At Least Six Nucleotides Preceding the AUG Initiator Codon Enhance Translation in Mammalian Cells

Sequences flanking the AUG initiator codon influence its recognition by eukaryotic ribosomes. From a comparison of several hundred mRNA sequences, $\mathrm{CC}\genfrac{}{}{0pt}{}{\mathrm{A}}{\mathrm{G}}\mathrm{CCAUGG}$ emerged as the consensus sequence for initiation in higher eukaryotes. Systematic mutagenesis of a cloned preproinsulin gene confirmed the facilitating effect of A or G in position −3 (i.e. 3 nucleotides upstream from the AUG codon), C in positions −1 and −2, and G immediately following the AUG codon. The analysis of a new set of mutants now reveals that sequences slightly farther upstream are also influential, the optimal context for initiation being $\mathrm{(GCC)GCC}\genfrac{}{}{0pt}{}{\mathrm{A}}{\mathrm{G}}\mathrm{CCAUGG}$. Possible mechanistic implications of the repeating GCC motif are discussed.

The features that identify initiation sites in eukaryotic mRNAs are gradually coming into focus. Discovery of the m^{7}G cap (Shatkin, 1976) and the monocistronic character of most eukaryotic mRNAs (Shih & Kaesberg, 1973) were important breakthroughs. Later, when the list of published mRNA sequences grew long enough to reveal trends, two additional characteristics were noticed: in most cases, the AUG codon that lies nearest the 5′ end of the message is the functional initiator codon (Kozak, 1983), and the nucleotides surrounding the initiator codon are non-random: 95% of the published eukaryotic mRNA sequences have a purine, most often A, in position −3 (i.e. 3 nucleotides upstream from the AUG codon, which is numbered +1 to +3) and G predominates in position +4 (Kozak, 1981, 1984). Recent experiments with a cloned preproinsulin gene revealed that translation declines five to tenfold when either the A in position −3 or the G in position +4 is replaced by a pyrimidine, and translation decreases 20-fold when pyrimidines are substituted in both of those positions (Kozak, 1986).

The aforementioned surveys also revealed a preponderance of C residues in positions −1, −2, −4 and −5, although those positions were less-conserved than A in position −3. When the significance of the flanking C residues was tested by mutagenesis, there was a small decrease in translation upon substituting T for C in positions −1 and −2, but no effect upon mutating positions −4 and −5 (Kozak, 1986). That result left open the possibility that C in positions −4 and −5 might yet facilitate translation when part of a larger motif. Indeed, in both surveys of eukaryotic mRNA sequences (Kozak, 1981, 1984), G occurred more frequently than any other nucleotide in position −6. I undertook the following experiments to see if translation would improve when GCC was introduced as a unit in positions −6 to −4.

The experiments were carried out with derivatives of a simian virus (SV40)-based shuttle vector that expresses the rat preproinsulin II gene from the SV40 early promoter (Lomedico & McAndrew, 1982). Casette mutagenesis was used to vary the sequence near the initiator codon, and the yield of proinsulin from each mutant plasmid was monitored during short-term transfection of COS-1 cells as described (Kozak, 1986). The parental plasmid p255/11, shown in Figure 1, has a unique *Hin*dIII site upstream from the ATG codon that initiates preproinsulin and a *Bam*HI site eight nucleotides downstream from the ATG codon. (The initiator codon will be written as ATG in the experimental portion of this paper, inasmuch as all recombinant techniques and sequencing were performed at the DNA level.) As illustrated in Figure 1, the "B series" mutants were constructed by replacing the small *Hin*dIII–*Bam*HI fragment of p255/11 with a synthetic oligonucleotide; an ATG codon on the inserted oligonucleotide now initiates the translation of preproinsulin. The construction of mutants B133 and B146 has been described in detail (Kozak, 1986), and identical procedures were followed for B150 through B166.

