shown in Table I. A 22-fold range in expression levels is observed among the various mutants. UAU and CUU (Table Ia), the expression levels are twice that of CAA. In the three mutants, UUC, UCA and AGG, the levels are 10-fold lower than that of CAA. The expression level of all other mutants, except GAA, varies from 0.1 to 1.0.

The following trends can be deduced from these data: in most cases (six out of ten) where the −2 and −3 position is constant and only the −1 position is different, a U residue at the −1 position is the most favorable base for expression levels, exemplified by the series CUN, GCN, UAN, AAN, UGN and AGN (N symbolizes any base). However, this rule does not hold for UUN and GGN and actually is reversed for AUN and CCN. The expression levels in all these exceptional series are relatively low. Apparently the bases UU, GG, AU and CC at the −3 and −2 position, respectively, have a dominating negative effect on the translation efficiency. Possibly, some triplets of the AUN series and also UUG might be used as a start codon (reviewed in Kozak, 1983). This may interfere with initiation at the proper AUG and thus may cause the relatively low expression levels. In all cases presented here, a U residue preceding the start codon is more favourable for expression than a C residue at the −1 position.

The very low level in the variant AGG could be explained by arguing that a false Shine-Dalgarno sequence is recognized; however, this argument does not hold for the mutants of the GAN series which, along with the preceding stop codon (UAG), also have a false Shine-Dalgarno sequence.

In the following, a comparison is made of some mutants that differ significantly in the expression levels but in only one base of the −1 triplet.

Comparison of mutants that differ only at position −1

In four out of the seven cases, a U to G change at the −1

Table Ia. Distribution of the − 1 triplet mutants according to their β-galactosidase levels

Relative β-gal. level:	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	2.0	2.1	2.2
– triplet:	UUC UCA AGG	UGG CUG AUU	UUG CCU CGG GCC GGC	UUU UCC GAG GGU	UCG UAG AUA AUG GCA	ACG AAC AGA GAC	CCG	UAA CCA CAC CAG CGC		UAC UGU CAA AAU AGU GCU			GAA	UAU		CUU

Table Ib. Relative β-galactosidase levels of the − 1 triplet variants arranged according to their sequence

Sequence	Relative β-gal. level	Sequence	Relative β-gal. level	Sequence	Relative β-gal. level	Sequence	Relative β-gal. level
UUU	0.4	UCU	–	UAU	2.0	UGU	1.0
UUC	0.1	UCC	0.4	UAC	1.0	UGC	–
UUA	–	UCA	0.1	UAA	0.8	UGA	–
UUG	0.3	UCG	0.5	UAG	0.5	UGG	0.2
CUU	2.2	CCU	0.3	CAU	–	CGU	–
CUC	–	CCC	–	CAC	0.8	CGC	0.8
CUA	–	CCA	0.8	CAA	1.0	CGA	–
CUG	0.2	CCG	0.7	CAG	0.8	CGG	0.3
AUU	0.2	ACU	–	AAU	1.0	AGU	1.0
AUC	–	ACC	–	AAC	0.6	AGC	–
AUA	0.5	ACA	–	AAA	–	AGA	0.6
AUG	0.5	ACG	0.6	AAG	–	AGG	0.1
GUU	–	GCU	1.0	GAU	–	GGU	0.4
GUC	–	GCC	0.3	GAC	0.6	GGC	0.3
GUA	–	GCA	0.5	GAA	1.3	GGA	–
GUG	–	GCG	–	GAG	0.4	GGG	–

All of the β-galactosidase levels were normalized to that found in the mutant CAA. The β-galactosidase levels were determined according to Putnam and Koch (1975). Each determination of the intracellular β-galactosidase level was performed on three independent colonies containing the same mutant plasmid. The assay of each culture was carried out in triplicate. The average and the standard deviation of all six samples were calculated. The numerical value of the standard deviation was <10% of the value for the β-galactosidase level. On each day nine mutants were analyzed in this way, two of which (CAA and UUU) were assayed on each occasion as internal standards.