The sequence preceding the ATG triplet was systematically varied to assess the contribution of nucleotides in positions −4 through −9. Figure 2 illustrates the results obtained with the first seven constructs. Formally, one can think of B153 as having been derived from B146 by transposing two nucleotides: GC in positions −7 and −6 of B146 is replaced by CG in B153. That rearrangement, which creates the desired GCC motif in positions −6 to −4 of B153, increased the yield of proinsulin about threefold. In mutant B150, the GCC motif was duplicated in positions −9 to −4, producing a

0022-2836/87/160947 04 $03.00/0

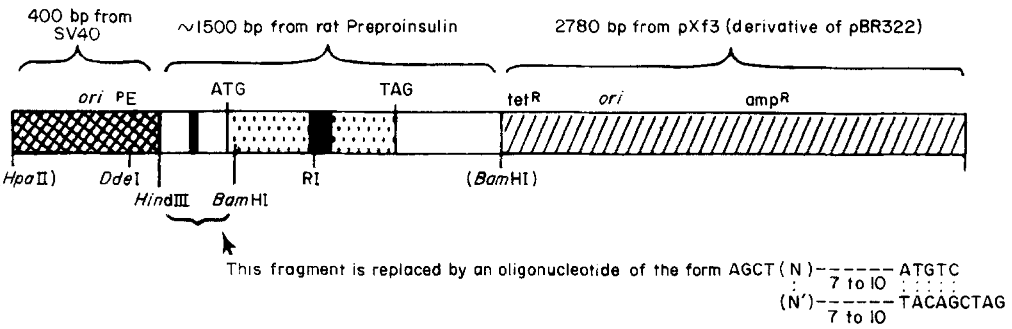
400 bp from SV40
~1500 bp from rat Preproinsulin
2780 bp from pXf3 (derivative of pBR322)
ori PE
ATG
TAG
tetR
ori
ampR
(HpaII)
DdeI
HindIII
BamHI
RI
(BamHI)
This fragment is replaced by an oligonucleotide of the form AGCT(N)------ATGTC
7 to 10
(N')------TACAGCTAG
7 to 10

Figure 1. Structure of the shuttle vector that encodes preproinsulin. The parental plasmid p255/11 is shown in linear form at the top. Below is a generalized version of the ATG-containing oligonucleotide that was inserted to generate mutants B150 through B166. The precise sequence preceding the ATG codon, shown here as $(N)_{7 \text{ to } 10}$, is defined for each mutant in Figs 2 and 3. bp, base-pairs.

further slight improvement in translation. (Although the difference between B150 and B153 was reproducible, the effect is small and may not be significant.) The yield of proinsulin from the control plasmid B152 was 80 to 90% lower than from B150, indicating that the stimulatory effect of GCCGCC is position-dependent: G in positions −6 and perhaps −9 enhances translation, whereas G