The average β-galactosidase level as determined with seven independent assays of the mutant CAA is 990 and that of UUU is 358 units. The average of the seven ratios of these seven assays is 0.36 (standard deviation is 0.02).

position lowers the expression levels 4-fold or more. In the three remaining cases (UUN, AUN and CCN), with a U to G change at − 1, the expression level is already low apparently for reasons other than the particular base at − 1. As suggested before, this comparison indicates that the combinations UU, AU or CC at − 3 and − 2 are affecting expression negatively. Interestingly, whereas AU at position − 3 and − 2 is unfavorable, these same two bases at positions − 2 and − 1, respectively, are favorable in both variants (UAU and AAU) available. These findings point to a favorable effect on translation of an A residue at position − 2.

Comparison of mutants that differ only at position − 2

In all five cases available with an A to U change at the − 2 position, the expression level is decreased by at least 2-fold. The inversion of the U and A in UAU, resulting in AUU, lowers the expression 10-fold. The series UAU (2.0) > AAU (1.0) > UAA (0.8) > UUU (0.4) > AUU (0.2) is of particular interest since it shows not only that those with an A at position − 2 are high expressors, but also that the context of this A residue at − 2 is important.

In four out of six cases with an A to G change at − 2, the expression is reduced; in two cases expression is unaltered, but such a change never enhances the expression. Similarly, an A to C change at − 2 lowers or has no effect. In no instance does such a change enhance the expression. In all four mutants with a C to U change at − 2, expression is lowered. The variant, CUU, which is the highest of all mutants, is the only exception of this comparison. Of the mutants with a C to G change at the − 2 position, four lower the expression level and one, which is already low, has no effect. In summary, the order for the second base in terms of its favorable effect on expression is A > C > G ≥ U. Table I also shows that none of the variants with an A at − 2 expresses extremely poorly. Apparently an A at − 2 can counteract to some extent the negative effect of certain neighboring bases.

Comparison of mutants that differ only at position − 3

In all three mutants with UU in the − 2 and − 1 position the − 3 base affects the expression level strongly. Although a C-residue at − 3 yields high expression, a U or A at − 3 results in very low levels of expression. With CA at the − 2, − 1 positions, a G at − 3 increases and a U at − 3 lowers expression level. In all other cases the − 3 position has a moderate effect on the expression levels.

Some examples have been reported (Hall *et al.*, 1982; Iserentant and Fiers, 1980; Gheysen *et al.*, 1982) showing that mRNA secondary structure can affect translation efficiency. To examine the question of whether our results could be explained in terms of secondary mRNA structure, we analyzed

Table II. Secondary structure analysis of the −1 triplet variants

IIa. Possible regions with secondary structure as can be found in all the −1 triplet variants

Region	5′	3′	Length	ΔG (kcal/mol)
1	1	21	6	−8.40
2	4	26	5	−8.40
3	22	64	7	−12.60
4	27	44	6	−7.00

IIb. Possible regions with secondary structure (in addition to that shown in IIa) as can be found in the five indicated variants. The relative β-galactosidase levels shown here are taken from Table I

Mutant	5′	3′	Length	ΔG (kcal/mol)	Relative β-gal. levels
AAC	40	55	6	−7	0.6
AAU	40	55	6	−7	1.0
ACG	52	81	8	−9.4	0.6
CCG	54	68	5	−11.7	0.7
CUU	6	57	5	7	2.2

The secondary structure analysis was performed on the 5′ region of the modified *lac*Z mRNA sequence as shown in the legend to Figure 1. The 5′ border and the 3′ border of the stem-structures (the numbers as shown refer to the sequence shown in the legend to Figure 1), are indicated, together with the length of the stem involved and its stability as calculated using the rule of Tinoco *et al.* (1973).

the first 90 bases of the 5′ region of the mRNAs in all the mutants with a computer program in which rules as described by Tinoco *et al.* (1973) were used to detect secondary structure. We used an upper limit of $\Delta G = -7$ kcal/mol for the stability of the stem structure as a threshold for its detection. We assumed that any stem structure with a free energy higher than −7 kcal/mol is unlikely to exist in protein-free mRNA. From this analysis it appeared that the 5′ region of all mutant mRNAs except AAC, AAT, ACG, CCG and CTT showed an identical pattern (Table IIa). These exceptional mutants each have one additional and unique calculated region of secondary structure (Table IIb). This analysis shows that most mutants, although their expression levels differ greatly, have an identical calculated secondary structure pattern. Among the mutants that have an additional secondary structure, no correlation can be found between the expression levels and the stability or location of these additional calculated stem structures. Clearly, such analysis does not reveal a trivial explanation for the observed differences in expression levels. In fact, it is questionable whether any of the calculated stem structures shown in Table II is stable enough to exist by itself in the absence of additional proteins. We realize, however, that the program used falls short in predicting secondary structures that might be stabilized *in vivo* by the presence of (ribosomal) proteins.

Discussion

Iserentant and Fiers (1980) and Hall *et al.* (1982) have suggested that secondary structure plays an important role in determining the translational efficiency. Although the genetic and biochemical evidence from the experiments of Hall *et al.* (1982) is quite convincing, no proof is yet available of the actual existence of such proposed secondary mRNA structures. The notion that secondary structure can play a role in the translatability of certain mRNAs cannot be dismissed; however, it is difficult to explain our results in that way. We believe that the primary structure itself is an underestimated source of information with regard to the translational efficiency of mRNAs. Experiments bearing on this question are underway to fuse some of our mutant ribosome binding sites to other messengers.

The data presented here show that the three bases preceding the start codon can affect the translational efficiency of the β-galactosidase mRNA over a 20-fold range. Previously (de Boer *et al.*, 1983a, 1983b), we showed that the four bases following the Shine-Dalgarno sequence affect the translational efficiency at least 5-fold. Recently, Matteucci and Heyneker (1983) described a dozen mutations scattered throughout the spacer region between the SD-sequence and the ATG of the chimeric bovine growth hormone mRNA. Some of these mutations affected the expression levels of up to 10-fold. Similarly, a number of other mutations in the spacer region of a variety of *E. coli* and phage mRNAs have been described (for a summary, see Gold *et al.*, 1981). Although some of these mutations affect expression levels, others do not. The only conslusion that can be drawn from these observations is that the primary sequence of the spacer region contains important information that, at least partly, determines the translational efficiency of a mRNA.

Recent studies (Scherer *et al.*, 1980; Gold *et al.*, 1981; Stormo *et al.*, 1982; Matteucci and Heyneker, 1983; de Boer *et al.*, 1983a, 1983b) indicate that G residues in general affect translation negatively. Exceptions to this trend can be found, however. A to G, and U to G mutations in the lambda cII mRNA are known which have no effect on the translational efficiency (Wulff *et al.*, 1980). It should be realized that the effect of mutations within the spacer regions can be mRNA specific, making it difficult to compare and interpret the effect of the reported mutations in the different mRNAs. In addition, the effect of a particular mutation can depend on its context, i.e. its neighboring bases. We suggest that such a context effect plays a role in the three bases before the ATG and possibly throughout the entire spacer region. This is based on our observation that a U residue at the −1 position and A residue at the −2 position is favorable for expression in most but not all cases, depending on the neighboring bases.

Two reports (Eckhardt and Luhrmann, 1981; Ganoza *et al.*, 1982) show that *in vitro* the base immediately adjacent to the 5′ side of the start codon affects the formation rate of ternary initiation complexes with 30S particles and fMet-tRNA. A U residue favors initiation complex formation more than any other base in that position. In these studies purified components and short (3−7 bases long) synthetic oligonucleotides containing an AUG were used. It has been suggested (Eckhardt and Luhrmann, 1981; Ganoza *et al.*, 1982) that this favorable effect of the U-residue in the −1 position might be caused by a 4-bp codon-anticodon interaction in which the A residue immediately adjacent to the 3′ side of the anticodon on the fMet-tRNA is involved. Our *in vivo* data support their *in vitro* data.