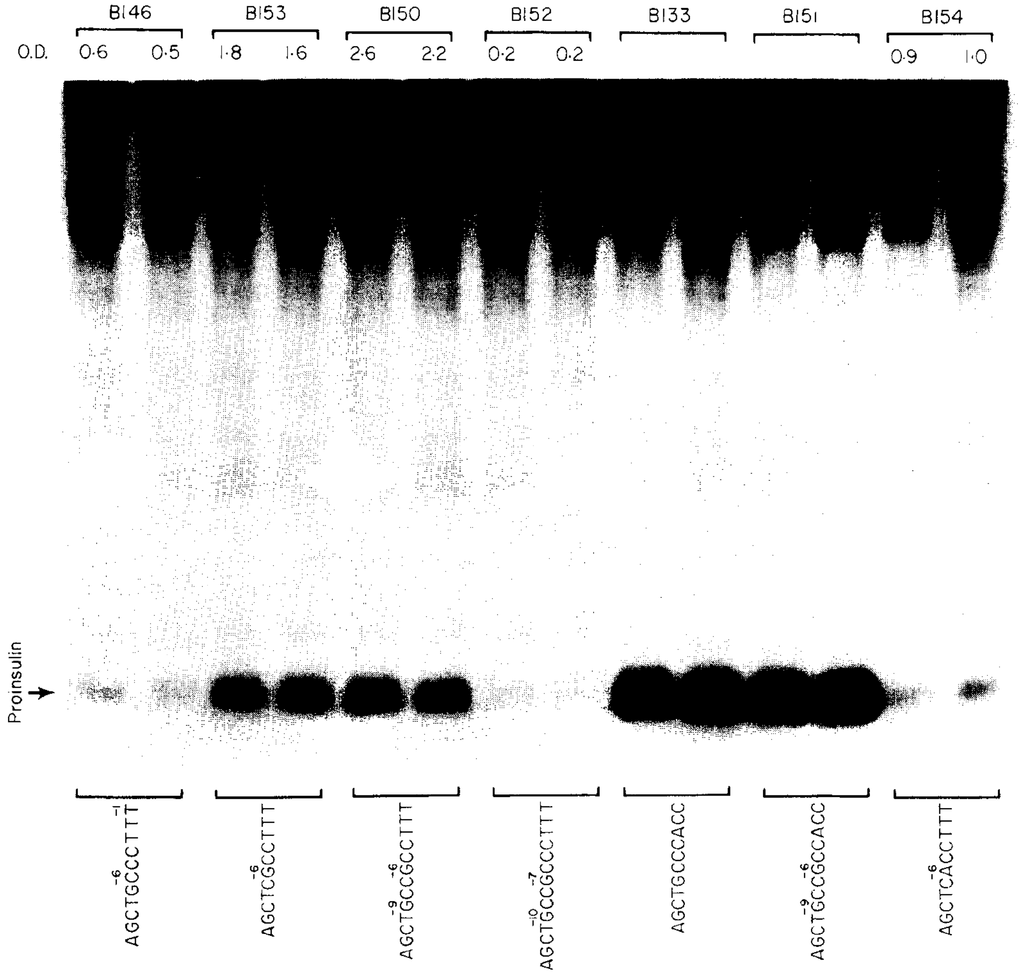

Figure 2. Effects of mutations in positions −6 to −10 on the synthesis of proinsulin in COS cells. At 48 h after transfection with the indicated mutant, the cells were incubated for 3 h with 1 ml of medium containing 0·25 mCi of [^{35}S]cysteine (1000 Ci/mmol). The cytoplasmic lysate from about 4×10^5 cells was used for immunoprecipitation with anti-bovine insulin antiserum (Miles Laboratories). The Figure shows a fluorogram of the immunoprecipitated proteins after polyacrylamide gel electrophoresis (Kozak, 1986). Adjacent lanes (bracketed) show the results from duplicate plates of cells that were transfected with the same plasmid. For each mutant, the sequence of the inserted oligonucleotide is shown up to the ATG codon that initiates synthesis of preproinsulin. The relative optical density (O.D.) of the proinsulin band is indicated above each lane, except for B133 and B151, which are overexposed here and therefore were quantified after a shorter exposure.