Of all the sequenced *E. coli* mRNAs as listed in Gold *et al.* (1981), a U residue at the −1 position is used in nearly 30% of the cases and an A residue at that position is used in nearly 50% of the sequences shown. However, those mRNAs that have a U residue at −1 are not necessarily translated efficiently. For example, the *trp*R and *lac*I mRNAs both have a U residue adjacent to the AUG and these mRNAs are inefficiently expressed. On the other hand, most ribosomal protein mRNAs (which are expressed efficiently) have an A residue at this position. This lack of a simple relationship between the nucleotide 5′ of the AUG and the efficiency of these mRNAs can be understood more easily if not only the specific base at a particular position but also the context of this base must be taken into account. Indeed, our results show that the effect of the neighboring base 5′ of the AUG is more complex than suggested by the *in vitro* studies (Eckhardt and Luhrmann, 1981; Ganoza *et al.*, 1982). For these reasons and considering the profound effect of mutations anywhere else within the spacer region, we doubt that an interaction between the anticodon loop of the initiator tRNA and the extended initiation codon, if existing, has any major role in modulating the translational efficiency. Evidence has been presented suggesting that the fMet-tRNA may not be involved at all in the first step of protein initiation, but that the primary event (Van Dieijen *et al.*, 1978; Van Duin *et al.*, 1980) is the binary complex formation between the 30S subunit and the mRNA, which is then followed by the fMet-tRNA addition (for a discussion of this issue, see Gold *et al.*, 1981). Thus we believe that the effect on translation of the − 1 base is exerted at the level of the 30S-mRNA interaction and not at the level of the fMet-tRNA/mRNA interaction.

The Shine-Dalgarno sequence shows a variable complementarity with the 3′ end of 16S rRNA. No obvious correlation has been found between the extent of complementarity and the efficiency of translation. For instance, some very efficient messengers like the coat cistrons of the RNA bacteriophages MS2 and Qβ have only a three to four nucleotide complementarity with 16S rRNA. Yet these mRNAs form stable initiation complexes with 30S ribosomes in the absence of fMet-tRNA (Van Duin *et al.*, 1980). In addition, the Trp leader mRNA has a three base complementarity of unusual composition, yet when fused to the human growth hormone gene (Goeddel *et al.*, 1979), for example, it results in a highly efficient translation system (de Boer *et al.*, 1982). In agreement with these findings is our recent report (de Boer *et al.*, 1983b) which shows that an increase in the complementarity from 4 to 8 or 13 bp results in a 50% decrease in expression level. Taken together, these observations indicate that the actual nature of the Shine-Dalgarno sequence plays only a minor role in determining the translational efficiency.

This leaves us with the conclusion that most, if not all, of the information necessary to set the efficiency of a particular mRNA lies in the RNA sequence outside the SD-region and the start codon. The effect of all of the described mutations in the regions on either side of the SD-sequence and the start codon can probably account for the 1000-fold variation in efficiency observed among *E. coli* and phage mRNAs. The class of mutations described here can account for a 20-fold variation and the mutations in the spacer region described by Matteucci and Heyneker (1983) and others (Gold *et al.*, 1981) can account for a 10-fold variation in efficiency. Using the methods described here, we obtained 60 mutants in the three bases immediately following the ATG, and here too we find a 20-fold variation in efficiency (in preparation). Furthermore, Kastelein *et al.* (1983a, 1983b) showed that deletions in the regions 5′ of the SD-region of the MS2 replicase gene can lower expression up to 100-fold.

It is likely that mutations in the ribosome binding site outside the SD-sequence and the AUG affect the interaction between ribosomal proteins and the mRNA, and that this interaction determines the initiation frequency. The AUG and the SD-sequence could merely be reference points instructing the ribosome where to start protein synthesis. The interaction between ribosomal proteins and the mRNA is likely to be very complex and involves several points of close contact, possibly involving more than one ribosomal protein. An explanation of our observations must be sought in the nature of the interaction between mRNA and ribosomal proteins. This complex protein/RNA interaction must be flexible enough to allow the small ribosomal subunit to recognize all of the different *E. coli* ribosome binding sites, and yet must be selective enough to distinguish real from pseudo-ribosome binding sites.