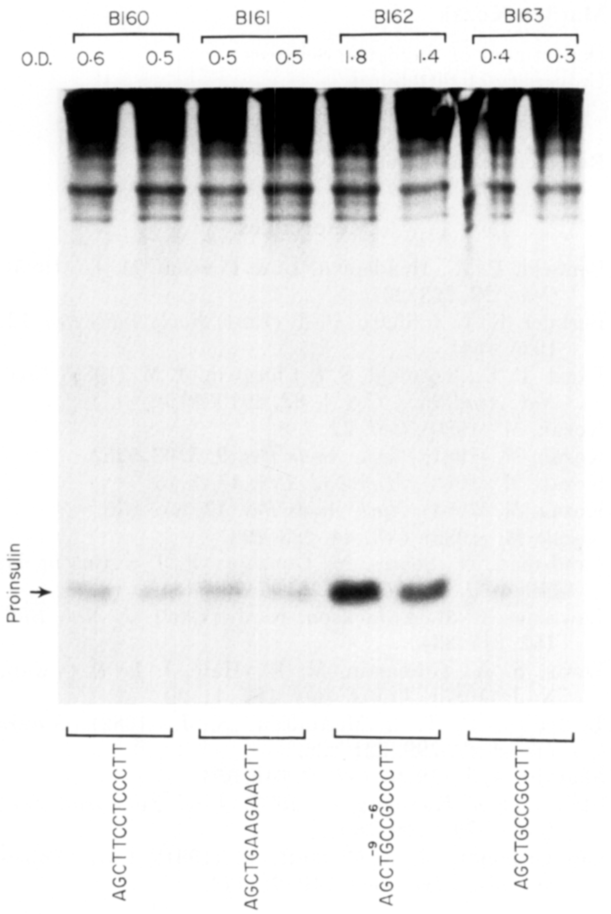

Figure 3. Enhanced synthesis of proinsulin requires G in positions −6 and −9 as well as C in the flanking positions. Details are given in the legend to Fig. 2.

residues in positions −7 and −10 have no effect. Whereas the stimulatory effect of the properly positioned GCC motif was easily detected in constructs that had an unfavorable sequence in positions −3 to −1 (B150 and B153), constructs that already had the optimal ACC sequence in positions −3 to −1 showed no further improvement when GCC was introduced upstream (B151 *versus* B133 in Fig. 2). This is consistent with our previous observation, that a high basal level of proinsulin synthesis due to the A in position −3 makes it difficult to detect the smaller contribution of other nucleotides (Kozak, 1986).

The foregoing conclusions were confirmed and extended by analyzing a few additional mutants. B166, B165 and B162 are matched constructs with no, one or two copies of GCC in positions −9 to −4; with those constructs, the GCC motif again improved translation about threefold (data not shown), confirming the effect shown in Figure 2 for B146, B153 and B150. Comparison of B162 with B161 reveals that G residues in positions −6 and −9 enhance translation only when C residues occur in positions −4, −5, −7 and −8 (Fig. 3). Comparison of B162 with B160 shows that C residues in positions −4, −5, −7 and −8 enhance only when G occurs in positions −6 and −9. The facilitating effect of the GCCGCC motif in B162 was once again dependent on its position; there was no stimulation when GCCGCC was shifted one nucleotide closer to the ATG codon in B163 (Fig. 3). All of these constructs expressed similar amounts of preproinsulin mRNA in transfected cells, as quantified by dot-blot hybridization.

In these experiments, as in earlier studies (Kozak, 1986), I have systematically mutagenized a cloned preproinsulin gene to identify local sequences that influence recognition of the AUG initiator codon. Although purines in positions −3 and +4 have the strongest influence, mutations in each of positions −1, −2, −4, −5 and −6 affect the yield of proinsulin, the optimal sequence for initiation being $\mathrm{(GCC)GCC}^{\mathrm{A}}_{\mathrm{G}}\mathrm{CCAUGG}$. (Although GCC in positions −9 to −7 always enhanced translation, the stimulation was slight and better evidence is needed to confirm its contribution. Translation clearly is not diminished by having an "extra" GCC motif in that position.) As explained previously (Kozak, 1980, 1986), the recognition sequence preceding the AUG codon is not believed to function as a direct "entry site" for ribosomes. Rather, the 40 S ribosomal subunit is thought to bind initially at the capped 5′ terminus and scan the mRNA sequence until an AUG is encountered; local context apparently modulates the efficiency with which the first AUG codon halts the migration of 40 S subunits.