Materials and methods

E. coli K5598 (W3110 laci^q LacZl8 LacZ::Tn9 lacL8) was obtained from Dr. H. Miller. *E. coli* K12 294 *end*A, thi$^-$, hsr$^-$, hsm$_k^+$ was used for standard transformation procedures. The β-galactosidase assay was performed according to Putnam and Koch (1975). Cells were grown to OD_{600} of 0.35 and the β-galactosidase activity was corrected for minor differences in the cell density. In a number of control experiments, the β-galactosidase activity was calculated per mg of protein (as determined with the Bio-Rad assay kit) instead of per cell density. The same relative β-galactosidase levels were obtained in both cases. All the data presented here are based on β-galactosidase levels expressed in units [nmol of O-nitrophenol hydrolyzed/min/unit of cell density (OD_{600})]. The nitrocefin assay for plasmid-encoded β-lactamase was done as described by the manufacturer (Glaxo Research Ltd., Greenford, Middlesex, UK). Cells were grown to OD_{600} of 0.35 and 1 ml of cells were spun down. The pellet was resuspended in 50 μl 0.1 M sodium-phosphate buffer, pH 7.0. To 20 μl of this cell suspension, 20 μl lysozyme solution (2.5 mg/ml) was added and subsequently incubated on ice for 30 min. Following this, the suspension was frozen and thawed three times. The cell debris was pelleted and the supernatant was diluted 50- to 500-fold with the same phosphate buffer to a final volume of 1 ml. 50 μl nitrocefin solution (1 mg/ml dissolved in dimethylsulphoxide-0.1 M phosphate buffer) was added and the kinetics of the reaction was followed at 482 nm during the first 4 min after nitrocefin addition.

$$1 \text{ unit of } \beta\text{-lactamase} = \frac{\Delta OD_{482}}{OD_{600} \text{ of cells}} \times 1000.$$

Acknowledgements

We acknowledge Dr. M. Casadaban for donating the plasmid pMC1403, Dr. H. Miller for donating the strain *E. coli* 5598, Drs. R.A. Kastelein, M.Shepard and L.J.Comstock for critically reading the manuscript, Peter Ng for purifying the synthetic DNA fragments, Jeanne Arch for typing the drafts and the final manuscript, and Alane Gray for drawing the figure.

References

Casadaban,M., Chou,J. and Cohen,S. (1980) *J. Bacteriol.*, **143**, 971-980.

Crea,R. (1980) *Nucleic Acids Res.*, **8**, 2331-2348.

de Boer,H.A., Comstock,L.J., Yansura,D.G. and Heyneker,H.L. (1982) in Rodriguez,A.L. and Chamberlain,M.J. (eds.), *Promoters: Structure and Function*, Praeger Scientific Publishing Company, pp. 462-481.

de Boer,H.A., Hui,A., Comstock,L.J., Wong,E. and Vasser,M. (1983a) *DNA*, **2**, 231-235.

de Boer,H.A., Comstock,L.J., Hui,A., Wong,E. and Vasser,M. (1983b) in Papas,T.S., Rosenberg,M. and Chirikjian,J.G. (eds.), *Gene Ampification and Analysis*, Vol. 3, Expression of Cloned Genes in Prokaryotic and Eukaryotic Cells, Elsevier Publishing Co., Amsterdam, pp. 103-116.

de Boer,H.A., Comstock,L.J. and Vasser,M. (1983c) *Proc. Natl. Acad. Sci. USA*, **80**, 21-25.

de Rooij,J.F.M., Wille-Hazeleger,G., Van Deursen,P.H., Serdijn,J. and Van Boom,A. (1979) *Recl. Trav. Chim. Pays-Bas*, **98:11**, 537-548.

Eckhardt and Luhrmann (1981) *Biochemistry (Wash.)*, **20**, 2075-2080.

Ganoza,M.C., Sullivan,P., Cunningham,C., Hader,P., Kofoid,E.C. and Neilson,T. (1982) *J. Biol. Chem.*, **257**, 8228-8232.

Gheysen,D., Iserentant,D., Derom,C. and Fiers,W. (1982) *Gene*, **17**, 55-63.