Because the sequence $\mathrm{(GCC)GCC}^{\mathrm{A}}_{\mathrm{G}}\mathrm{CCAUGG}$ is internally repetitious, it is interesting to consider the possibility that the unit of recognition is just the three-nucleotide motif ${}^{\mathrm{A}}_{\mathrm{G}}\mathrm{CC}$. In other words, that triplet could have a facilitating effect when it precedes the AUG codon by three, six (or perhaps 9) nucleotides; an effect that is greatest when ${}^{\mathrm{A}}_{\mathrm{G}}\mathrm{CC}$ immediately precedes the AUG triplet, and that fades as it moves farther upstream. The mechanism by which ${}^{\mathrm{A}}_{\mathrm{G}}\mathrm{CC}$ facilitates initiation in unknown, but the rigid positional (i.e. phase) requirement seems like an important clue: although ${}^{\mathrm{A}}_{\mathrm{G}}\mathrm{CC}$ enhances when the purine is positioned three, six (or perhaps 9) nucleotides upstream from the AUG codon, there is no stimulation when the sequence is shifted by one nucleotide to the left or right (mutants B152 and B163 in Figs 2 and 3; see also B141 and B143 in Kozak, 1986). A possible explanation is that ${}^{\mathrm{A}}_{\mathrm{G}}\mathrm{CC}$ phases scanning, instructing the 40 S ribosomal subunit to read the next three nucleotides as a triplet rather than sampling all three frames. If that notion is correct, tandem repetition of the GCC motif might be especially favourable, since the phase established by the first GCC triplet in an array would be propagated, perhaps with increasing precision, until the 40 S ribosome reached the AUG codon. In some eukaryotic mRNAs, the GCC motif is indeed repeated from four (Faust *et al.*, 1985; Landsman *et al.*, 1986) to six times (Benfield *et al.*, 1985; Lewis *et al.*, 1985). The only such case in which translational efficiency has been studied is a late adenovirus transcript that encodes polypeptide IX: the sequence at the initiation site is GCCGCC-GCCGCCAUG (van Ormondt & Galibert, 1984) and that message is translated with extraordinary efficiency (Berkner & Sharp, 1984; Lawrence & Jackson, 1982).

Although nucleotides in each of positions −1 through −6 enhance translation, positions −3 and +4 clearly have the dominant influence: with A in position −3 or G in positions −3 and +4, translational efficiency is 60 to 90% of what one could achieve by including the entire consensus sequence (Kozak, 1986), which may explain why only about 5% of the eukaryotic mRNAs that have been sequenced contain a perfect consensus sequence. It is interesting that, of 78 vertebrate mRNA sequences that were selected because they conform to the consensus in at least four of the five positions from −1 to −5, 70% have G in position −6 (unpublished results). That circumstantial evidence, together with the mutational analysis described herein, encourages the hypothesis that G in position −6 is part of the ideal context for initiation in mammalian cells.

This work was supported by grant GM33915 from the National Institutes of Health.

Marilyn Kozak

Department of Biological Sciences
University of Pittsburgh
Pittsburgh, PA 15260, U.S.A.

Received 3 March 1987

References

Benfield, P. A., Henderson, L. & Pearson, M. L. (1985). *Gene*, **39**, 263–267.

Berkner, K. L. & Sharp, P. A. (1984) *Nucl. Acids Res.* **12**, 1925–1941.

Faust, P. L., Kornfeld, S. & Chirgwin, J. M. (1985) *Proc. Nat. Acad. Sci., U.S.A.* **82**, 4910–4914.

Kozak, M. (1980). *Cell*, **22**, 7–8.

Kozak, M. (1981). *Nucl. Acids Res.* **9**, 5233–5252.

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

Kozak, M. (1984). *Nucl. Acids Res.* **12**, 857–872.

Kozak, M. (1986). *Cell*, **44**, 283–292.

Landsman, D., Soares, N., Gonzalez, F. J. & Bustin, M. (1986). *J. Biol. Chem.* **261**, 7479–7484.

Lawrence, C. B. & Jackson, K. J. (1982). *J. Mol. Biol.* **162**, 317–334.

Lewis, S. A., Gilmartin, M. E., Hall, J. L. & Cowan, N. J. (1985). *J. Mol. Biol.* **182**, 11–20.

Lomedico, P. T. & McAndrew, S. J. (1982). *Nature (London)*, **299**, 221–226.

Shatkin, A. J. (1976). *Cell*, **9**, 645–653.

Shih, D. S. & Kaesberg, P. (1973). *Proc. Nat. Acad. Sci., U.S.A.* **70**, 1799–1803.

van Ormondt, H. & Galibert, F. (1984). *Curr. Topics Microbiol. Immunol.* **110**, 73–142.

Edited by S. Brenner