Goeddel,D.V., Heyneker,H.L., Hozumi,T., Arentzen,R., Itakura,K., Yansura,D.G., Ross,M.J., Miozzari,G., Crea,R. and Seeburg,P.H. (1979) *Nature*, **281**, 544-548.

Goeddel,D.V., Yelverton,E., Ullrich,A., Heyneker,H., Miozzari,G., Holmes,W., Seeburg,P.H., Dull,T., May,L., Stebbing,N., Crea,R., Maeda,S., McCandliss,R., Sloma,A., Tabor,J.M., Gross,M., Familletti, P.C. and Pestka,S. (1980) *Nature,* **287**, 411-416.

Gold,L., Pribnow,D., Schneider,T., Shinedling,S., Singer,B.W. and Stormo, G. (1981) *Annu. Rev. Microbiol.,* **35**, 365-403.

Hall,M.N., Gabay,J., Debarbouille,M. and Schwartz,M. (1982) *Nature,* **295**, 616-618.

Iserentant,D. and Fiers,W. (1980) *Gene,* **9**, 1-12.

Kastelein,R.A., Berkhout,B. and van Duin,J. (1983a) *Nature,* **305**, 741-743.

Kastelein,R.A., Berkhout,B., Overbeek,G.P. and van Duin,J. (1983b) *Gene,* **23**, 245-254.

Kozak,M. (1983) *Microbiol. Rev.,* **47**, 1-45.

Matteucci,M.D. and Heyneker,H.L. (1983) *Nucleic Acids Res.,* **11**, 3113-3121.

Miller,J.H. (1972) *Experiments in Molecular Genetics,* published by Cold Spring Harbor Laboratory Press, NY, pp. 356-359.

Putnam,S.L. and Koch,A.L. (1975) *Anal. Biochem.,* **63**, 350-360.

Ray,P.N. and Pearson,M.L. (1974) *J. Mol. Biol.,* **85**, 163-175.

Ray,P.N. and Pearson,M.L. (1975) *Nature,* **253**, 647-650.

Scherer,G.F.E., Walkinshaw,M.D., Arnott,S. and Morre,D.J. (1980) *Nucleic Acids Res.,* **8**, 3895-3907.

Shepard,H.M., Yelverton,E. and Goeddel,D. (1982) *DNA,* **1**, 125-131.

Shine,J. and Dalgarno,L. (1974) *Proc. Natl. Acad. Sci. USA,* **71**, 1342-1346.

Steitz,J.A. (1979) in Goldberger,R.F. (ed.), *Biological Regulation and Development,* Plenum Publishing Corp., NY, pp. 349-399.

Steitz,J.A. (1980) in Chambliss,G., Craven,G.R., Davis,J., Davis,K., Kahan, L. and Nomura,M. (eds.), *Ribosomes: Structure, Function and Genetics,* University Park Press, Baltimore, pp. 479-495.

Steitz,J.A. and Jakes,K. (1975) *Proc. Natl. Acad. Sci. USA,* **72**, 4734-4738.

Steitz,J.A. and Steege,D.A. (1977) *J. Mol. Biol.,* **114**, 545-558.

Stormo,G.D., Schneider,T.D. and Gold,L.M. (1982) *Nucleic Acids Res.,* **10**, 2971-2996.

Tinoco,I., Borer,P.N., Dengler,B., Levine,M., Uhlenbeck,O.C., Crothers, D.M. and Gralla,J. (1973) *Nature,* **246**, 40-41.

van Dieijen,G., Zipori,P., van Prooijen,W. and van Duin,J. (1978) *Eur. J. Biochem.,* **90**, 571-580.

van Duin,J., Overbeek,G.P. and Backendorf,G. (1980) *Eur. J. Biochem.,* **110**, 593-597.

Wulff,D.L., Beha,M., Izumi,S., Beck,J., Mahoney,M., Shimatake,H., Brady,C., Court,D. and Rosenberg,M. (1980) *J. Mol. Biol.,* **138**, 209-230.

Received 19 December 1983